The New Encyclopædia Britannica

Volume 23

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

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

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



THE UNIVERSITY OF CHICAGO

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

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

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

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

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

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

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

© 1991
By Encyclopædia Britannica, Inc.

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

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

Printed in U.S.A.

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

CONTENTS

1 LIGHT
29 LIGHTING AND LIGHTING DEVICES
38 LIMA
42 LINCOLN
49 LINGUISTICS
81 LISBON
86 The Art of LITERATURE
224 The History of Western LITERATURE
229 LOCKE
234 The History and Kinds of LOGIC
291 LONDON
307 LOS ANGELES
313 The LOW COUNTRIES
364 LUTHER
373 MADRID
378 MALTA
382 MAMMALS
503 MANCHESTER
507 MANILA
511 MAO ZEDONG
515 MAPPING AND SURVEYING
538 MARKETING AND MERCHANDISING
552 MARKETS
569 MARSEILLE
573 Marx and MARXISM
586 MASKS
594 The Foundations of MATHEMATICS
603 The History of MATHEMATICS
634 MATRIX THEORY
639 MATTER: Its Properties, States, Varieties, and Behaviour
725 MAXWELL
727 Principles, Methods, and Instruments of MEASUREMENT AND OBSERVATION
792 MEASUREMENT THEORY
798 MECCA AND MEDINA
802 MECHANICS: Energy, Forces, and Their Effects
884 MEDICINE
939 MEDITERRANEAN SEA
944 MELBOURNE
947 MEMORY
956 MENTAL DISORDERS and Their Treatment
975 METABOLISM

Light

Light, a basic aspect of the human environment, cannot be defined in terms of anything simpler or more directly appreciated by the senses than itself. Light, certainly, is responsible for the sensation of sight. Light is propagated with a speed that is high but not infinitely high. Physicists are acquainted with two methods of propagation from one place to another, as (1) particles and as (2) waves, and for a long time they have sought to define light in terms of either particles or waves. In the early 19th century a wave description was favoured, though it was difficult to understand what kind of wave could possibly be propagated across the near-vacuum of interstellar space and with the extremely high speed of 300,000 kilometres per second (186,000 miles per second). In the latter half of the 19th century a British physicist, James Clerk Maxwell, showed that certain electromagnetic effects could be propagated through a vacuum with a speed equal to the measured speed of light. Thus, in the second half of the 19th century, light was described as electromagnetic waves (see ELECTROMAGNETIC RADIATION). Such waves were visualized as analogous to those on the surface of water (transverse waves) but with an extremely short wavelength of about 500 nanometres (one nanometre is 10^{-9} metre). The analogy is valid up to a certain point but the experimental results obtained at the end of the 19th century and in the early years of the 20th century revealed properties of light that could not have been predicted from knowledge that was obtainable about other waves. These results led to the quantum theory of light, which in its primitive form asserted that, at least in regard to its emission and absorption by matter, light behaves like particles rather than waves. The results of certain important experiments on the spreading of light into shadows and other experiments (on the interaction of beams of light) that supported the wave theory found no place in a particle theory. For a time it was believed that light could not be adequately described by analogy with either waves or particles—that it could be defined only by a description of its properties. A reconciliation of wave and particle concepts did not emerge until after 1924.

Two properties of light are, perhaps, more basic and fundamental than any others. The first of these is that light is a form of energy conveyed through empty space at high velocity (in contrast, many forms of energy, such as the chemical energy stored in coal or oil, can be transferred from one place to another only by transporting the matter in which the energy is stored). The unique property of light is, thus, that energy in the form of light is always moving, and its movement is only in an indirect way affected by motion of the matter through which it is moving. (When light energy ceases to move, because it has been absorbed by matter, it is no longer light.)

The second fundamental property is that a beam of light can convey information from one place to another. This information concerns both the source of light and also any objects that have partly absorbed or reflected or refracted the light before it reaches the observer. More information reaches the human brain through the eyes than through any other sense organ. Even so, the visual system extracts only a minute fraction of the information that is imprinted on the light that enters the eye. Optical instruments extract much more information from the visual scene; spectroscopic instruments, for example, reveal far more about a source of light than the eye can discover by noting its colour, and telescopes and microscopes extract scientific information from the environment. Modern optical instruments produce, indeed, so much information that automatic methods of recording and analysis are needed to enable the brain to comprehend it.

From the standpoint of wave motion, blue light has a somewhat higher frequency and shorter wavelength than red. In the quantum theory, blue light consists of higher energy quanta than the red.

The subject of light is so wide and its associations are so numerous that it cannot be accommodated within one article of reasonable length. There are three main divisions of the subject of light: physical optics, physiological optics, and optical instrumentation. This article deals primarily with physical optics, treating the nature and behaviour of light. It also discusses the interaction of light with matter and describes such phenomena as luminescence in considerable detail. Although electromagnetic theory is considered here, further elucidation may be obtained in the article ELECTROMAGNETIC RADIATION. The article SENSORY RECEPTION includes the physiological and psychological aspects of light, while the section *Optical Engineering* in the article OPTICS treats the theory and technology of lenses, mirrors, and optical systems. The experimental evidence that led to the quantum theory of radiation is included in the present article along with a brief statement of some of the basic ideas. The quantum theory of radiation, however, is so closely associated with the quantum theory of matter that the two must be considered together, as is done in the section on *Quantum mechanics* of the MECHANICS article. (R.W.Di./Ed.)

The article is divided into the following sections:

General considerations 1
  Historical survey 1
  Basic concepts of wave theory 3
  Light spectrum 6
  Velocity of light 7
Interference and diffraction phenomena 8
  Interference 8
  Diffraction 11
Polarization and electromagnetic theory 13
  Polarized light 13
  Electromagnetic-wave character of light 15
The interaction of light with matter 16
  Reflection and refraction 16
  Dispersion and scattering 17
  Mechanical effects of light 18
Quantum theory of light 19
  Photons 19
  The wave–particle nature of light 21
Luminescence 22
  Sources and process 22
  Early investigations 22
  Phosphorescence and fluorescence 23
  Luminescence excitation 23
  Luminescent materials and phosphor chemistry 24
  Luminescence physics 25

General considerations

HISTORICAL SURVEY

From 500 BC to AD 1650. In this period, there were innumerable confusions and false starts toward an understanding of light. Sometimes an idea was stated, though not clearly, and then almost forgotten for centuries before it reappeared and was generally accepted. The uses of plane and curved mirrors and of convex and concave lenses were discovered independently in China and in Greece. References to burning mirrors go back almost to the start of history, and it is possible that Chinese

and Greek knowledge were both derived from a common source in Mesopotamia, India, or Egypt. The formulation of general empirical laws and of speculation about the theory of light derives mainly from Mediterranean (Greek and Arab) sources. Pythagoras, Greek philosopher and mathematician (6th century BC), suggested that light consists of rays that, acting like feelers, travel in straight lines from the eye to the object and that the sensation of sight is obtained when these rays touch the object. In this way, the more mysterious sense of sight is explained in terms of the intuitively accepted sense of touch. It is only necessary to reverse the direction of these rays to obtain the basic scheme of modern geometrical optics. The Greek mathematician Euclid (300 BC), who accepted the Pythagorean idea, knew that the angle of reflected light rays from a mirror equals the angle of incident light rays from the object to the mirror. The idea that light is emitted by a source and reflected by an object and then enters the eye to produce the sensation of sight was known to Epicurus, another Greek philosopher of Samos (300 BC). The Pythagorean hypothesis was eventually abandoned and the concept of rays travelling from the object to the eye was finally accepted about AD 1000 under the influence of an Arabian mathematician and physicist named Alhazen.

Pythagorean hypothesis of light

Angles of incidence and of refraction—*i.e.,* the change in direction of a light ray going from one transparent medium to another—were measured by an astronomer, Ptolemy, in the 1st century in Alexandria. He correctly deduced that the ray is bent toward the normal (*i.e.,* the direction perpendicular to a boundary plane, such as the plane separating air and water) on entering the denser medium. A Dutchman, Willebrord van Roijen Snell, discovered the so-called sine law that gives the index of refraction (a measure of the change in direction) for light in a transparent medium. The laws of reflection and refraction were brought together by a 17th-century French mathematician, Pierre de Fermat, who postulated that the rays of light take paths that require a minimum time. He assumed that the velocity of light in a more dense medium is less than that in a less dense one in the inverse ratio of the indices of refraction.

Fermat's principle

The idea of rectilinear propagation of light—that is, that it travels in a straight line—was applied in a practical sense to drawing and painting long ago. Euclid was familiar with the basic idea, but the main theory was developed by Leonardo da Vinci, and a complete description of shadows was given by the Danish astronomer Johannes Kepler in 1604. Kepler also was the first to apply the laws of rectilinear propagation to photometry (the measurement of light intensities).

From 1650 to 1895. At the beginning of this period, the result of the conflict between the corpuscular theory and the wave theory was in doubt. At the end of the period, the wave theory was generally accepted and seemed capable of explaining all known optical phenomena though, with hindsight, it can now be seen that there were some important difficulties.

Diffraction—*i.e.,* the spreading of light into shadows—was first observed in Italy in the 17th century. In England, a worker, who independently noticed diffraction, also observed the interference colours of thin films, which are commonly seen today in an oil film on a wet road surface or in the iridescent colours of a butterfly's wing. He believed that light consists of vibrations propagated at great speed. Christiaan Huygens, of Holland, greatly improved the wave theory. In England, Sir Isaac Newton did not attach much importance to the small amount of spreading of light, and he knew that strictly rectilinear propagation could not be reconciled with the wave theory. Polarization phenomena (which can be accounted for by transverse wave motion in a single plane) discovered in the 17th century by a Danish physicist, Erasmus Bartholin, and by Huygens were not consistent with the theory of longitudinal waves (waves vibrating in the direction of propagation, like compression waves in a coiled spring), which was the only wave theory then considered. Newton therefore supported the corpuscular theory, although he did not reject the wave theory completely. He accepted a concept of a luminiferous ether, and he postulated that the particles had "fits of easy reflection" and "fits of easy transmission"; *i.e.,* he assumed that they changed regularly between (1) a state in which they were reflected at a glass surface and (2) a state in which they were transmitted. He thus introduced periodicity—one of the basic ideas of wave theory—in a form that anticipates the quantum mechanics. Newton, using a glass wedge, or prism, discovered that white light can be separated into light of different colours and took the first steps toward a theory of colour vision.

In the century following his death the great authority of Newton was quoted to uphold the corpuscular theory and to oppose the wave theory in a way that he probably would not have approved. It was not until the 19th century that the work of Thomas Young of England; Augustin-Jean Fresnel, François Arago, and Armand-Hippolyte-Louis Fizeau, all of France; Irish scientist Humphrey Lloyd; and German physicist Gustav Kirchhoff established the transverse-wave concept of light; *i.e.,* light is a wave vibration at right angles to the direction of travel. A universal medium pervading all space and called the ether was supposed to be some kind of elastic solid. This made it possible to accept the transmission of light through a vacuum, but there was no completely satisfactory theory of the ether or of the way in which light is modified by transparent materials like glass. The necessity for an elastic solid disappeared when Maxwell proposed an electromagnetic theory of light. He stated the laws of electromagnetism in a clear mathematical form and generalized the concept of an electric current. From his equations he predicted the existence of transverse electromagnetic waves having a constant speed *c in vacuo.* The constant *c* had a value of 300,000 kilometres per second and was derived from measurements on electrical circuits. It was known from the work of Ole Rømer, a Danish astronomer; Jean-Bernard-Leon Foucault of France; and others that the velocity of light was not much different from the velocity constant *c.* A.A. Michelson, a physicist in the United States, measured the velocity of light and showed that it is equal to *c* within a small margin of experimental error. This result, together with the work of a German physicist (Heinrich Rudolf Hertz) on electromagnetic waves of larger wavelength, confirmed Maxwell's predictions (see ELECTROMAGNETIC RADIATION). The existence of a connection between electromagnetism and light had, indeed, been demonstrated in England much earlier in the century by Michael Faraday, who observed the rotation of the plane of polarization of a beam of light by a magnetic field (Faraday effect).

Transverse wave concept established

From 1900 to the present. Maxwell's theory is a theory of waves in a continuous (*i.e.,* infinitely divisible) medium. The energy of the waves is also infinitely divisible so that an indefinitely small amount can be emitted or absorbed by matter. Classical physical theories of the 19th century had predicted that in such a system the energy in equilibrium would be distributed so as to give an equal amount to each mode (frequency) of vibration. Because a continuous medium has an infinite number of modes of vibration, and the atoms (which constitute matter) have only a finite number, all the energy of the universe would be transformed into waves of high frequency. Maxwell understood this difficulty, which was later most clearly stated in the Rayleigh–Jeans law (after two English physicists, Lord Rayleigh and Sir James Hopwood Jeans) of the radiation of a blackbody (a body in which the intake and output of energy are in equilibrium). The German physicist Max Planck demonstrated that it is necessary to postulate that radiant-heat energy is emitted only in finite amounts, which are now called quanta. At first, it was hoped to retain, without modification, the theory of light as electromagnetic waves in free space and to use the quantum concept only in relation to the interaction between radiation and matter. In 1905, however, Einstein showed that, in the photoelectric effect, light behaves as if all the energy were concentrated in quanta—*i.e.,* particles of energy now called photons. In the same year, Einstein published the theory of relativity, which modified the whole of physics and gave a special role to the velocity constant *c.* Because light, in some situations, behaves like waves and, in others, like particles, it is necessary to have

The role of the velocity of light in relativity theory

a theory that predicts when and to what extent each kind of behaviour is manifested. The main development of the quantum mechanics, which does precisely this, took place between 1925 and 1935.

Light from ordinary sources is emitted by atoms the phases of which are not correlated with one another, so that there is a random irregularity or incoherence between the waves emitted from different atoms. This places severe restrictions on the conditions under which the periodicity associated with wave theory can be observed. In England, Lord Rayleigh appreciated this effect and knew that, by the use of pinholes or slits and light of a narrow range of wavelength, effectively coherent light could be produced. For a long time, interest in this topic lapsed. About 1935 Frits Zernike, a Dutch physicist, and others extended the theory of coherence to include the concept of partial coherence. This appeared to be of practical importance only in a few rather special applications (*e.g.*, in the Michelson stellar interferometer; see below *Interference*). A theory of stimulated emission, attributable to the work of Einstein and an English physicist, Paul A.M. Dirac, postulated that under certain conditions atoms could be made to radiate in phase so that highly coherent radiation could be maintained indefinitely. The practical realization of these conditions, previously thought to be impossible, was achieved in 1960.

A second major development in the theory of light in this century is the application of so-called Fourier transform methods (a mathematical treatment of light waves) to a wide range of optical problems and, especially, to the transfer of information in optical systems (see OPTICS).

Information transfer in optical systems

Today, the theory of light has again reached a point at which all known terrestrial phenomena are included in one logical theory. The known unsolved problems concern the transmission of light over the vast distances of intergalactic space. Here the theory of light impinges on the science of cosmology.

BASIC CONCEPTS OF WAVE THEORY

In this section on the wave theory of light, those properties of light that are consistent with a wave theory are described using a minimum of mathematical formulation. It is convenient to introduce the basic concepts of wave theory in relation to mechanical systems. Below, in the section on *Interference,* and beyond, it will be necessary to consider results obtained by more sophisticated mathematical methods, such as Fourier analysis.

General characteristics of waves. *Periodicity in time and space.* If one end of a stretched rope is vibrated, a wave will run along the rope. Figure 1 (top) represents a profile of the wave—*i.e.*, a "snapshot" of the displacement of the rope from its normal position. It gives the variation of this displacement (indicated by ξ) at different points (z) along the axis of propagation for one specific instant of time. Similarly, Figure 1 (bottom) shows the variation with time of the displacement at one arbitrary point on the axis. In Figure 1 (top) the distance between successive crests is constant and is called the wavelength (λ). Similarly, the constant time between crests in Figure 1 (bottom) is called the period (τ). The temporal frequency ($\nu_t = 1/\tau$) is the number of vibrations per unit time and the spatial frequency or wave number ($\nu_s = 1/\lambda$) is the number of waves per unit length. The wave shown in Figure 1 (top) may be represented by the cosine of an angle (φ) to give the displacement for a particular point on the axis at any instant of time:

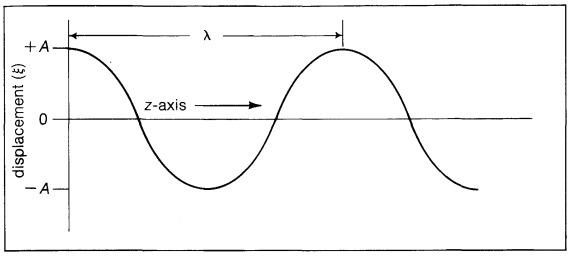


Figure 1: *Wave profiles.*
(Top) Variation with position at one time. (Bottom) Variation with time at one place (see text).

$$\xi = A \cos \Phi = A \cos 2\pi (\nu_t t - \nu_s z), \qquad (1)$$

in which ξ is the displacement at any point z on the axis at a time t, A is the amplitude (the maximum displacement); the angle φ (phi) in this case is equal to $2\pi(\nu_t t - \nu_s z)$ and is called the phase angle, or simply, the phase.

Energy. The energy per unit volume (W) stored in a wave motion is proportional to the square of the amplitude (A) so that, with a suitable choice of units, $W = A^2$.

Phase velocity. Any one crest moves forward a distance λ in a time τ; *i.e.*, with a velocity b of the wavelength divided by the period or the temporal frequency divided by the spatial frequency,

$$b = \frac{\lambda}{\tau} = \frac{\nu_t}{\nu_s} = \lambda \nu_t. \qquad (2)$$

The velocity b is called the phase velocity because the phase angle φ will remain constant when the time t changes by an incremental amount t_0 and z changes by $z_0 = bt_0$. (This may be seen by substituting $t = t_0$ and $z = z_0$ in the expression for this phase and using $b = \nu_t/\nu_s$.)

Relation of phase velocity to the medium

The velocity of light in vacuum (denoted by c) is the same for all frequencies; all colours travel through space with the same speed. The phase velocity (denoted by b) in a material medium, on the other hand, depends on the medium and on the temporal frequency and, hence from equation (2), on the wavelength.

Wave surfaces. Two-dimensional waves are formed by vibrating (dipping) the end of a rod up and down in the surface of a liquid. Waves spread from the point of origin (where the rod contacts the surface) and, at any moment, the phase at any point on a circle is the same; *i.e.*, if, at a given moment, the wave is at a maximum at one point on a circle then it is at a maximum everywhere on this circle, and the circle as a whole is a wave crest. Similarly, a trough is found at all points on another circle (the radius of which is $\lambda/2$ greater than that of the first circle). As the waves progress farther and farther from the origin, they become less strongly curved about the origin so that, at great distances, they are approximately plane waves.

Light waves are propagated in three dimensions and, for waves from a point source in an isotropic medium (*i.e.*, one in which the speed is the same along any radius), the phase is constant over spherical surfaces drawn about the point source as a centre. The surfaces of constant phase are called wave surfaces, and waves are called plane, spherical, ellipsoidal, and so on according to the shapes of the wave surfaces.

Reflection and refraction. The similarity between the behaviour of light waves and the surface waves of a liquid may be demonstrated with the so-called ripple tank. For reflection of a train of surface waves incident on a flat object, it may be readily observed that the angle of reflection is equal to the angle of incidence. For waves that are refracted in passing from one medium of the ripple tank in which the phase velocity is b_1, to another in which the phase velocity is b_2, measurements of angles of incidence (θ_i) and refraction (θ_r) of the surface waves verify Snell's sine law of refraction; *i.e.*, that the ratio of the sines of the angle of incidence and refraction is a constant, or

$$\frac{\sin \theta_i}{\sin \theta_r} = \frac{b_1}{b_2} = n_{12}, \qquad (3)$$

in which the constant n_{12} is called the index of refraction from medium 1 to medium 2. The index of refraction (n)

from vacuum to a material medium is called the index of the medium and, for transparent mediums is always greater than unity (one). When n_{12} is less than unity, as happens when light is refracted as it passes from glass into air, the refracted ray grazes the surface if $\sin \theta_i = n_{12}$, θ_i being the angle of incidence in the glass. At angles of incidence greater than this critical angle there is total reflection; *i.e.*, light, instead of penetrating into the air, is reflected back into the glass.

Dispersion. Newton found that, when a beam of white light is refracted by a glass prism, it is dispersed, or split, into beams of different colours. This phenomenon is now interpreted in the following way: the velocity of light in glass varies fairly rapidly with its wavelength, whereas its velocity in air varies little; thus the index of refraction and hence the angle of refraction depend on wavelength. A beam of white light, containing as it does a wide range of wavelengths, is thus dispersed by a glass prism so that light of one wavelength emerges from it in a different direction from light of another wavelength. Because colour depends on wavelength, the emergent light forms a spectrum (see Plate). All material mediums are, to some extent, dispersive (*i.e.*, phase velocity varies with the temporal or spatial frequency).

Wave groups. When a stone is dropped into a quiescent pond, a few waves may be seen travelling out from the point of impact. This group of waves maintains its identity as it is propagated over a considerable distance, although it finally dies away. The velocity of the group as a whole is called the group velocity. Careful observation shows that the group velocity is less than the phase velocity. Individual waves may be seen to appear at the back of the group, advance through it, and die out as they reach the front of the group. In a nondispersive medium the group velocity is equal to the phase velocity, while in a dispersive medium it may be greater than, less than, or equal. For light waves, the group velocity is almost always less than the phase velocity.

Group velocity

Interference. When two or more wave motions are present at the same place and time, the simplest assumption is that the resultant displacement (ξ_R) is the algebraic sum of the individual displacements (ξ_1, ξ_2, ξ_3, etc.), *i.e.*,

$$\xi_R = \xi_1 + \xi_2 + \xi_3 + \cdots + \xi_N. \qquad (4)$$

Nearly all observations on light are in accord with this equation, which is a statement of the principle of superposition. These phenomena constitute the subject of what is known as linear optics. The possibility that additional phenomena might be observed at high intensities of light has long been accepted, and the use of lasers in the attainment of the necessary high intensities has led to the discovery of frequency doubling and other effects that cannot be predicted from equation (4). These new observations constitute the material of nonlinear optics (see OPTICS). Equation (4) is valid for all the phenomena of interference, diffraction, etc., which will be described in this article.

Two waves are said to be coherent if their phase difference remains constant during a period of observation. Figure 2 shows two equal coherent plane waves travelling across the same space, with the wave fronts inclined at a small angle α, AB representing a surface corresponding to a crest of one wave. (The surface must be assumed to be perpendicular to the page.) C_1 D_1, C_2 D_2, etc., represent surfaces that correspond to crests of the other wave. The intermediate dotted lines represent troughs. At points such as P_1 (and P_2, P_3, . . .), a crest of one wave coincides with a crest of the other and according to the principle of superposition the displacement is twice that of either wave alone. At points Q_1, Q_2, etc., a crest of one wave meets a trough of another; so the displacements being equal and opposite, the resultant is zero. Thus, an observer looking at a plane that is perpendicular to the page and passes through AB sees a series of straight lines through P_1, P_2, P_3, etc., representing large displacement and a series of lines through Q_1, Q_2, Q_3, etc., representing zero displacement.

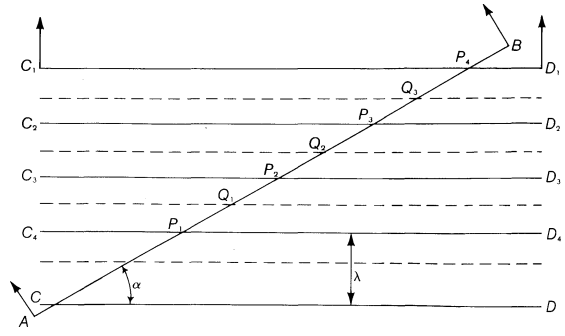


Figure 2: Interference of two plane waves *AB* and *CD* with directions inclined at an angle α. The crests of *CD* are represented as C_1D_1, C_2D_2, etc., and the troughs are shown as broken lines (see text).

There are many ways in which coherent beams of light can be made to cross at an angle of about one part in a thousand. The eye (or a low-power magnifier) can be focussed on a plane such as that through AB. The resulting parallel light and dark lines are called interference fringes (Figure 3). From Figure 2 it may be seen that the separation (d) of two bright fringes is λ/α or 1,000 λ if $\alpha = 0.001$. When α has this value, $d = 0.5$ millimetre for blue-green light and this would imply that λ is about 0.5×0.001 or 1/2,000 part of a millimetre (this is usually written 500 nanometres).

Interference fringes

Milward T. Rodine

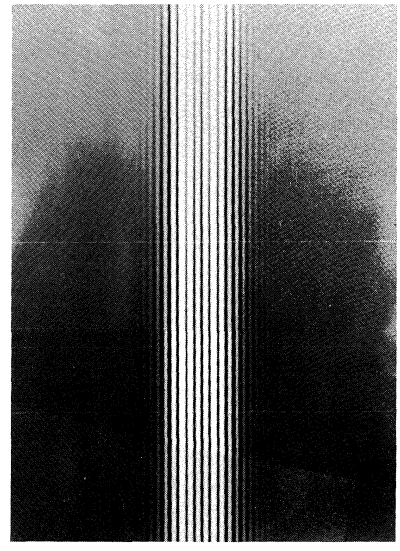


Figure 3: Two-beam interference fringes from Young's double slits or Fresnel's biprism (see text).

In this experiment the spatial periodicity of the light waves (about 2,000 waves per millimetre) has been made to produce fringes with periodicity of about two per millimetre. The spatial periodicity of a light wave is too high for the human eye, and it cannot be magnified directly. Interference methods effectively magnify it so that the resultant fringes can be seen by eye or with a convenient magnification. The following method of producing interference fringes, developed by Thomas Young, is now called Young's experiment.

In the arrangement shown in Figure 4, light of one wavelength passes through a slit S producing semicylindrical waves that are intercepted by two other slits P_1 and P_2. The two slits P_1 and P_2 act as secondary sources of coherent, semicylindrical waves the combined effect of which is observed on the plane perpendicular to the page and designated AB. In a typical case the separation (a) of P_1 and P_2 is a millimetre and the distances l_1 and l_2 are each about a metre. The slits are a centimetre or so long but are much less than a millimetre wide. They are accurately parallel to one another and, as represented in the drawing, are at right angles to the page. Because the waves from P_1 and P_2 are indirectly derived from the same small source, they are coherent. When they cross plane AB they are

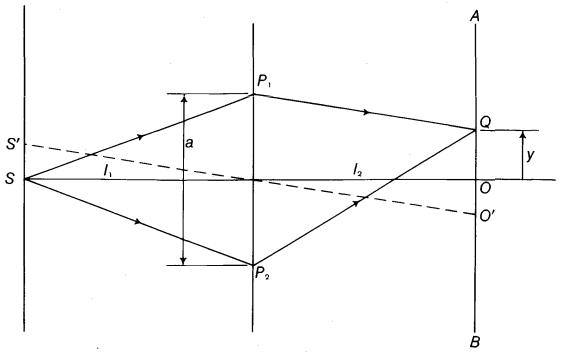


Figure 4: Young's experiment (see text).

nearly plane because of the large radius, and they intersect at an angle α equal to 0.001. It may be shown that the intensity (I) for these fringes varies from point to point along the line AB in the way shown in Figure 5 (curve A), which is in accord with the equation

$$I = 2A^2(1 + \cos\varepsilon) = 2I_0(1 + \cos\varepsilon)\,, \qquad (5)$$

in which A is the amplitude of either wave, I_0 is the intensity of one wave acting alone and the phase difference $\varepsilon = 2\pi ya/\lambda l_2$. Bright fringes are seen in positions for which $\varepsilon = 2\pi p$ or $y = p\lambda/l_2 a$ (in this case p is a whole number, which may be positive, zero or negative—0, ±1, ±2, ±3, etc.). Because cos ε varies from −1 to +1, I varies from $4I_0$ to zero. The average, in accordance with the law of conservation of energy, is $2I$.

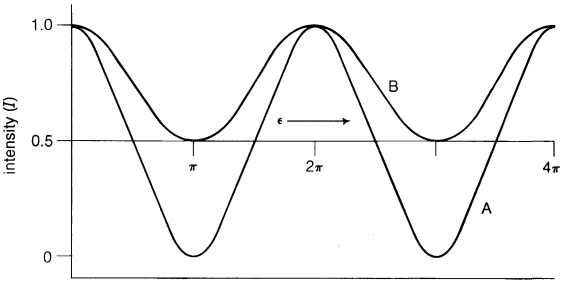


Figure 5: Interference fringes obtained in Young's experiment (see text).

Diffraction. Plane waves that pass through a restricted opening emerge as divergent waves. When the opening is less than one wavelength in diameter the emergent wave is nearly spherical. Whenever a beam of light is restricted by holes or slits or by opaque obstacles that block out part of the wave front, some spreading occurs at the edges of geometrical shadows. This effect, called diffraction, is also obtained with transparent obstacles that cause an irregularity in the wave front. Diffraction can be demonstrated by allowing a parallel beam of light to fall on a grating consisting of an array of equally spaced narrow slits. If the extent of physical separation of two adjacent slits is e, then the path difference between any two adjacent rays emitted in a direction symbolized by θ is $e \sin\theta$, and if this path difference is an integral number (p) of wavelengths,

$$e \sin\theta = p\lambda, \text{ or } \nu_s \sin\theta = pg\,, \qquad (6)$$

in which ν_s is the spatial frequency ($1/\lambda$) and g is the number of lines per unit width of the grating, then the waves from different slits have phases that differ by angles of $2p\pi$, and they reinforce one another. Thus, when lenses are employed with a grating, sharp lines are obtained for each wavelength at values of θ corresponding to integral values of p. If white light is used, each line is drawn out into a spectrum of wavelengths because the direction of reinforcement depends on the wavelength.

Polarization. In the propagation of waves on a rope or across the surface of a liquid the displacement (as shown in Figure 1) is in a direction perpendicular to the direction of propagation and the waves are said to be transverse. Sound waves in a gas consist of alternate dilation and compression and the displacement is in the direction of propagation. The waves are longitudinal. If a beam of longitudinal waves is propagated in a vertical direction, there is nothing to distinguish one azimuthal plane from another—everything that is true for an east-west plane is equally true for a north-south plane. With transverse waves the displacement may be in the east-west plane; in that case, there is no component in the north-south plane, and this should manifest itself in the form of a property that depends on the azimuth. Such an effect is called an azimuthal property. An ordinary beam of light from a thermal source does not exhibit any azimuthal property, but experiments show that light can have an azimuthal property and must be represented by transverse waves.

Azimuthal property

If an unsilvered glass plate has an index of refraction equal to 1.5 and the angle of incidence of a beam of light is 57°, about 15 percent of the light will be reflected from the two glass surfaces of the plate (Figure 6); this percentage will not be altered when the glass plate is rotated about an axis parallel to the beam of light so as to change the azimuth of the plane of reflection. If a second mirror (G_2), parallel to the first (G_1), is used to reflect the beam in the same plane as that of the original reflection, about 30 percent of the light incident on the second plate of glass will be reflected; but if the second plate is turned so as to reflect the light in a plane perpendicular to that of the first reflection—*i.e.*, out of the plane of the page—hardly any light will be reflected. Thus, after the first reflection, the beam of light will have acquired an azimuthal property—it will be reflected more strongly when the transverse displacement is in one azimuthal plane than when in another. Further tests will show that the transmitted light has a complementary azimuthal property; it is more strongly reflected in the perpendicular plane—though the difference is less marked.

These results may be understood if ordinary light consists of a mixture of transverse waves with displacements in all azimuthal planes but only one component is reflected from a glass surface when the angle of incidence is 57°. The reflected light is said to be plane-polarized because all of the displacement of the wave is in one azimuthal plane. The transmitted light (about 85 percent of the whole) contains about 50 parts of a component that is polarized in a perpendicular plane and about 35 parts of light that is polarized in the same way as the reflected light. It is more strongly reflected in the plane of the page, but because it is only partially polarized, the azimuthal effect is less.

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

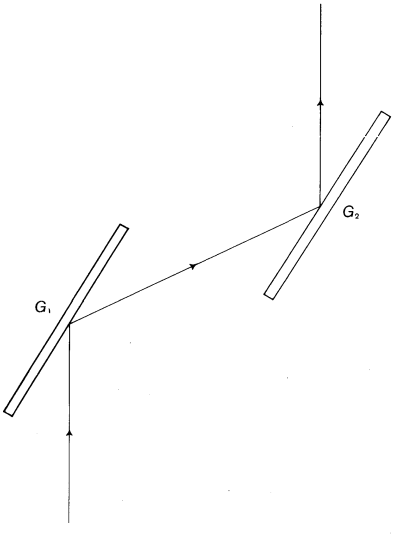


Figure 6: *Malus' experiment.* Successive reflections at two unsilvered mirror surfaces, G_1 and G_2 (see text).

The above experiments do not show whether or not the reflected light has its displacement in the plane of reflection or perpendicular to it. It is a matter of choice whether the reflected light is said to be polarized in or perpendicular to the plane of reflection. Some controversy (and some

difference of nomenclature) that formerly led to confusion was removed by the electromagnetic theory (see below). In this theory light is represented by two vectors (quantities that can be represented graphically by arrows that point in the field directions), a magnetic vector in the plane of reflection and an electric vector perpendicular to it. Confusion is avoided by specifying the plane of the electric vector instead of speaking of the plane of polarization.

The azimuthal property of reflected light at the surface of any medium—glass, plastic, a liquid—is most strongly manifested when the angle of incidence is so chosen that its tangent is equal to the index of refraction; that is, it satisfies Brewster's law (after Sir David Brewster, a British physicist), which states that, at the polarizing angle, the incident and refracted beams make an angle of 90° with one another: $\tan \theta_i = n$, in which θ_i is the angle of incidence, called the polarizing angle, and n is the index of refraction of the medium. Nevertheless, there is some azimuthal difference after reflection at any angle except $\theta_i = 0$ or $\theta_i = 90°$. Other ways of producing polarized light are described in a later section.

Optical activity

It is found that the plane of polarization of a beam of polarized light is rotated when the beam is passed through certain mediums (especially sugar solutions). These mediums are said to be optically active. Most mediums do not normally rotate the plane of polarization, but do so when there is a magnetic field in the direction of propagation (the Faraday effect).

The wave equation. The expression for a plane wave, given in equation (1) and showing the relationship between displacement (ξ), the time span (t), and distance (z) along the wave, may be differentiated twice with respect to t and z; that is, to find out how the displacement changes with position and time. This operation yields the partial differential equation:

$$\frac{\partial^2 \xi}{\partial z^2} = \frac{1}{b^2} \cdot \frac{\partial^2 \xi}{\partial t^2}, \tag{7a}$$

in which b is the phase velocity. For a three-dimensional wave the analogous expression is

$$\frac{\partial^2 \xi}{\partial z^2} + \frac{\partial^2 \xi}{\partial y^2} + \frac{\partial^2 \xi}{\partial x^2} = \frac{1}{b^2} \cdot \frac{\partial^2 \xi}{\partial t^2}. \tag{7b}$$

There are many solutions of this basic equation. Some correspond to the sinusoidal plane waves, which have already been considered. Others correspond to groups of plane waves that differ slightly either in direction, or wavelength, or both. Yet another solution of the general wave equation is:

$$\xi = \frac{A}{r} \cos 2\pi(\nu_i t - \nu_3 r), \tag{8}$$

in which r is the magnitude of a radius vector drawn from the origin and A is a constant. This represents spherical waves.

Energy of a beam of light. The energy in a small volume (dV), through which plane waves are passing, is proportional to the product of the square of the amplitude (A), or its energy per unit volume (W), and the small volume; that is, $A^2 dV = W dV$. The rate of transport of energy across a surface normal to the direction of propagation is proportional to the product of the energy per unit volume, the phase velocity, and a small area (dS) normal to the direction of propagation, or $WbdS$. For spherical waves, the rate of transport is inversely proportional to r^2, *i.e.*, $(A/r^2)dS$. Because the area of a sphere is $4\pi r^2$ in which r is its radius, this equation implies that the total energy crossing any sphere surrounding a point source is independent of the radius. Thus, inverse-square law for the intensity of radiation at a distance r from a point source is in accord with the law of conservation of energy—the total energy of a wave remains the same even though the wave is spread over a greater area.

Doppler–Fizeau effect. The length of a wave train emitted in one second by a stationary light source is equal to the velocity of light (c) times one second, which in itself is equal to the product of its frequency (ν_i) times its wavelength (λ)—*i.e.*, $c = \nu_i \lambda$. If the source moves away from the observer with a velocity (v) that is small compared with the velocity of light, then the length of the wave train increases so as to be numerically equal to the sum of the two velocities ($c + v$) and the number of waves remains the same. The wavelength λ increases to λ' by a factor $(c + v)/c$; that is $\lambda' = (1 + v/c)\lambda$. This change was discovered by an Austrian physicist, Christian Doppler, in the 19th century in relation to sound waves and subsequently applied to light waves by Fizeau. It is called the Doppler–Fizeau effect. The Doppler–Fizeau effect is easily observed when part of the light from a gas laser is allowed to be scattered by a moving body and mixed with a little unscattered light. It is known from the study of sound waves that the beat frequency is equal to the difference between the frequencies of the two waves that are mixed. Although the frequency of light waves is extremely high (more than 10^{14} per second), the beat frequency may be a megahertz (10^6 cycles per second), which is easily detected by radio amplifiers, or even a few hundred cycles per second, which the human ear can detect. Thus, just as interference fringes provide a periodic phenomenon in which two light waves combine to produce fringes of low spatial frequency, so the Doppler–Fizeau effect produces beats the temporal frequency of which is a known, but very small, fraction of the temporal frequency of the light waves. In this way the periodicity of light in both space and time is exhibited and measured.

Beat frequency

LIGHT SPECTRUM

It was seen, in the preceding section, that white light can be dispersed into a spectrum by refraction, by diffraction, or by interference. Newton showed that if a suitably oriented slit is used to select a small region of the spectrum, the light that passes through the slit is much more homogeneous than the original white light, and he was unable to observe any further dispersion when passing this light through a second prism. Delicate methods of interferometry nevertheless show that this light is never entirely of one wavelength, however fine the slit, but covers a range ($\Delta\lambda$) of wavelengths. The ratio of the wavelength divided by this range, which measures the purity of the spectrum, may be a few thousand for a spectrum formed by a prism and up to a million for a spectrum formed by a large diffraction grating. It is never infinite, as it would be if $\Delta\lambda$ were zero.

The spectrum of a hot body such as the solar photosphere is continuous (every wavelength is represented); but a German physicist, Joseph von Fraunhofer, early in the 19th century observed that the solar spectrum contains numerous dark lines appearing at certain wavelengths, which are attributed to wavelengths originally emitted by inner layers of the Sun but then absorbed by various elements (in gaseous form) in the cooler outer layers (see Plate). Emission spectra produced by electric sparks and arcs contain sharp bright lines which are characteristic of the elements in the electrodes.

In monochromatic light, colour and wavelength are associated. Nevertheless, as Newton said, "the rays, to speak properly, are not coloured." Colour is a sensation in the human mind. Light of one wavelength can stimulate the visual system so that a certain colour sensation (*e.g.*, red) is produced. The way in which the visual system analyzes colour is entirely different from the way in which physical instruments form a spectrum (see SENSORY RECEPTION).

There are a number of ways in which spectra are produced in nature. The rainbow is the most striking of these. The primary rainbow is formed by reflection and refraction of light in raindrops. The rays emerging from the drops are spread out, but for any given wavelength there is a minimum angle of deviation and there is a concentration of energy at this angle. For green light the minimum angle of deviation is about 138° and an observer with his back to the Sun sees the bow at an angle of 42° to the direction of the Sun's rays. Because of the dispersion of water, the angles for different wavelengths are not exactly the same, and the red is seen on the outside and blue on the inside of the bow. A weaker rainbow is formed by

Primary and secondary rainbows

rays that have been twice reflected. In this the colours are reversed. Still weaker supernumerary bows are caused by diffraction in droplets. A rainbow may be regarded as a spectrum of the Sun, but the purity is low.

VELOCITY OF LIGHT

The accepted value of the velocity of light (c) in vacuum is 299,792.5 ± 0.4 kilometres per second, or 186,282.4 ± 0.2 miles per second (see Table 1). The velocity is the same for all wavelengths over the whole range of the electromagnetic spectrum from radio waves to gamma rays. Methods of measurement are of three types: (1) measurement of the time (T) in which a group of waves covers a known distance (l); (2) measurement of the frequency (ν_f) and wavelength (λ) of monochromatic waves; and (3)

Table 1: The Constant c
(in kilometres per second)

	year	value
Derived from measurements of the velocity of light		
Michelson	1927	299,796 ± 4
Michelson, Pearson and Pease	1935	299,774 ± 11
Value accepted in 1941	1941	299,773 ± 3
Bergstrand	1951	299,793.1 ± 0.2
Bergstrand (mean value)	1957	299,792.9 ± 0.2
Derived from measurements on radio waves		
Essen (10^4 MHz)	1950	299,792.5 ± 1
Froome (2.4 and 7.5 × 10^4 MHz)	1951–8	299,792.5 ± 0.1
Value adopted by 12th General Assembly of the Radio-Scientific Union	1957	299,792.5 ± 0.4
Derived from electrical measurements		
Rosa and Dorsey (ratio of units)	1907	299,788 ± 30
Mercier (Lecher wires)	1923	299,795 ± 30

indirect methods, such as measurement of the change of frequency or wavelength (Doppler–Fizeau effect) when a beam of light is reflected from a mirror moving with a known velocity.

Methods of type (3) have, so far, given an accuracy of only a few percent. Methods of type (2) cannot be used for light waves because the frequency is about 1.5 × 10^{14} hertz and is too high to be measured directly. The remainder of this section will review measurements of the velocity of light by methods of type (1) and compare the results of the best measurements with the results obtained for radio waves by methods (1) and (2).

Astronomical measurements. In 1676 Rømer made careful measurements of the times at which satellites of Jupiter were eclipsed by the planet. The times observed did not agree with those calculated on the assumptions of a constant period of rotation and of instantaneous transmission of light. Starting at a time when the Earth was at its nearest to Jupiter, the apparent period increased and the eclipses became increasingly later than the calculated times as the Earth receded from Jupiter. Similarly, the period shortened when the Earth was moving toward Jupiter. The observed times were consistent with a finite velocity of light such that the time for it to transverse the Earth's orbit is about 1,000 seconds. Taken with modern values of the size of the Earth's orbit, the derived value of the velocity is 298,000 kilometres per second. It is remarkable that this first measurement was even of the correct order; the most important conclusion was that the velocity of light is finite. An English astronomer, James Bradley (died 1762), obtained a similar value by the so-called aberration method, based on the apparent motion of stars as the Earth travels in its orbit about the Sun.

Early terrestrial experiments. In terrestrial experiments by method (1), the beam of light is periodically marked either by interrupting it at regular intervals or by modulating it (alternately increasing and decreasing its intensity). The marked beam is transmitted to a distant mirror and the return beam passes through the apparatus that interrupts or modulates the outgoing beam and then to a detector. If the time required for transmission to the distant mirror and return is $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, . . ., times the period of the interrupter (or modulator), then the amount that reaches the detector is small. It is usual to adjust either the path length or the period of the interrupter or modulator until the light registered by the detector is a minimum. In the earlier experiments a mechanical chopper was used as interrupter and the eye was the detector. The later experimenters used electronic modulators and photoelectric detectors.

The apparatus used by Fizeau in 1849 is shown in Figure 7, in which M_1 is a partially reflecting mirror and M_2 is a fully silvered mirror. As the speed of the wheel (which has 720 teeth) was increased from zero it was found that the light was first eclipsed by a tooth when the speed was about 12.6 revolutions per second—*i.e.*, when the time to make the round trip was 560 microseconds (0.00056 second), the length of the double path being 17.3 kilometres (about 10 miles). The chief error in the measurement lay

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

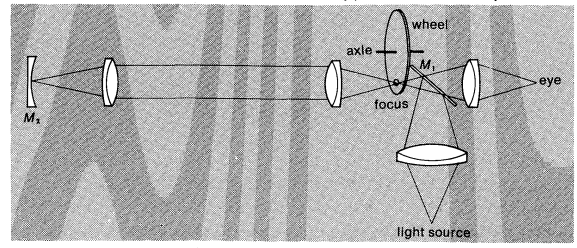


Figure 7: Fizeau's method for measuring the velocity of light.

in the difficulty of determining the exact speeds at which the light received by the eye at E was at a minimum. Essentially the same method was used by others between 1874 and 1903. The accuracy gradually improved, and it was shown that the velocity is between 299,000 and 301,000 kilometres per second.

In 1834 Sir Charles Wheatstone of England suggested a method incorporating a rotating mirror for interrupting the light that was later developed by Arago (1838) and Foucault (1850). It was considerably improved by Michelson, who made measurements from 1879 to 1935. Use of rotating mirror

Michelson's measurements. Figure 8 shows the arrangement used in 1927. The mirror M_3 is a little above the plane of the diagram, and M_3' is a little below. Light from the source S passes to one face of the octagonal mirror M_1 and then to M_2, M_3, and M_4. From M_4 it goes to the mirror M_5 at a distance of about 35 kilometres (about 22 miles). It returns via M_6, M_4, M_3', and M' to the octagon. An image of S is seen in an eyepiece at E. The octagonal mirror rotated at 528 revolutions per second. It turned through approximately one-eighth of a revolution during

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

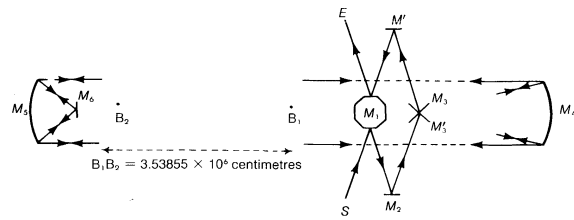


Figure 8: Michelson's Mount Wilson experiment, 1927.

the transit of the light. If the rotation were exactly one-eighth of a revolution the image would be undisplaced from the position it had when the mirrors were stationary. In some of Michelson's experiments the speed of rotation was slowly changed until this condition was obtained. In others the speed and distance were fixed, and a small displacement of the image was measured.

It is difficult to estimate the accuracy of Michelson's 1927 and 1935 experiments, and it is no longer important to do so in view of the more accurate measurements made since 1945. His most important contribution to the measurement of the velocity was the proof that the velocity agreed with Maxwell's prediction to better than one part in a thousand. This gave confidence to those working on applications of the electromagnetic theory.

The electro-optical shutter. This device, based on the Kerr effect (see below), makes it possible to modulate a beam of light at frequencies more than 10,000 times the highest frequency of interruption used by Michelson and obtain values in reasonably good agreement with each other and with Michelson's later work. This method was greatly improved by E. Bergstrand in Sweden, who reduced the random errors by a factor of more than 30 and obtained a value for the velocity of light of 299,793.1 kilometres per second.

Radio-frequency measurements. The velocity of electromagnetic waves of radio frequency in vacuum has been measured by several methods. An English physicist, Louis Essen, measured (1950) the resonance frequency of a cavity resonator (an electromagnetic device) whose dimensions were also determined with high accuracy. Keith Davy Froome, a physicist in England, measured (1952 and 1958) the wavelength in air, corresponding to a known frequency, using a microwave interferometer. The results of these and other measurements are in agreement with those of Bergstrand to within a few parts per million. The velocity of radio waves in vacuum is thus equal, within this accuracy, to the velocity of light. The velocity of gamma rays is also the same, within the much lower accuracy of this last measurement. Table 1 summarizes the measurements of the velocity constant (c) and shows that there is now satisfactory agreement between results obtained over a wide range of conditions.

Recognition of the velocity of light as a fundamental constant

Since the publication of the special theory of relativity (1905), the constant c has been recognized as one of the fundamental constants of modern physics. For this reason, attempts will undoubtedly be made to measure it with even greater precision. The use of lasers may help, but a major improvement will require the establishment of better standards of length and time than those now available.

Velocity in material mediums. All measurements of the velocity of light involve interruption or modulation of a beam of light so as to form groups of waves and the velocity measured is the group velocity. The difference in magnitude between the wave velocity and the group velocity of light in air is only about one part in 50,000, but in most glasses and in some liquids it is much larger. Michelson obtained 1.758 for the ratio of the velocity in air to the velocity in carbon disulfide. The inverse ratio of their indices of refraction is 1.64 and the value calculated from this for the ratio of group velocities is 1.745 for wavelength 580 nanometres, close to Michelson's observations. Bergstrand found that the ratio of the velocity *in vacuo* to the velocity in a certain glass was 1.550 ± 0.003. The refractive index of the glass was 1.519, but the ratio of c to the group velocity was 1.547. The experimental results thus agree with those calculated on the assumption that the measured velocity is the group velocity.

Interference and diffraction phenomena

INTERFERENCE

Quasi-monochromatic waves. A perfectly monochromatic wave, represented by equation (1), has constant amplitude and is not limited in space or in time. Sources of light (other than lasers) emit waves the amplitude of which varies with time. For example, a single undisturbed atom emits a damped wave (Figure 14A). Under favourable conditions the damping is so weak that 10^7 waves are emitted before the amplitude has fallen to half its initial value and the change of amplitude is not significant over a distance of several thousand wavelengths. Wave trains of this type are said to be quasi-monochromatic. Superposition of these waves gives interference when the path difference is not too large.

Photometric summation. Figure 5, curve A, shows the way in which the intensity of light varies from place to place when two monochromatic or quasi-monochromatic waves overlap. The intensity at a point in the region where the waves overlap may be expressed as the sum of two terms: (1) the sum of the intensities of each wave acting alone ($2I_0$ if each alone would give intensity I_0); (2) a term representing the interference of the waves. The second term varies from point to point along the direction of propagation between the values $-2I_0$ and $+2I_0$. Thus the total intensity varies from $4I_0$ (*i.e.*, twice the intensity sum) to zero. Now, when a large number of waves from different sources cross a certain space, the fringes caused by the interference of each pair of waves have their maxima in different places and the overall result is that, at any point, the interference terms are positive nearly as often as they are negative and their total sum is nearly zero. In this case the resultant intensity at any point caused by a number of sources is just equal to the sum of the intensities (at that point) of each source acting alone. This is the law of photometric summation and is used by illumination engineers in calculating the illumination on a surface that receives light from various sources. Interference fringes are obtained only when experimental conditions are such that the interference fringes caused by light emitted from different atoms all have their maxima in the same places (or near to the same places). The interference term then becomes a significant fraction of the summation term. This may be achieved either (1) by using two secondary sources (such as the two slits used in Young's interference experiment), which are both derived from the same primary source, or (2) by using a laser in which the source atoms are stimulated in such a way that the phase relations between them remain constant during the period of observation.

Visibility of interference fringes. The distribution of intensity in interference fringes, shown in Figure 5, curve A, represents an ideal that is closely approached in some experiments, but generally the distribution is such that the fluctuations that constitute the fringes are superposed upon a nearly uniform background. Michelson defined the visibility of fringes as the difference between the maximum and minimum intensity of a fringe divided by their sum, or

$$V = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}, \qquad (9)$$

in which V is the visibility, I_{max} is the maximum intensity, and I_{min} is the minimum. The fringe visibility is thus always between zero and one. When the minimum intensity is zero, the visibility equals one. Obviously, fringes for which V is less than one are obtained when waves of unequal amplitude are superposed because the weaker cannot, at any point, annul the stronger. It is also found, however, that even when the intensities are equal, the visibility is usually less than one (as shown in Figure 5, curve B). Further consideration of Young's slit experiment leads to recognition of two conditions that must be fulfilled to obtain fringes of high visibility. These relate to their geometrical condition and spectral range.

Two conditions for high visibility

Geometrical conditions. In the arrangement shown in Figure 4, the centre of the fringe system is at a position O on the screen, on the straight line from the source slit S to a position midway between P_1 and P_2 (the slits are all assumed to be extremely narrow). If slit S is moved to S' then the centre moves to O'. If, instead of moving the slit S to this new position, it is gradually widened, the intensity at any point Q is found by adding the intensities of waves emitted by atoms behind different parts of the slit. Because the fringes on plane $O'OQ$ produced by light from different parts of slit S are not in register, there cannot be zero intensity at any point in the pattern. As slit S is widened the fringes gradually become blurred—*i.e.*, the visibility falls from unity to zero. If $l_1 = l_2$, no fringes are seen when the width of slit S is about equal to the distance (d_f) between successive fringes.

Spectral range. In the case in which the slit S is extremely narrow and the light is not all of exactly one wavelength, the path difference and the phase difference will be zero at the centre of the fringe system for all wavelengths, so that for all wavelengths there is maximum intensity at the centre O of the fringe system. Because the separation of the fringes is proportional to the wavelength, the fringes produced by light of different wavelengths gradually go out of register as the path difference is increased. With white light, one clear fringe is seen in the centre. A few coloured fringes are seen on either side because the

eye makes a certain degree of separation of the colours. If a filter is used to restrict the light to a band of say 50 nanometres wide, then about ten fringes may be seen on either side, and this number is increased if the wavelength range is further restricted.

These two causes of reduced visibility differ in that the geometrical condition affects all parts of the fringe system equally and the effect of the spectral range increases as the path difference increases. In discussing these phenomena it has been assumed, in accordance with the preceding discussion, that the intensity of light from different atoms obeys the law of photometric summation. It is also assumed that the photometric law applies when different wavelengths are superposed.

Coherence. When two beams of light can interact so as to produce interference fringes the visibility of which is unity, they are said to be perfectly coherent. When their interaction produces no fringes (but only photometric summation) they are said to be noncoherent or incoherent. An elaborate mathematical theory of coherence recognizes that coherence and noncoherence are extreme cases—between them lies "partial coherence." Zernike, who contributed a great deal to the development of the subject, defined the degree of coherence γ_{12} of two sources as equal to the visibility of the fringes obtained in the most favourable circumstances using light from these sources. It has been shown that the visibility of the fringes obtained in Young's experiment depends on the width of the slit S_1, and the following mathematical relation has been derived:

Partial coherence

$$\gamma_{12} = \frac{\sin(2\pi ad/l_1)}{2\pi ad/l_1}, \qquad (10)$$

in which d is the width of the slit S_1 (Figure 4). If d is gradually increased from zero, γ_{12} falls from one (for d equal to zero) to zero for $d = l_1a/\lambda$, a value equal to the separation of the fringes when $l_1 = l_2$. When the width d is further increased, fringes are again seen but they are of low visibility and are reversed (*i.e.*, there is now a dark fringe in the centre).

For the case of a slit source being inaccessible for measurement, its angular width (d/l_1) can be determined by measuring the visibility of the fringes while a, the separation between P_1 and P_2, is varied. Michelson used this method to obtain the angular diameter d of a star (serving as the slit source) from measurements of the visibility of interference fringes formed in the focal plane of a telescope that receives light from two small mirrors mounted in front of the telescope's objective, separated from each other by a distance a.

Coherence length

The concept of coherence that has been applied to light from two pinholes may be extended to a beam of light considered as a whole. A roughly parallel beam of light is incident on a thin sheet of metal normal to the direction of propagation. Then two pinholes may be made in the sheet at A and B (Figure 9) and the visibility of the resulting fringes measured so as to obtain the mutual coherence γ_{AB}. If A and B are initially coincident and are slowly separated then γ_{AB} falls gradually from one to zero. It is possible to define a region of coherence around any point A such that if point B lies within this region the coherence is good ($\gamma_{AB} > 0.7$). Similarly, by devices such as that described in the next section it is possible to measure the mutual coherence between A and a point A' that is, as it were, downstream from A and to define a "coherence length" l_c such that coherence is good when $AA' < l_c$. When, (1) the region of coherence extends across the whole beam of light, and (2) the coherence length is large, the beam is said to be highly coherent because the mutual coherence between any two points such as B and A' is high. What qualifies as a "large coherence length" depends on the type of source and the conditions of the experiment; ten centimetres is a large coherence length for the kind of source considered in the next section, but a well-stabilized gas laser may give a beam with a coherence length of many metres.

In the wave equation for light already cited, the displacement and phase angle, represented by the variables ξ and ϕ, were used to specify a wave motion, but, for light, these

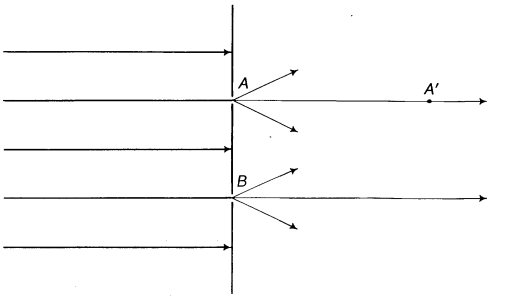


Figure 9: Two pinholes in an opaque sheet to illustrate mutual coherence between points A, A', and B.

quantities are not observable nor can they be inferred from any observations—because of the high frequency of the wave motion. The coherence γ_{12} and the phase difference, however, are observable quantities that characterize sources and beams of light. This makes them important both in theory and in practice.

Two-beam interference. In the Michelson interferometer, shown in Figure 10, the incident wave W is divided at the beam splitter BS so that part of the light is transmitted and part is reflected. After reflection at M_1 and M_2 the two parts form the wave fronts W_1 and W_2. These are copies of W, and, because corresponding points are superposed, coherence is obtained even with an extended source. The light from source S, selected by the filter at FF', is quasi-monochromatic. The plane R represents the image of M_2 that would be seen by reflection in BS. The phase differences between W_1 and W_2 are the same as if W_2 had been reflected from R, which is called the reference plane. M_1 may be traversed normal to itself and may also be tilted with respect to the reference plane. A compensating glass plate C having the same thickness as BS is used so that both wave fronts will pass through a total of three thicknesses of glass.

Fringes will be formed when M_1 is adjusted to be exactly parallel to R and separated from R by a small distance e. For a hollow cone of rays, each ray will be incident on M_1 and on R at an angle θ. After passing through the instrument on their return trip, these rays will be focussed into a circular ring in focal plane FF' of the lens L. At each point on this ring two waves will be superposed and their path difference will be $2e \cos\theta$. Bright rings are obtained for values of θ such that $2e \cos\theta = p\lambda$, in which p is an

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

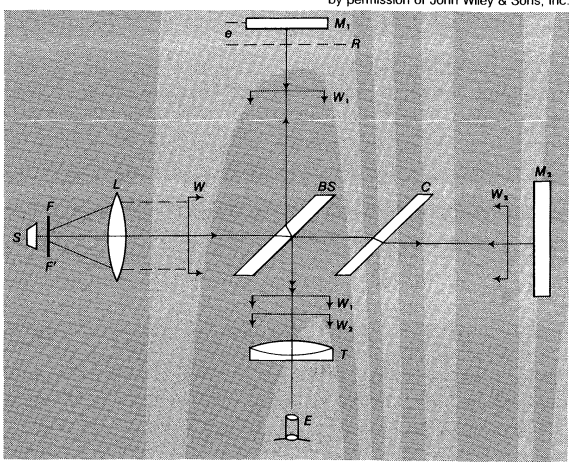


Figure 10: The Michelson interferometer.

integer. The appearance of these fringes is similar to that of Newton's rings (see Plate).

These fringes are known as fringes of equal inclination because any one ring corresponds to a set of rays that all have the same inclination, θ, to the mirror M_1. They are conveniently observed by focussing an eyepiece E on the plane F. Because the lens T and the eyepiece E constitute a telescope focussed for an infinite distance, the fringes are said to be localized at infinity.

The apparatus may also be adjusted so that the mirror M_1 is inclined to R, the image plane of mirror M_2, and nearly in coincidence with it. The incident light is rendered nearly parallel and normal to plane R. If the telescope is removed, straight line fringes can be seen by an observer who focusses his eye on the region between M_1 and R. A bright fringe is the locus of points for which $2l_p = p\lambda$. These fringes are called fringes of equal thickness.

Fringes of equal thickness may be formed by reflection at the two glass surfaces bounding an air film between two glass plates (Figure 11). Strictly speaking, this arrangement does not give two-beam interference because multiple reflections occur, as shown in the figure. Only the two beams A and B, however, need be considered for present purposes. Beams like B' and C', caused by multiple reflections, are weak, and unless the glass plates are fairly thin and of high optical quality, fringes formed by beams reflected from the outer surfaces of the glass plates are close together and are of poor visibility. If the arrangement is such that one of the plates is truly planar and that

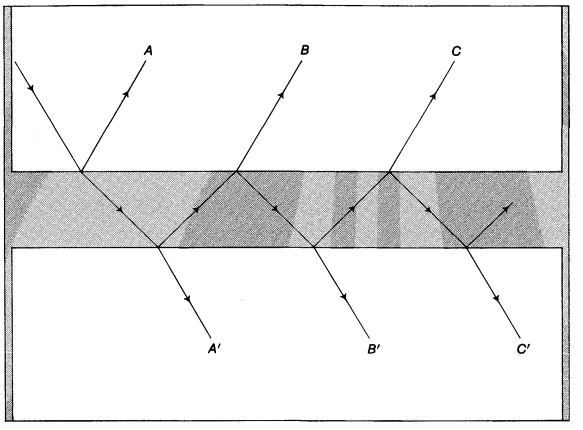


Figure 11: Interference in a thin film of air between two pieces of glass. The rays A, B, C, etc. interfere with each other, as do rays A', B', C', etc.

the other is spherical, as is the case for a convex lens lying on a glass plate, the resulting fringes of equal thickness are circles centred at the point of contact. They are known as Newton's rings (see Plate). In this situation, in which one surface is plane and the other is not, the fringes form a contour map of the nonplanar surface. They are then called contour fringes. This is a useful method for testing the flatness of a surface and determining the location of irregularities.

Newton's rings

Multiple-beam interference. If the two inner surfaces of the plates shown in Figure 11 are coated so as to make them reflect 80 percent or more of the incident light, then the resulting interference pattern will be caused by the superposition of many beams. Figure 12 shows an arrangement for producing the fringes of constant inclination by multiple beam interference. The amplitudes of successive beams are proportional to r, r^2, r^3, etc. (r is the ratio of the intensity of the reflected light to that of the incident light for one reflection). The phase differences are ε, 2ε, 3ε, etc., in which $\varepsilon = (4\pi e \cos\theta)/\lambda$. These fringes are much sharper than those obtained with two-beam interference (see Plate).

With a large number of beams the intensity is extremely high when they are all in phase ($\varepsilon = 0$), but, even when the phase difference between any two successive beams (*e.g.*, the first and the second) is quite small, the phase difference between the first and say the thirtieth beam is so large that the later beams in the series are in opposition to the earlier beams. Thus the intensity is relatively small except when the value of ε is close to one of the values $2p\pi$ (in which p is an integer). Multiple-beam fringes of constant inclination were used by Charles Fabry and Alfred Pérot in France for resolution of spectral lines having only small differences of wavelength. Multiple-beam fringes of constant thickness have been used by an English physicist, Samuel Tolansky, to detect surface irregularities down to less than a nanometre.

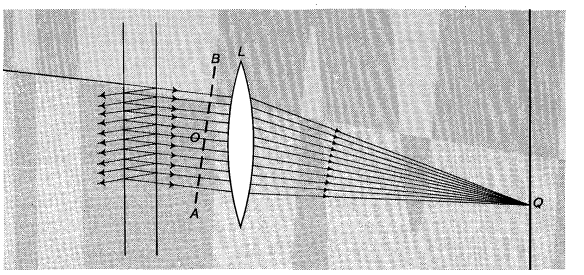


Figure 12: Multiple beam interference. Lens L concentrates all beams at focus Q with same phase differences they had while crossing a plane AB normal to OQ.

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

Wave groups. If two pendulums that have frequencies ν_t per minute and $(\nu_t + 1)$ per minute are started together, they will gradually go out of step; after half a minute they will be moving in opposite directions and after a minute they will be together again. Over a long time they will move together once every minute. In a similar way, when two waves of slightly different frequency are moving in the same direction, they are sometimes in phase and sometimes out of phase so that the resultant is sometimes large and sometimes small, as shown in Figure 13. Two waves may be considered for which the spatial frequencies are ν_s and $(\nu_s + \Delta\nu_s)$ and temporal frequencies ν_t and $(\nu_t + \Delta\nu_t)$. The fluctuation represented by the envelope (dotted line in Figure 13) is called the beat wave. It has a temporal frequency equal to the difference ($\Delta\nu_t$) of the temporal frequencies of the constituent waves and a spatial frequency $\Delta\nu_s$) equal to the difference of the spatial frequencies. It is therefore propagated with a velocity $U = \Delta\nu_t/\Delta\nu_s$. Many physical problems involve groups of waves that include a range of frequencies. It is found that, even in a dispersive medium, a group is propagated over a considerable distance as a recognizable unit. The velocity of this recognizable group is $U = d\nu_t/d\nu_s$.

Velocity of groups of waves

There is a certain kind of wave group for which the variation of the displacement with distance (z) along the path of propagation may be represented by the expression

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

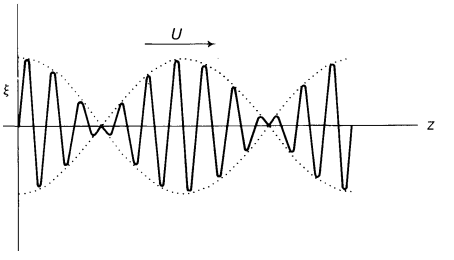


Figure 13: A simple beat wave of amplitude ξ moving with velocity U (see text).

$h(z) \cos \nu_0 z$, in which $h(z)$ is a function that varies with z much more slowly than $\cos \nu_0 z$—*e.g.*, in Figure 13, $h(z)$ would be the function represented by the dotted line, and ν_0 is the spatial frequency of the individual waves represented by the full line. These waves are called modulated waves. If $h(z)$ varies extremely slowly with z, the modulated wave is quasi-monochromatic in the sense described above; *i.e.*, it departs little (over distances long enough to contain many wavelengths) from a monochromatic wave. A modulated wave is completely described when $h(z)$ and ν_0 are known. It is also completely described when the amplitudes and phases of the various waves that make up the group are known. These are given by a function a dependent on the frequency ν_s, $a(\nu_s)$. Because $h(z)$ and $a(\nu)$ both describe the same wave group, there must be a relation between them. A mathematical theorem of a French mathematician, Jean-Baptiste-Joseph Fourier, gives this relation, making it possible to calculate either $h(z)$ or $a(\nu)$ when the other is known. The average density at z is equal to $W(z)$, which is proportional to $(h[z])^2$. The energy per unit frequency range near ν_s is $G(\nu_s)$, which is proportional to $(a[\nu_s])^2$.

When $h(z)$ varies very slowly with z, $a(\nu)$ is large for a range of ν close to ν_0 and falls rapidly to near zero outside this range, as shown in Figure 14B. If this range in which $G(\nu_s)$ is large is ν_R, and if l_R represents a range of z over which $h(z)$ varies very little (as shown in the Figure), then it is found that ν_R and l_R are inversely proportional to one another and that their product is of the order of magnitude of unity. This represents the fact that the longer the wave train the more closely its properties agree with those of the ideal monochromatic wave, which is infinitely long and has a precisely defined frequency.

Undisturbed atoms emit exponentially damped waves

From (A) R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

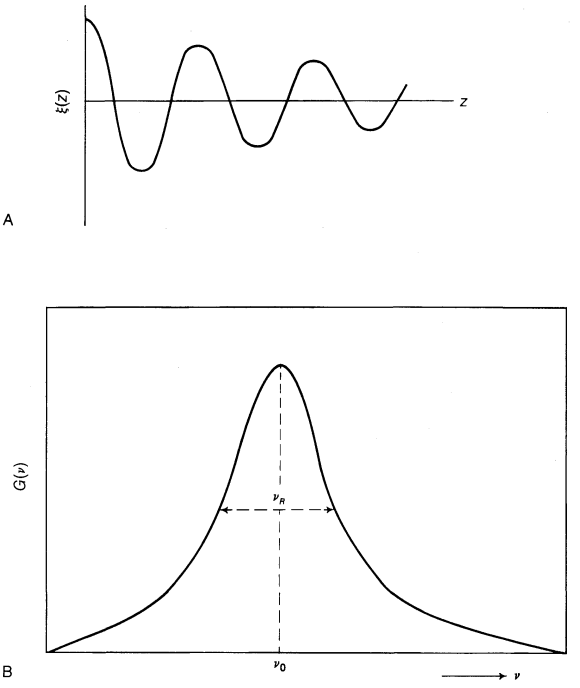


Figure 14: *Damped waves.*
(A) Amplitude, $\xi(z)$, as a function of distance, z. (B) Energy, $G(\nu)$, as a function of frequency, ν. The figure shows strong damping. For light emitted by free atoms, l_R would encompass 10^7 waves or more and ν_R would be correspondingly smaller (see text).

the length of which is usually 10^7 waves or more so that ν_R is a small fraction of ν_0. Collisions increase the damping by a factor that is proportional to the pressure. The observed radiation is also modified by the Doppler–Fizeau effect, because the atoms that emit the light do not all have the same velocity. This increases the range $\Delta\nu_R$. Even when the effects of collision damping and Doppler–Fizeau effect are combined, the value of l_R for the wave trains emitted in low pressure electrical discharges is still about 10^5 wavelengths, and most of the energy is confined within a frequency range of order $10^{-5}\ \nu_0$ (corresponding to a wavelength range of less than 0.01 nanometre). These quasi-monochromatic waves are called wave groups. The light emitted by high-pressure lamps or by luminescent solids extends over a much wider range of frequency, and wave-group theory has little useful application to problems concerning non-monochromatic light from these sources.

Wave groups in a dispersive medium. In vacuum, all components of the group have the same phase velocity, and therefore the phase relations between different members of the group are constant. A group advances as a unit without any change of the modulation function $h(z)$. In a dispersive medium, the phase relations change and $h(z)$ changes as the wave train advances, but this change is slower than might be expected. Over considerable distances the group is propagated as a recognizable whole with the group velocity U. The change of $h(z)$ is small for passage through a gas and also for groups (for which ν_R is small) that represent sharp spectral lines. Thus these wave groups are propagated virtually unchanged through an optical instrument or, if they arrive from the Sun or stars, through the Earth's atmosphere.

Propagation of wave groups from stars

In Young's experiment, Figure 4, the fringes have maximum visibility at the position O, corresponding to zero path difference and zero phase difference for all wavelengths. If a thin sheet of mica is inserted in front of slit P_1, the centre (or position of maximum visibility) is displaced upward. It was at one time thought that the new centre would be found at the point corresponding to zero phase difference for the mean wavelength in the wave group—*i.e.*, the point calculated for equal times from slits P_1, and P_2 allowing for the fact that the phase velocity in mica is less than that in air. This did not agree with experimental observation. It was found that the new centre is situated at the position where the times from P_1 and P_2 are equal when the group velocity is used to calculate the time required to traverse the piece of mica. At this position the wave train from P_1 exactly overlaps the wave train from P_2. At any other position, part of each wave train cannot take part in the interference because it does not coincide with any part of the other wave train. This light that cannot interfere forms a uniform background and so reduces the visibility of the interference fringes.

If white light is used and a fairly thick piece of mica is inserted, no fringes are obtained. This is because the wave train has changed shape so much in passing through the mica that it can no longer match the wave train that has travelled through air.

Michelson, using the apparatus shown in Figure 10, studied the decrease in visibility of interference fringes as the path difference between the two wave trains is increased. The reduction in visibility as the path difference increases may be assigned either (1) to the fact that the parts of the wave trains that overlap are decreasing or (2) to the increasing difference between the positions of the bright fringes for different wavelengths in the group. As was seen above, the length of the wave trains and the range of the wavelength are inevitably linked, and so these alternatives (1) and (2) do not constitute two different theories. They are just two different ways of visualizing how wave trains (or wave groups) interfere.

DIFFRACTION

Theory of diffraction. Huygens assumed that every point on a wave front may be regarded as a source of spherical wavelets the envelope of which is the position of the wave front at a later time. Huygens was thus able to account for rectilinear propagation and for the laws of reflection and refraction. Fresnel added the hypothesis that the wavelets can interfere and this led to a theory of diffraction. Figure 15 shows how a coherent, monochromatic wave from a point source P falls on the screen

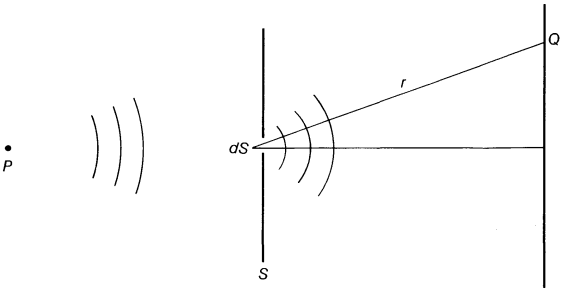


Figure 15: Principle of Fresnel's theory of diffraction (see text).

S, which is opaque except for an aperture dS. Fresnel assumed that the amplitude ($d\xi_Q$) of the wavelet at Q, originating from a small area dS, is:

$$d\xi_Q = a\frac{A}{r}f(\chi)dS, \qquad (11)$$

in which A is the amplitude of the incident wave, a is a constant, r is the distance from dS to Q, and $f(\chi)$ is a function of χ, the inclination factor; this factor was introduced by Fresnel because he believed that the effect of the element dS would be greater in the forward direction

($\chi = 0$) than in an inclined direction. The total effect at Q was obtained by superposing the wavelets from all parts of the aperture, allowing for phase differences caused by a variation of r and also for variation of the inclination factor, $f(\chi)$. Fresnel developed an ingenious method of dividing S into a series of zones of equal area and calculating the total effect as the sum of a simple series. This method applies only to circular apertures and obstacles and then only to points on the axis of symmetry, but Fresnel also developed integrals that are more generally applicable.

Fresnel zones

Fresnel predicted that there should be a bright spot at the centre of the shadow of a circular obstacle. The experimental verification of this unexpected result gave confidence in Fresnel's wave theory of diffraction.

Fraunhofer diffraction. When the source and pattern screen are sufficiently far from the slit, the phase differences corresponding to different parts dS of the slit opening vary linearly with x and y coordinates in the plane of the aperture (Figure 16). This situation is obtained when two spherical lenses L_1 and L_2 are introduced with source P at the focus of L_1 and Q in the focal plane of L_2. Spherical waves emanating at the focus of a lens are rendered plane wherever they encounter the lens. Plane waves are made spherical by a lens. They have the same radius of curvature as the focal length of the lens. The wave falling on S is a plane wave, and the total effect at Q may be regarded as caused by a plane wave leaving S. The same result is obtained if L_1 and L_2 are replaced by a single lens (situated near S) that forms an image of P at Q_0. This

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

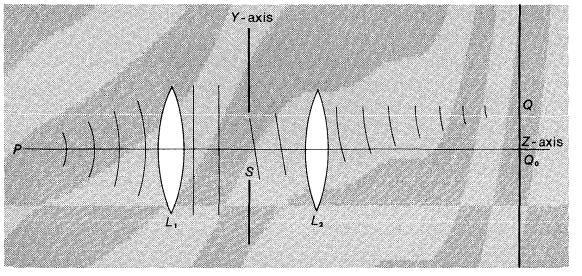


Figure 16: Arrangement for Fraunhofer (far-field) diffraction. The opening at S diffracts light from source P onto plane Q (see text).

is known as far-field diffraction, or Fraunhofer diffraction, and is thus distinguished from near-field, or Fresnel, diffraction. It should be understood, however, that there is only one physical theory of diffraction that is derived from the ideas of Huygens and Fresnel. Fraunhofer diffraction is of great practical importance especially in regard to the performance of optical instruments.

Groups of waves with different directions. When a plane wave is incident upon a slit as shown in Figure 16 (so that its width is limited), the emergent light may be represented by a group of plane waves. All the waves of this group have the same spatial frequency but differ in regard to direction of propagation. It is possible to define a range of angles (in the plane of the page) within which most of the light is found. If this range is θ_R and the width of the slit is w, then it is found that sin θ_R is inversely proportional to w; *i.e.*, the narrower the slit, the greater is the angular spread—sin θ_R is roughly equal to λ/d. Using Fourier's theorem it is possible to derive an equation that gives the amplitude and phase of the light diffracted in any direction as a function of the width of the slit. Extension of the calculation to diffraction by apertures or obstacles of any shape involves more lengthy mathematics but no new physical principle.

Angular power spectrum. The energy diffracted in any direction is proportional to the square of the corresponding amplitude. This energy expressed as a function of the angles that define the direction is called the angular power spectrum. It may be measured and is found to agree with that calculated when the width of the slit (or, more generally, the shape and size of the apertures) is known. There are many problems in which it is desired to carry out an inverse calculation—*i.e.*, to calculate the shape and size of the apertures from measurements of the angular power spectrum. Unfortunately, this is not, in general, possible because measurement of the angular power spectrum does not give the phase of the diffracted light. It is found, however, that measurement of the angular power spectrum yields a function (known as the auto-correlation function) of the size and shape of the obstacles or apertures responsible for the diffraction. In X-ray analysis, this function gives important information about symmetry. A complete picture of the crystal may often be obtained by combining calculations from the angular power spectrum with information derived from other sources.

It is found that when diffraction is due to a number of apertures (or obstacles) that are similar in size, shape, and orientation, the angular power spectrum (G) is the product of two factors, F and f, in which F (called the form factor) depends only on the properties of the individual aperture and f (called the structure factor) depends only on the arrangement or spacing of the elements. When the apertures are irregularly arranged, f is just equal to N (the number of apertures). Thus the diffraction halos produced by an irregular distribution of small similar objects have the same intensity distribution as the pattern for a single particle. This principle is used in a device called an eriometer to determine the size of blood corpuscles and may also be used to calculate the average size of the small particles that cause a halo around the Moon.

Diffraction by similar apertures

When N similar elements are arranged in a regular pattern, the structure factor may vary from zero to N^2. A diffraction grating (a plate having parallel lines engraved across its surface) with N lines is such a pattern, and, for any of the directions θ_p defined by equation (6), the light from all elements (lines) is in phase; the amplitude is N times that given by a single element, and thus the energy and the structure factors are proportional to the total number of lines squared—*i.e.*, $f = N^2$.

Limits of resolution. Diffraction spreads the light in optical images; so that if two objects are too close to each other, the gap between them cannot be distinguished. The distribution of intensity with radius in the image of a point source is shown in Figure 17. Rayleigh showed, theoretically and experimentally, that the images of two point sources are just resolved when their separation is such that the centre of the pattern due to one image falls on the first minimum of the pattern due to the other (Figure 18). This implies that a telescope with a perfect objective of diameter D can just resolve two stars whose angular separation is $1.2\ \lambda/D$. Qualitatively, this agrees with a calculation that shows that most of the energy in the diffraction pattern of an aperture of width d lies within an angular range $\pm\lambda/d$.

The angular separation of the maxima resulting from light

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

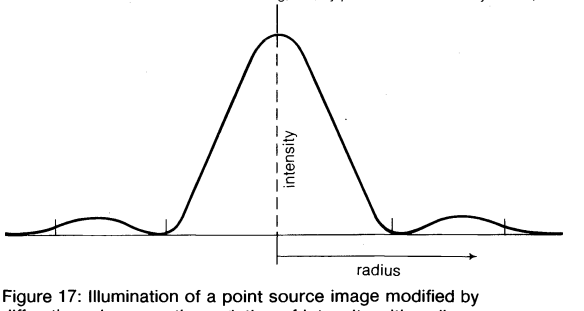


Figure 17: Illumination of a point source image modified by diffraction, shown as the variation of intensity with radius.

of two wavelengths λ and $\lambda + \Delta\lambda$ in a spectrum formed by a diffraction grating is obtained by differentiating equation (6), resulting in $\Delta\theta = p\Delta\lambda/e \cos\theta$. These maxima are just resolved if $\Delta\theta = \lambda/M$, in which M is the width of the beam diffracted by the grating. For a grating of N lines, this width is $Ne \cos\theta$, and the resolving power $R = \lambda/\Delta\lambda = pN$. A grating ten inches wide, for example, with 10^4 lines per inch and $p = 10$, has a resolving power $R = 10^6$.

The limit of resolution for a microscope depends on conditions of illumination and is at best about half a wavelength ($\lambda/2$), or about 250 nanometres for visible light.

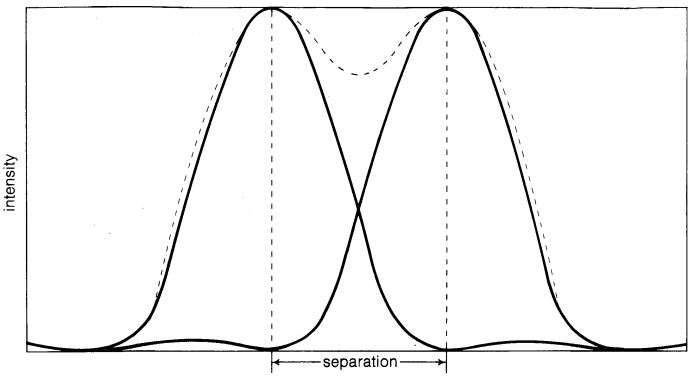


Figure 18: Overlapping images of two point sources. Full lines show how intensity varies with distance from a separate source, dashed line shows combined intensity.

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

Polarization and electromagnetic theory

POLARIZED LIGHT

Interference of polarized light

Interaction of plane-polarized beams. Fresnel and Arago, using an apparatus based on Young's experiment (Figure 4), investigated the conditions under which two beams of plane polarized light may produce interference fringes. They found that: (1) two beams polarized in mutually perpendicular planes never yield fringes; (2) two beams polarized in the same plane interfere and produce fringes, under the same conditions as two similar beams of unpolarized light, provided that they are derived from the same beam of polarized light or from the same component of a beam of unpolarized light; (3) two beams of polarized light, derived from perpendicular components of the same beam of unpolarized light and afterwards rotated into the same plane (*e.g.*, by using some device such as an optically active plate) do not interfere under any conditions.

Result (1) is to be expected because two displacements in perpendicular planes cannot annul one another, and result (2) is also easily understood. Result (3) shows that mutually perpendicular components of unpolarized light in a beam are non-coherent. Their phase difference varies in time in an irregular way. Unpolarized light has a randomness, or lack of order, as compared with polarized light (implying an entropy difference). This order (or lack of order), rather than the azimuthal property, is the most fundamental difference between polarized and unpolarized light. Perfectly monochromatic light is perfectly coherent and completely polarized.

Superposition of polarized beams. Two coherent beams of plane polarized light may be thought of as propagated in the Oz direction, one with its vector along Ox and the other with the electric vector along Oy; *i.e.*, the two vibrations are at right angles to each other as well as to the direction of propagation (Figure 19). If the beams have amplitudes a_x and a_y and phases ε_x and ε_y, then, in general, the resultant vibration (R_1, R_2, and R_3) may be represented in magnitude and polarization by a vector, or arrow, the tail of which touches the axis of propagation Oz while the point moves round the ellipse (Figure 19). It goes round once when the phase angle φ (see equation [1]) changes by 2π—*i.e.*, at any given place when t changes by y or for any one time when z changes by λ. The beam is said to be elliptically polarized. If the phase difference is $\pi/2$, then the axes of the ellipse are equal to a_x and a_y and are along O_x and O_y.

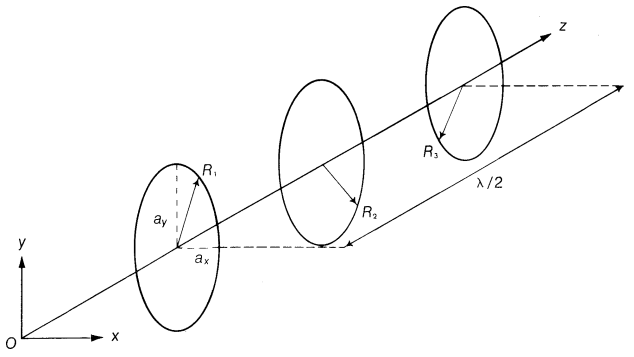


Figure 19: Progression of elliptically polarized wave (see text).

Elliptically polarized light may be regarded as the most general type of polarized light. If the amplitudes of the two waves are equal, $a_x = a_y$, and the phase difference is still $\pi/2$, the ellipse becomes a circle and the light is said to be circularly polarized. If the phase difference ε_{xy} is not equal to $\pi/2$, the resultant is still elliptically polarized light, but the axes of the ellipse no longer coincide with the axes of coordinates. If the phase difference $\varepsilon_{xy} = 0$ or π, the ellipse shrinks to a straight line and the light is said to be plane-polarized. If the representative vector, when viewed by an observer who receives the light, rotates in a clockwise direction, the light is said to be right-handed (or positive) elliptically polarized light. The opposite sense of rotation corresponds to left-handed (or negative) elliptically polarized light.

In the above analysis, elliptically polarized light is regarded as the resultant of two beams plane-polarized in perpendicular planes. Conversely, it is possible to regard plane-polarized light as the resultant of two beams of elliptically (or circularly) polarized light of the same wavelength, provided that the ellipses are similar in orientation and eccentricity, but one beam is right-handed and the other left-handed.

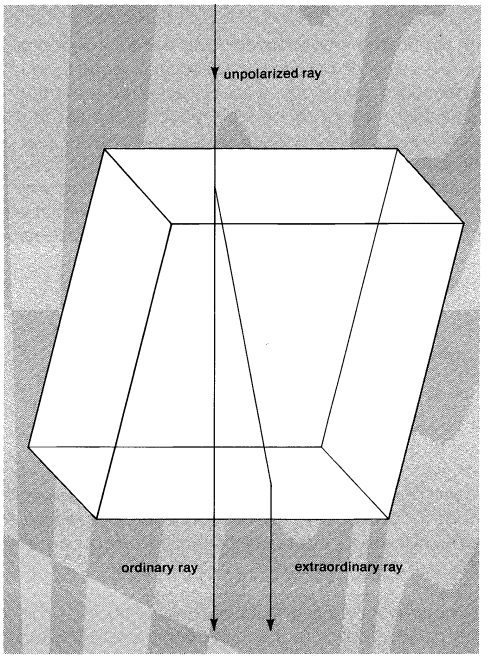


Figure 20: Double refraction showing two rays emerging when a single light ray strikes a calcite crystal at right angles to one face (see text).

Double refraction. In the 17th century Bartholin showed that a ray of unpolarized light incident on a plate of calcite, unlike glass or water, is split into two rays, as shown in Figure 20. One ray, called the ordinary ray, is in the plane containing the incident ray and the normal to the surface. If the angle of incidence is varied, this ray is found to obey Snell's law of sines, equation (3). The other ray, called the extraordinary ray, is not in general coplanar with the incident ray and the normal; also, for it, the ratio of sines is not constant. The fact that Snell's law is not obeyed in certain directions implies that the velocity of light in such a medium, called anisotropic, depends on the direction of travel in it. The two rays are polarized in mutually perpendicular planes. This is known as double refraction, or birefringence.

Birefringence

In order to apply Huygens' method of constructing wave fronts (see above *Theory of diffraction*), it is necessary to assume that, in an anisotropic medium, the wave surface

from a point source consists of two sheets, or surfaces (Figure 21). The observation that one ray obeys both laws of refraction implies that one sheet must be a sphere—like the wave surface in an isotropic medium. Huygens assumed that the other sheet is an ellipsoid of revolution that touches the sphere either internally (Figure 21A) or externally (Figure 21B). There is one velocity of propagation for the direction defined by the line through the two points of contact (called the optic axis) and two velocities for any other direction corresponding to light polarized in two mutually perpendicular planes.

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

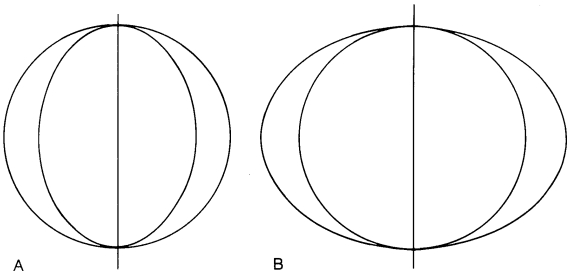


Figure 21: Wave surface for (A) positive and (B) negative uniaxial crystals.

Crystals for which the wave surface has the form shown in Figure 21A are said to be positive uniaxial crystals, and Figure 21B refers in a similar way to negative uniaxial crystals. Huygens thought all crystals were uniaxial, but later observations showed that the general form of the wave surface is more elaborate and is biaxial.

When a parallel beam of plane-polarized light is incident normally upon a thin crystal plate and the crystal is rotated about an axis normal to the plate, two orientations will be found in which a single beam of plane-polarized light emerges; for one orientation, the light is called the ordinary ray, and for the other orientation, the extraordinary ray. Two lines may be drawn on the plate (or on its mount) to indicate the direction of the electric vector of the incident beam when the orientation is such that plane-polarized light emerges. The directions of these lines are called privileged directions for the given anisotropic plate; they are perpendicular to one another.

When light waves pass through an isotropic plate of thickness d, the phase change is $2\pi nd/\lambda$ (n is the index of refraction). For an anisotropic plate there are two indices, n_o and n_e, corresponding to the two privileged orientations, and there are corresponding phase changes, $2\pi n_o d/\lambda$ and $2\pi n_e d/\lambda$. If a beam of light, polarized in some plane other than one of the two privileged orientations, is incident on the plate, it may be resolved into two wave components (lying in two mutually perpendicular planes) that emerge with a phase difference ε_{oe} equal to $2\pi(n_e - n_o)d/\lambda$. The phase difference ε_{oe} is called the retardation of the plate. When the retardation ε_{oe} is equal to $\pm\pi/2$, the plate is called a quarter-wave plate, and when ε_{oe} is equal to $\pm\pi$, it is called a half-wave plate.

The maximum difference of $(n_e - n_o)$ for calcite, a crystal that is strongly birefringent, is about 0.17. Quarter-wave plates are often made by splitting sheets of mica. Inasmuch as the relevant difference of indices is about 0.004, a mica quarter-wave plate is about 60 wavelengths, or 0.03 millimetre thick.

When a parallel beam of plane-polarized light is incident normally on a quarter-wave plate, the emergent light is circularly polarized if the plane of the electric vector bisects the angle between the privileged directions. Otherwise the emergent light is elliptically polarized, the axes of the ellipse being parallel to the privileged directions of the quarter-wave plate. In a similar way, a suitably oriented quarter-wave plate may be used to convert elliptically polarized light to plane-polarized light.

Induced birefringence effects

Many substances that are normally isotropic become weakly birefringent when subjected to shear stress (photoelastic effect; see Plate). Birefringence is also induced by an applied electric field (Kerr effect). A restricted class of crystals that are normally birefringent shows a large change in birefringence when an electric field is applied. Weak birefringence is produced by an applied magnetic field.

Production of polarized light. Unpolarized light may be separated into two components, polarized in perpendicular planes, either (1) by reflection or (2) by refraction. A single reflection at the polarizing angle produces a well-polarized reflected beam of low intensity (*i.e.*, most of the waves are polarized in one plane) and a strong transmitted beam that is only partially polarized. A pile of glass plates provides many surfaces for reflection and thus gives a much stronger reflected beam that is still well polarized; it also considerably improves the degree of polarization of a transmitted beam.

A Nicol prism (Figure 22) is made by cutting a calcite crystal in a suitable plane relative to the crystal axes and cementing the two parts together with Canada balsam. The critical angle for total reflection is less for the ordinary ray than for the extraordinary ray. The crystal is cut at an angle (relative to the crystal axes) such that, for a cone of incident rays of divergence about 24°, the ordinary ray is totally reflected, whereas the extraordinary ray is transmitted. Both rays are almost completely polarized. A prism of this type with air instead of Canada balsam as the separation film was devised by Foucault for polarization of ultraviolet radiation.

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

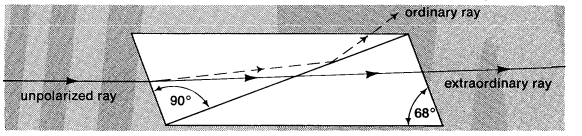


Figure 22: Nicol prism (see text).

Many crystalline substances have different absorption coefficients for the ordinary and extraordinary rays, but the difference is not large enough to provide a useful method of polarization. A series of artificial materials that polarize by absorption has been developed. Complex crystals are produced in a plastic matrix and oriented by stretching. One material of this type, called Polaroid H, transmits about 80 percent of plane-polarized light with the electric vector in one plane and less than 1 percent if the vector is in a perpendicular plane. This type of film offers the most convenient method of obtaining an intense beam of light that is nearly 99 percent plane-polarized.

Analysis of polarized light. Any polarizer has a preferred plane such that if an incident plane-polarized beam has the electric vector in this plane, a high intensity is passed, whereas the device rejects light polarized with the electric vector in a plane perpendicular to the preferred plane. This property implies that any polarizer may be used as an analyzer to test whether or not a given beam of light is plane-polarized. The analyzer is rotated in the beam and if the intensity I of the transmitted light is zero for some orientation, the beam is plane-polarized with the electric vector in a plane perpendicular to that for which the intensity is zero. If there is a position for which the analyzer gives a non-zero minimum, then either (1) the light is elliptically polarized or (2) it is partially polarized. Also if the intensity is not affected by rotating the analyzer, then the light is either (3) unpolarized or (4) circularly polarized. These ambiguities between (1) or (2), and (3) or (4), may be removed and the analysis of a mixture may be completed by using an analyzer and a quarter-wave plate together.

Stokes parameters and Poincaré sphere. Any mixture of different kinds of polarized light may be reduced to an equivalent amount of one kind of polarized light characterized by the orientation and length of axes of an ellipse. It may, of course, happen that the ellipse degenerates to a straight line or a circle. An alternative way of characterizing polarized light is by means of the Stokes parameters, named after a British physicist, George Gabriel Stokes, and defined as follows:

$$S_0 = \langle a_x^2 \rangle + \langle a_y^2 \rangle; \qquad S_1 = \langle a_x^2 \rangle - \langle a_y^2 \rangle;$$
$$S_2 = 2\langle a_x a_y \cos \varepsilon_{xy} \rangle; \qquad S_3 = 2\langle a_x a_y \sin \varepsilon_{xy} \rangle,$$

in which the polarized light is regarded as the resultant

of two beams polarized with vectors a_x and a_y along the coordinates Ox and Oy, and ε_{xy} is the phase difference as in the section above. The brackets $\langle\rangle$ indicate time averages. If the light is completely polarized, then the square of S_0 is equal to the sum of the squares of the other three parameters. This relation was used by a French mathematician, Henri Poincaré, to produce an elegant representation of the properties of polarized light by means of a sphere, which is known as the Poincaré sphere. Each point on the sphere represents a different kind of polarized light; *e.g.*, the two poles represent right- and left-handed circularly polarized light, points on the equator represent plane polarized light; other points on the sphere represent different kinds of elliptically polarized light. The Stokes parameters constitute a matrix representation of polarized light. This kind of representation makes available certain powerful mathematical methods (which have been developed in other connections) for calculating how the state of a beam of polarized light changes on reflection or on passing through a crystal.

Colours of thin plates. When a parallel beam of white light (I_0) is passed through a thin slice of crystal placed between a polarizer and an analyzer, the intensity I of the emergent light may be expressed as the sum of two terms:

$$I = \text{white term} + \text{colour term}.$$

The white term represents the amount of light that would be transmitted if the crystal were not present. It is equal to $I_0 \cos^2 \theta$, in which θ is the angle between the preferred directions of polarizer and analyzer. The colour term depends on the relation between the directions of the crystal axes and the preferred directions of polarizer and analyzer. It is dependent on the retardation ($\varepsilon_{oe} = 2\pi d\,[n_e - n_o]/\lambda$) of the plate as well. Because this retardation varies with the wavelength, the transmission depends on the colour. Some parts of the spectrum are transmitted more than others, and the emergent light is coloured. The first term, on the contrary, does not depend on the properties of the crystal, and it represents white light. It can be reduced to zero by setting the analyzer and polarizer so that their preferred directions are at right angles ($\theta = \pi/2$). The colour is then obtained most strongly, because any transmission of light results from anisotropy (*i.e.*, birefringence) in the slice. This provides a sensitive method of detecting strains in glass and may also be used to detect induced anisotropy when mechanical or electrical stress is applied. This method is sometimes used to detect regions of large strain in beams or girders. A model of the beam is made in plastic and placed between polarizer and analyzer. On applying a load, fringes appear in the regions of most strain. Plate 1 shows a picture obtained in this way. The example is chosen for simplicity, whereas useful application to more complicated structures requires laborious calculations.

The Kerr shutter

A high-speed optical shutter called the Kerr shutter can be made by placing an isotropic material between crossed polaroids, one acting as a polarizer and the other as an analyzer, and applying an electric field. If a suitable material is used, the intensity of the transmitted light will reproduce the variation of electric field, even at frequencies over a megahertz. An improved shutter is produced by using the change of birefringence in some crystals.

Rings and brushes. The retardation of waves by a crystal plate depends on the relation between the perpendicular to the surface of the plate and the axes of the crystal from which it was cut. Observation of the colours of one plate with parallel light gives information in relation to one direction within the crystal. It is possible to obtain information about many directions within the crystal at one and the same time. This is done by taking the light that has passed through the polarizer and converging it strongly, using a microscope objective (reversed). It then passes through the crystal as a solid cone of light. It is rendered nearly parallel again by a second objective and passes through the analyzer. Finally it passes through an optical system that brings all rays that have passed through the crystal in one direction to a focus at a single point in a certain plane (a similar focussing is shown in Figure 12). Thus each point in this plane corresponds to a certain direction in the crystal and to a single value of retardation ε_{oe}. The differences in retardation ε_{oe} are caused partly by variation of $(n_o - n_e)$ with the direction in the crystal and, usually to a lesser extent, by a difference in the length of path within the crystal. This implies that, with monochromatic light, bright and dark fringes are observed. Each dark fringe is the locus of points for which the retardation is an odd number multiple of 180°—*i.e.*, ε_{0e} equals $p_0\pi$ (p_0 being an odd integer). If the slice is cut from a uniaxial crystal with its normal in the direction of the optic axis, the symmetry of the wave surface (Figure 21) requires that the fringes be circular rings (see Plate). The rings are interrupted at certain points for which the second term (in the expression of the intensity of the light transmitted by a crystal plate placed between polarizer and analyzer) is zero. This gives either a bright cross or a dark cross depending on the angle between the preferred directions of the polarizer and the analyzer; these crosses are known as brushes. The patterns obtained with a slice cut so that the normal is not in the direction of the optic axis (a crystal direction for $n_o = n_e$) or with biaxial crystals are complicated and sometimes beautiful (see Plate). Such patterns may be used to identify crystals present as fairly small inclusions in other materials.

ELECTROMAGNETIC-WAVE CHARACTER OF LIGHT

Maxwell's equations. Historically the theory of electricity and magnetism developed in the form of a number of empirical laws each of which was a generalization based on a series of experiments; *e.g.*, Coulomb's law dealt with the force between two stationary electric charges. Maxwell replaced all these laws by a single theory concisely stated in the form of a set of vector equations. It has been said that Maxwell's theory is Maxwell's equations, and indeed it is impossible to do justice to Maxwell's achievement without use of these equations. The content of these equations and their relevance to the theory of light will be described here in general terms. The reader may refer to standard texts on electromagnetic theory for the equations themselves.

Constitutive and field equations

Electric and magnetic fields are specified by means of the vectors E and H with which are associated the vectors D, B, and J (electric and magnetic induction and density of electric current). Maxwell's equations fall into two groups: (1) three constitutive equations and (2) four field equations.

All material bodies contain electrons. These are negative charges circulating around heavier nuclei that are positively charged. When an electric field is applied to a material body, the average positions of the negative charges relative to the positions of the positive charges are changed. This creates an internal electric field. Similarly the action of a magnetic field on a material changes the movement of the electrons and sets up an internal magnetic field. The constitutive equations state that effects within a material body are proportional to the applied fields so that the resultant fields within the body are proportional to the applied field. Certain constants are defined: ε is the dielectric constant, μ is the magnetic permeability, and σ is the electrical conductivity. There is no general agreement, however, concerning these constants, and therefore some authorities use a different nomenclature (for treatment of the constants in greater detail, see ELECTROMAGNETIC RADIATION and ELECTRICITY AND MAGNETISM). But these are basic properties of the material: for free space the values are ε_0, μ_0, and zero respectively.

The first of the four field equations quantifies certain properties of the electric induction at the boundary of a volume that contains a net positive or negative charge. The second states that, since there are no free magnetic poles, a certain integral of magnetic quantities is zero. The third equation states that, when the magnetic flux through a surface changes, electrical voltages appear on the boundaries of the surface. The fourth states that electrical currents in conducting materials and changes of the electric induction in nonconducting materials produce magnetic effects.

The field equations (like the constitutive equations) are linear equations: they state that certain quantities are proportional to one another; *e.g.*, the third equation states

that the electrical voltages are directly proportional to the rate of change of magnetic flux. No new constants (other than ε, μ, and σ), however, are introduced. In Maxwell's equations, electricity and magnetism are two aspects of one thing called electromagnetism. Maxwell's theory can indeed be stated in a way that does not mention electricity and magnetism separately.

Maxwell's new hypothesis

The three constitutive equations and the first three of the field equations are precise formulations of known empirical laws. In the fourth field equation Maxwell introduced a new hypothesis—that an electrical change in a nonconductor produces magnetic effects. This hypothesis—which can be verified by electrical experiments—leads to the theory of electromagnetic waves capable of being propagated through a vacuum.

Propagations of electromagnetic waves. When Maxwell's equations are combined, using standard mathematical methods, a new equation is obtained. This is similar in form to the wave equation (7) given above. It predicts the existence of electromagnetic waves, with well-defined properties which will now be described.

1. In free space (*vacuo*) electromagnetic waves are propagated with a phase velocity $c = (\varepsilon_0\mu_0)^{-1/2}$; the magnitude of c obtained from electrical measurements of $\varepsilon_0\mu_0$ is approximately equal to the measured velocity of light (see Table 1). Plane waves are propagated without attenuation. Spherical waves have an amplitude inversely proportional to the distance from a small source.

2. For transparent nonconducting mediums such as water or glass, the phase velocity is $b = (\varepsilon\mu)^{-1/2}$, and the index of refraction is $n = c/b = (\varepsilon_0\mu_0/\varepsilon\mu)^{-1/2}$. This relation holds for radio waves and, with moderate accuracy, for infrared radiation. For visible light it holds moderately well for some mediums but not for others. The constants ε and μ are associated with redistribution of charge in response to changes in the electromagnetic field. At low frequencies these movements follow the changes of the field in a simple way. At optical frequencies, when the field is reversing about 10^{14} times a second, the situation is more complicated. The effective value of μ at these frequencies is μ_0 and the effective value of ε is less than the value measured with static fields, or fields that vary more slowly—apart from the possibility of resonance, which will be considered later.

3. In conducting mediums, the electromagnetic waves are absorbed as they are propagated. This absorption is exponential; *i.e.,* the fraction absorbed in a thin layer of thickness d is βd, in which β is a constant the value of which can be calculated when ε and σ are known. For metals, β is very large so that a layer only a tenth of a micrometre in thickness absorbs about half of the incident light. This absorption is much stronger than that of nonconducting substances, which are normally regarded as opaque (such as ebony): these are usually quite transparent in thicknesses of a micrometre or so.

Although the simple theory is correct in predicting high absorption for metals in general, the values of β for different metals (calculated from ε and σ) are not correct. This is because the effective value of σ at high frequencies is not the same as that measured with direct current or low frequency alternating current.

4. The waves are transverse; the electric field vector (E) and the magnetic field vector (H) are perpendicular to one another and to the direction of propagation along the axis OZ as shown in Figure 23.

5. The two quantities E and H, the magnitudes of vectors E and H, fluctuate in phase, so that when E is a maximum H also is a maximum (Figure 23). This is implied in the equation

$$\varepsilon^{1/2}E = \mu^{1/2}H, \qquad (12)$$

which enables either E or H to be calculated when the other is known.

6. The electromagnetic field possesses energy, and the energy per unit volume (W) is equal to $\frac{1}{2}(\varepsilon E^2 + \mu H^2)$. In view of equation (12) above, this implies that for electromagnetic waves,

$$W = \varepsilon E^2. \qquad (13)$$

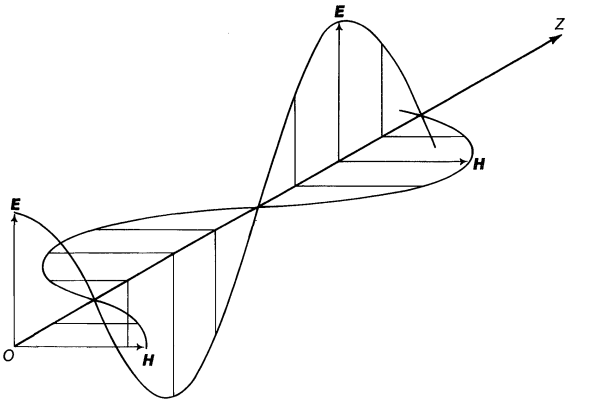


Figure 23: Electromagnetic wave, showing that electric field vector E and magnetic field vector H are in phase (see text).

The Poynting vector

7. The propagation of electromagnetic waves involves a transport of energy. A vector $S = E \times H$, called the Poynting vector (after an English physicist, John Henry Poynting), is used to describe this flow—*i.e.,* the energy flows in the direction of S (which is also the direction of wave propagation in an isotropic medium)—and the amount of energy crossing unit area (lying across the direction of flow) in unit time is equal to S—*i.e.*, to the magnitude of S. For electromagnetic waves $S = E \times H = (\varepsilon/\mu)^{1/2}E^2$, and this relation, combined with equation (13), implies that $S = W/(\varepsilon\mu)^{1/2} E^2$, or $S = bW$; thus, the mean value of S is b times the mean value of W—*i.e.*, the energy crossing unit area in unit time is equal to the velocity multiplied by the energy density.

The linearity of Maxwell's equations implies that the law of superposition applies to electromagnetic waves. The electromagnetic theory is therefore able to give a satisfactory account of all the simpler phenomena of interference and diffraction already described above. There are a few problems for which it is clearly superior to the "scalar-wave" theory that was described. These include the polarization of light transmitted through a slit the width of which is of the same order of magnitude as the wavelength and also the polarization of light scattered from microscopic particles. Calculations based on the electromagnetic theory are sometimes difficult mathematically, but the results are in agreement with experiment.

The interaction of light with matter

REFLECTION AND REFRACTION

When light is incident in a perpendicular direction on the surface of a piece of glass, part is transmitted and part is reflected. This happens at any surface that forms the boundary between two transparent mediums of different refractive indices. For a particle theory of light, this phenomenon constitutes a serious difficulty. If a beam of light consists of a stream of particles all alike, because like causes produce like effects, there is no good reason why some should go through the surface and some be turned back. If the particles are not all the same (*e.g.,* if some have more energy and are therefore able to get through), then it would be expected that all those particles that do penetrate one surface would be able to pass through a second. This is never observed. In fact, the fraction of light transmitted at the second surface is the same as the fraction transmitted at the first (for light of one wavelength). Newton was aware of this fundamental difficulty and introduced the ad hoc hypothesis of particles with periodic oscillation between "fits of easy reflection" and "fits of easy transmission."

A wave theory of light, on the other hand, offers an explanation that follows in a natural and logical way from its premises. When a wave passes from one medium to another certain conditions must be fulfilled at the boundary. These conditions arise because the frequency of the transmitted wave must be the same as that of the incident wave. The velocities of light in the two mediums are different, and therefore, because velocity is the product of

wavelength times frequency, the wavelength cannot be the same. By way of analogy, elastic waves can be imagined falling on a welded boundary between two solids—*e.g.*, steel and brass—with point P_1 just inside the first medium and P_2 extremely near P_1 and just inside the second medium. Then if P_1 and P_2 do not move exactly together the material at the boundary must tear. If the boundary is so strong that it does not tear, then the waves at the boundary must match each other in a most precise way. This match is possible only if there is a reflected wave as well as a transmitted wave: when the density and the elastic properties of the two mediums are known then the fraction reflected can be calculated.

For light, the fraction reflected can be calculated when the indices of refraction are known, and it is found that the fraction reflected when a beam of light is incident perpendicular to a boundary between air (refractive index nearly equal to 1) and a medium of refractive index n is $(n-1)^2/(n+1)^2$. Thus for glass (index 1.5) about 4 percent is reflected, but for diamond (index 2.42) 17 percent is reflected. This result is predicted by any reasonably self-consistent wave theory including the electromagnetic-wave theory. When light is incident on a surface at an angle other than perpendicularly, however, the calculation of the fraction of light reflected is more difficult. This fraction depends both on the direction of incidence and on the polarization of the light.

Fresnel relations

Fresnel derived certain expressions (known as Fresnel's relations) for the amount of light reflected under various conditions, using an elastic solid theory. To do this he assumed boundary conditions chosen ad hoc—*i.e.*, to produce the correct result. In the electromagnetic theory there is no choice in regard to the conditions that must be satisfied at the boundary of two dielectrics. The boundary conditions are fixed by the constitutive equations and by general principles such as the conservation of energy. They are derived from experiments on electricity, not from experiments on light.

These boundary conditions may be applied to electromagnetic waves with the following results:

1. The existence of a wave reflected so that the angle of reflection is equal to the angle of incidence and of a refracted wave the direction of which obeys Snell's law.
2. All the results that Fresnel had previously obtained for the amounts of light reflected and refracted under a variety of conditions, including Brewster's law.
3. The existence of total reflection and of certain observed results in regard to the change of polarization of light that has been totally reflected (when the incident light is polarized).
4. The amount of light reflected at the surfaces of metals (including the result that strong reflection is, in general, obtained with good conductors).
5. A detailed description of certain complicated results obtained when light is incident upon the surface of a transparent anisotropic crystal (*e.g.*, calcite).

DISPERSION AND SCATTERING

Oscillating dipoles. In electrostatics, a dipole consists of two equal and opposite charges situated a small distance apart. The line joining the charges is called the axis of the dipole, the dipole moment M is the name given to the product of one charge times separation. An oscillating dipole may be thought of as one in which charges move so that M fluctuates between $+M_0$ and $-M_0$ (*i.e.*, $M = M_0 \sin 2\pi\nu_t t$). Maxwell's theory shows that an oscillating dipole produces electromagnetic waves that are nearly spherical at distances large compared with the size of the dipole. The total energy emitted is proportional to the square of the ratio M_0/λ^2 (in which $\lambda = c/\nu_t$).

Scattering by free electrons. When an electromagnetic wave (with an electric vector of amplitude E_0) operates on a free electron, it causes the electron to oscillate so as to produce an oscillating dipole with a moment proportional to E_0/λ^2. This dipole emits waves in all directions. These waves are regarded as scattered light. The scattered light is strongest in directions perpendicular to the axis of the dipole, which is the same as the direction of the electric vector in the incident wave. It is polarized with its electric vector in the plane containing the dipole. The ratio of the energy scattered to the energy incident per unit area is: $k = (1/6\pi\varepsilon_0^2)\,(e^4/m^2c^4)$, in which e is the charge and m is the mass of the electron. The energy scattered is equal to the amount of energy incident upon an area equal to k. For this reason k is called the scattering cross section or the scattering coefficient.

Scattering by bound electrons. A bound electron is subject to a restoring force when displaced from its equilibrium position, similar to a weight on the end of a spiral spring. An electron bound to an atom can oscillate with a natural frequency ν_0. Its motion produces an oscillating dipole and electromagnetic waves are emitted. Energy lost in this way constitutes damping, and the oscillation decays like the well-known damped harmonic oscillator of classical mechanics. The curve of decay is similar to the curve of a damped wave shown in Figure 14A. When a light wave is incident, a bound electron oscillates. The strength of the dipole produced and, hence, of the light scattered depends on the frequency (ν_t) of the light wave and also upon the frequency ν_0 and the damping constant γ of the electron.

Dipole radiation

The scattering cross section (k') is found to be:

$$k' = \frac{\nu_t^4}{(\nu_0^2 - \nu_t^2)^2 + \gamma^2\nu_0^2}\,k\,; \qquad (14)$$

and it reaches a maximum value when ν_t is nearly equal to ν_0; *i.e.*, when the incident wave is in resonance with the natural frequency.

Molecular scattering. Because atoms and molecules contain bound electrons, they scatter light, but not just like single bound electrons. The scattered wave has an amplitude proportional to the amplitude (E_o) of the incident wave and inversely to the distance from the scattering centre; *i.e.*, the electric vector E_r of the scattered wave is $\eta E_o/r$, in which η is a constant. This value applies to the wave emitted at right angles to the electric vector of the incident light. The phase of the scattered wave is either the same as that of the incident wave or differs from it by an amount that is the same for each molecule. Figure 24 shows a beam of light incident upon a layer of gas at low pressure. The

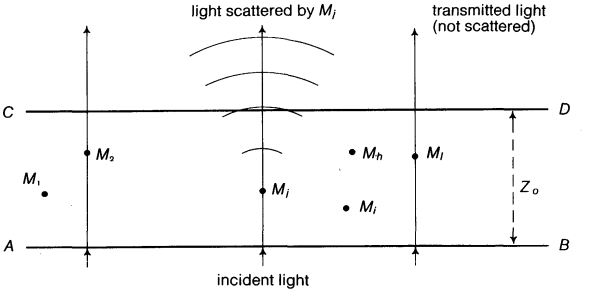


Figure 24: Scattering of light from molecules (see text).

molecules are irregularly placed so that the light scattered from different molecules, such as M_1, M_2, M_r, M_i, etc., in any direction (except the forward direction) is incoherent. The law of photometric summation (see above) applies, so that the total energy scattered by N molecules is just N times the energy scattered by one molecule.

A gas with N molecules per unit volume has an equivalent cross section (k_m),

$$k_m = \frac{8\pi}{3}\eta^2 N. \qquad (15)$$

In the forward direction the light scattered from different molecules is coherent because the total distance from the back edge of the gas layer (AB) to any molecule, such as M_j and M_l and thence to the front edge of the layer (CD) is equal to the thickness of the layer (z_0). Thus in the forward direction, the scattered waves can interfere with each other and with the wave representing the unscattered light. The resultant is a new wave in the forward direction but retarded—*i.e.*, with a phase lag $\eta\lambda N z_0$. In ordinary

wave theory, however, the velocity of light in a medium of refractive index n is c/n and the phase lag due to passage through a layer of thickness z_0 is equal to $2\pi(n-1)z_0$. Equating these two estimates of the lag requires

$$2\pi(n-1)=\eta\lambda^2 N, \tag{16}$$

and substituting for η from equation (16) into equation (15) implies that

$$k_m=\frac{32\pi^3}{3\lambda^4 N}(n-1)^2. \tag{17}$$

This equation was derived by Rayleigh. He also showed that the brightness of the blue sky is about that to be expected if the light is produced only by molecular scattering, and it is not necessary to suppose that there is any appreciable scattering by dust particles—in the high atmosphere and on a clear day. The factor $1/\lambda^4$ implies that the scattering coefficient for blue light ($\lambda = 400$ nanometres) is about six times larger than that for red light ($\lambda = 640$ nanometres) and this accounts for the blue colour. Laboratory measurements of the amount of scattered light agree with that calculated from equation (17). For the rare gases (*e.g.,* argon) the scattered light is almost completely plane-polarized; but other gases show less complete polarization (*e.g.,* 90 percent for carbon dioxide) as would be expected if their molecules possess an electric dipole moment. The light of the blue sky is polarized but polarization is reduced because some of the light has been scattered more than once.

Dispersion. Equation (17) implies that $(n-1)$ may be calculated if η is known; *i.e.,* if the ratio of the amplitude of the scattered wave to that of the incident wave can be calculated. Dispersion formulas for gases have been obtained by assuming that the atoms or molecules contain bound electrons having a response that varies with frequency in the manner described above. The methods of calculation involve the properties of atoms and molecules rather than those of light and are not considered in this article. Certain general results are readily obtained—*e.g.,* that for most transparent mediums the refractive index increases regularly toward the blue end of the spectrum (this is called normal dispersion). It is also found that in a region in which there is a strong absorption, the index varies rapidly, and over part of this region the dispersion is anomalous (*i.e.,* the index increases in the direction of larger wavelengths). This derives from equations (14) and (17).

Normal and anomalous dispersion

Scattering of larger particles. A particle that contains N atoms but is small compared with the wavelength of light has a dipole moment approximately N times that for a single atom. The scattered energy is then proportional to N^2, or to V^2, or to R^6 (if V is the volume, and R is the radius). Thus a particle 20 nanometres in diameter scatters about as much light as 10^{12} separate atoms. As the size of the particle increases, the scattering continues to increase, though more slowly. When the particle is larger than a wavelength it must be considered as a diffraction object. The ratio of effective cross section for scattering to actual cross section fluctuates as the size increases. The theory of this variation, in relation to size, refractive index, and absorption has been worked out and verified experimentally.

Scattering by macroscopic solids and liquids. In crystalline solids the atoms are regularly arranged so that the wavelets scattered in a given direction have regularly varying phases. They interfere to give a small resultant except in the direction of the forward wave. Scattering of light is small in relation to the number of atoms per unit volume and results largely from crystal defects. For X-ray wavelengths smaller than the interatomic distance, the coherent waves scattered in certain directions reinforce one another.

For amorphous solids and liquids the number of atoms in any volume having a diameter of about one wavelength is large and the scattering per atom, although usually larger than in crystals, is still small in relation to the number of atoms involved.

The Kramers–Kronig relation. Any satisfactory theory of dispersion must comply with the condition that the scattered wave can never appear in advance of the incident wave that produces it. Physicists Hendrick Anthony Kramers of The Netherlands and Ralph de Laer Kronig of Germany showed that this basic causality condition implies that the dispersion (*i.e.,* the variation of refractive index with frequency) and the absorption are not independent. They derived equations enabling the absorption to be calculated when the dispersion is known (for all frequencies) and vice versa. It is not surprising that a relationship should exist, because dispersion and absorption are each related to the resonators described above in connection with scattering by bound electrons. The relationship has been found of great importance in many branches of pure and applied physics.

MECHANICAL EFFECTS OF LIGHT

When light is emitted or absorbed there are three mechanical effects: (1) an exchange of energy, (2) an exchange of linear momentum, and (3) an exchange of angular momentum. These are manifested as (1) heating (or cooling), (2) a pressure, and (3) a torque. If the conservation laws apply, then light must possess energy, momentum, and angular momentum. All these effects are predictable either from classical electromagnetic theory or from quantum theory. They can all be explored in relation to either exchanges between light and macroscopic pieces of matter or interaction between radiations and atoms. In this section the first will be treated, leaving the latter for the next section.

Equation (13) is an expression for light energy according to electromagnetic theory, and every detector of light is an energy transducer (converter). There is a general theorem, valid for light and sound, predicting that radiation must exert a pressure equal to W (the energy per unit volume) on a body that absorbs it. Twice this pressure is produced if the light is totally reflected at normal incidence. In electromagnetic theory, when light falls at normal incidence on a metal surface, a current (proportional to the field E) is produced in the metal; and, because this occurs in the magnetic field of the light, there is a reaction proportional to the vector product of the electric and magnetic field strengths ($E \times H$) and normal to the surface—*i.e.,* a reaction proportional to the Poynting vector (S). For an insulator the induced current is the displacement current. Because the Poynting vector is proportional to W, the energy per unit volume, the pressure is proportional to W (further consideration shows that, in the usual units, it is equal to W).

Light pressure

The measurement of the linear momentum of light is difficult because, under normal experimental conditions, the pressure is less than 10^{-4} dyne per square centimetre. In Figure 25, light from powerful sources S_1 and S_2 falls

From R.W. Ditchburn, *Light* (1963); Interscience Publishing, Inc., by permission of John Wiley & Sons, Inc.

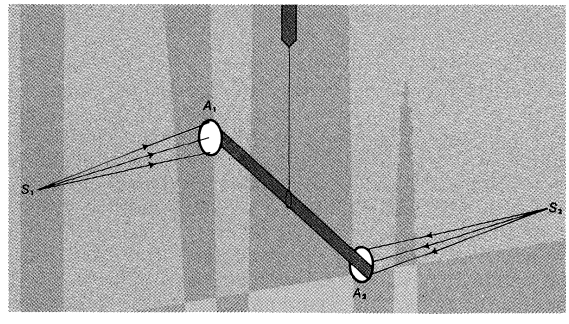
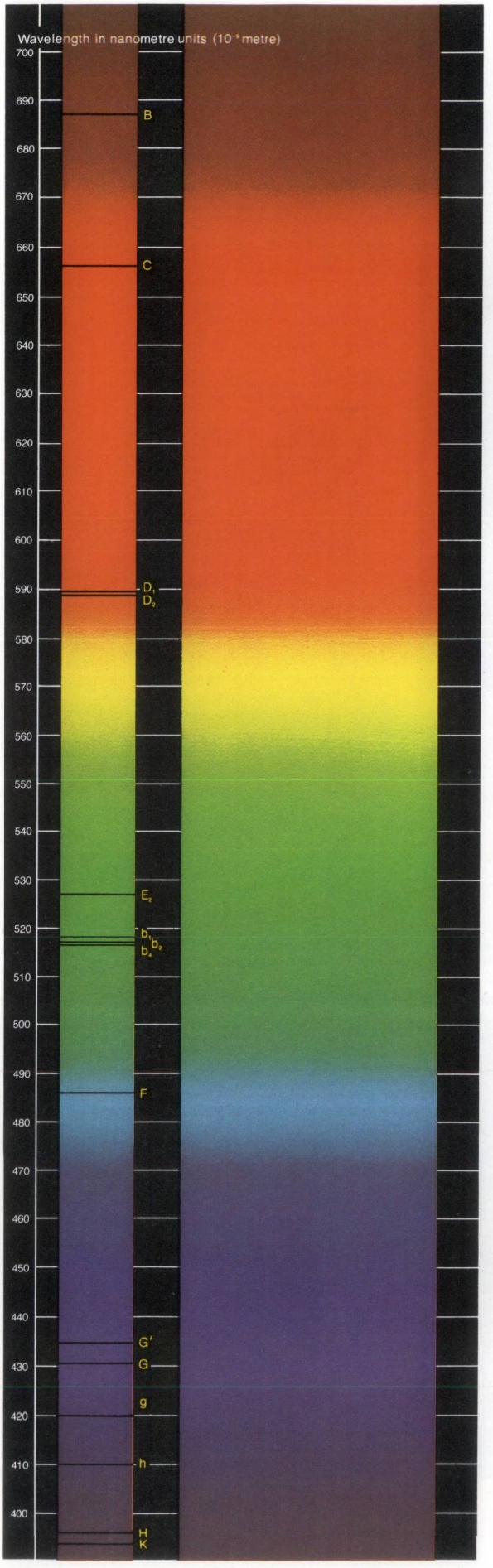
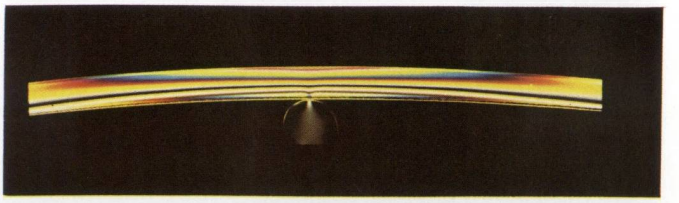


Figure 25: Apparatus for measurement of light pressure (see text).

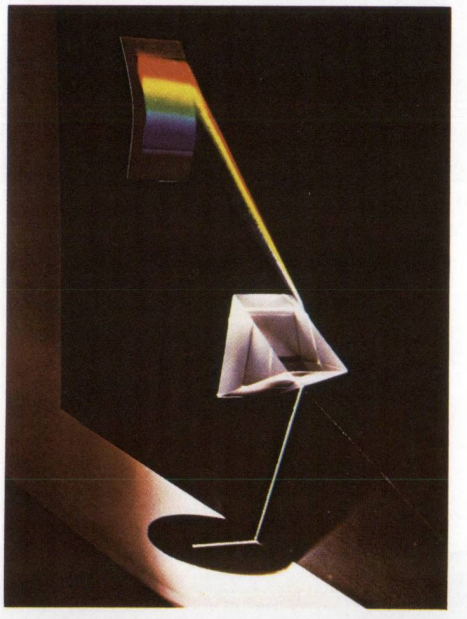
on the mirror vanes A_1 and A_2 that are at the ends of a rod suspended by a quartz fibre. The pressure on the vanes produces a couple or twisting of the fibre and hence a rotation of the system that is measured by the usual mirror-and-scale method. The true radiation pressure can be measured only when a certain thermal action, called the radiometer effect, has been eliminated. In the radiometer





Bending of transparent plastic bar between crossed Polaroids to show photoelastic strain patterns.

The visible solar spectrum (simulated), showing prominent Fraunhofer absorption lines.



Spectrum of white light by (above) a prism and (below) a diffraction grating. With a prism, the red end of the spectrum is more compressed than the violet end.

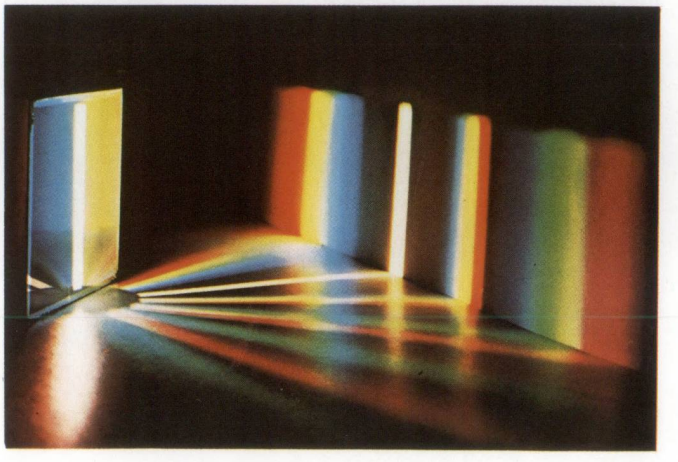
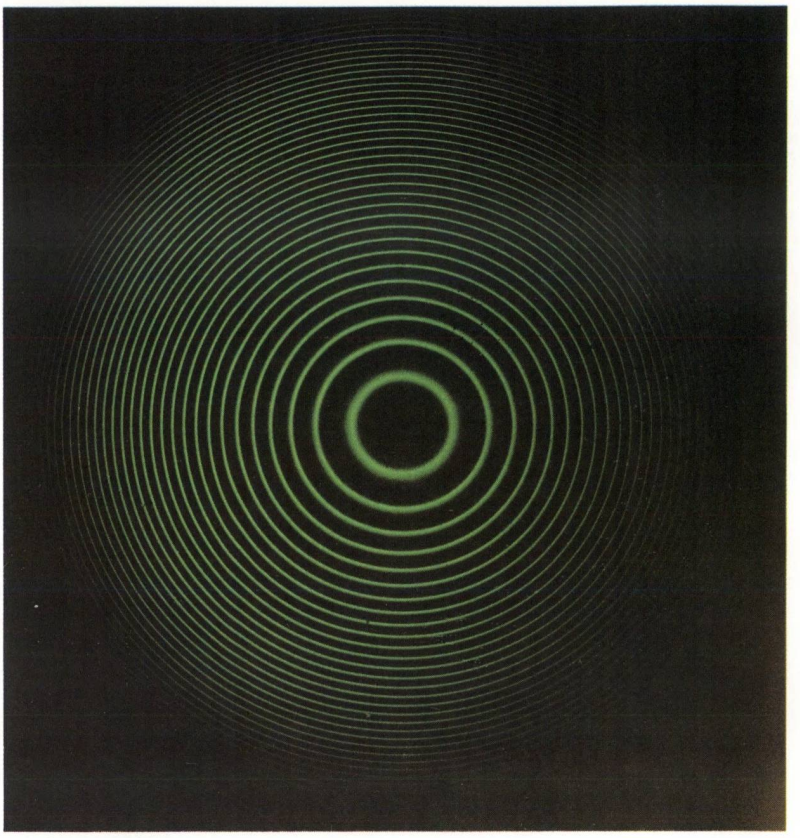
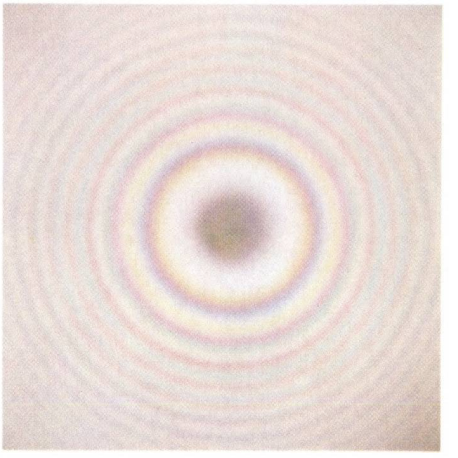


Plate 1: By courtesy of (bottom) Bausch & Lomb, Inc., Rochester, N.Y.; photographs, (left) EB Inc., (top) Charles D. Reilly—EB Inc., (centre) Fritz Goro—Life Magazine © 1944, Time Inc.

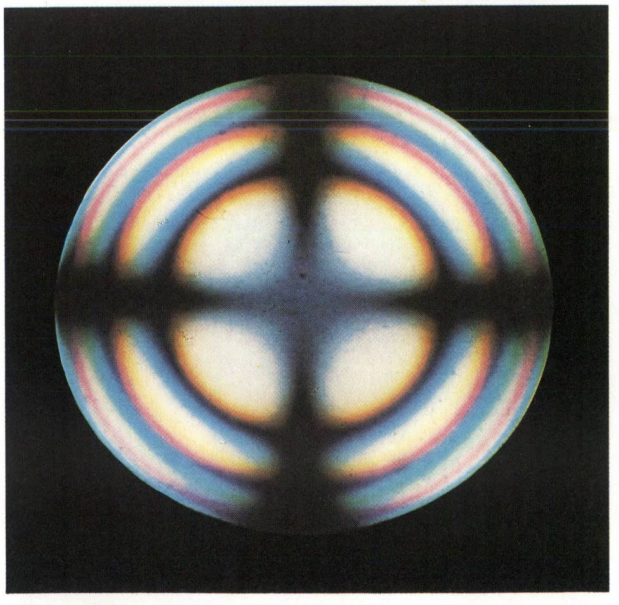


Fabry-Pérot fringes.

Newton's rings.



Interference patterns obtained with crystals in convergent plane-polarized light (polarizer and analyzer crossed): (left) uniaxial crystal cut normal to the axis; (right) biaxial crystal cut perpendicular to bisector of the axes.



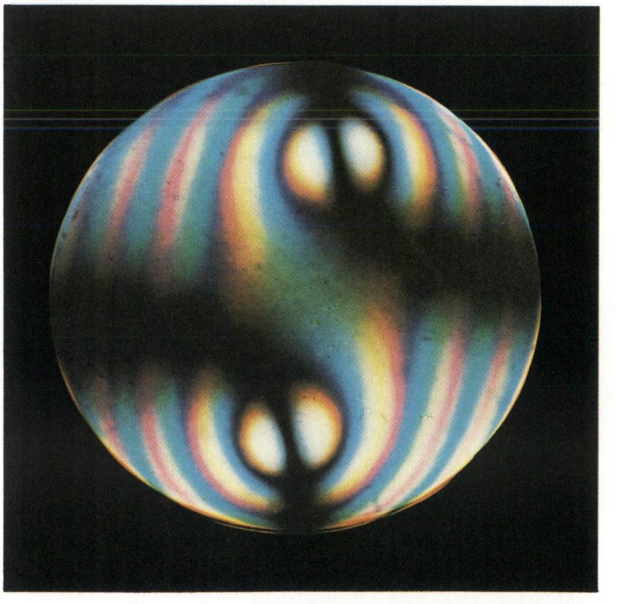


Plate 2: By courtesy of (top left, bottom left, bottom right) R.W. Ditchburn and Blackie & Son Limited, Glasgow and London; photograph, (top right) Charles D. Reilly—EB Inc.

effect, if radiation heats a surface, then molecules of gas that rebound from one side will move faster than those that rebound from the other, cooler side, and there will be a net pressure. In the period 1901 to 1905, when physicists were making the first accurate determination of light pressure, it was not possible to remove this effect by placing the system in a vacuum in which the number of molecules is at a minimum; instead, the effect was reduced by substituting the special cells for the vanes. Any radiometer effect within the cells produces no net force. The residual effects at outer surfaces are small and are the same whichever side of the cell is illuminated. The measured values (1) when radiation is absorbed and (2) when it is reflected agreed with theoretical values within 1 percent. By use of an alternative method of eliminating the radiometer effect, results in agreement with the theoretical values were also obtained.

It can be shown that a beam of circularly polarized light possesses angular momentum and thus should exert a torque upon a half-wave plate that reverses the rotation of polarization (*i.e.*, from left-handed to right-handed circularly polarized light or vice-versa). The experimental difficulties of measuring this torque (of about 10^{-11} dyne centimetre) are formidable and the overall error is about 10 percent. The result agrees with the theoretical value within this margin.

The density of radiation incident on the Earth's surface at its mean distance from the Sun is about 1.36 kilowatts per square metre (for a surface normal to the Sun's rays), and the corresponding light pressure is about 5×10^{-5} dyne per square centimetre. The total pressure on the Earth is about 6×10^{14} dynes, a minute fraction of the gravitational force exerted by the Sun.

Importance of the mechanical properties of light

The mechanical properties of light, difficult to measure in the laboratory, are important under certain astronomical conditions. Gravitational forces are proportional to the mass of a body and hence to the cube of its radius. For a body having large radius compared with the wavelength, the radiation pressure is proportional to the square of the radius. Whereas the attraction and repulsion may be equal and balanced for a medium size particle, for a smaller particle, the radiation pressure may be a large fraction of the whole force. For this reason, radiation pressure has a considerable effect in extending a comet's tail. In a star the outward flow of radiation from the centre produces a net force tending to expand the star, thus opposing the gravitational forces. Also, when radiation energy is transferred between two parts of a rotating star that are moving with different velocities, momentum transfers are involved. This produces an effect known as radiative viscosity because the resulting forces reduce the relative velocity of adjacent layers and so reduce the angular velocity of rotation.

Quantum theory of light

In the classical electromagnetic wave theory, the absorption of light is a continuous process and there is no lower limit to the amount of energy that an atom can absorb from light of given frequency. Energy can be exchanged, in infinitesimal amounts between radiation and atoms, in an enclosure that is in thermal equilibrium. The total energy of the enclosure should, in equilibrium at temperature T, be distributed so that each degree of freedom has an equal amount, kT, in which k is Boltzmann's constant. The number of possible modes of vibration for transverse waves with frequencies between ν and $\nu + d\nu$ is $8\pi\nu^2 d\nu$ per unit volume. (In this section all frequencies are temporal frequencies and ν is used for ν_t). If $\rho(\nu)d\nu$ is the radiation energy per unit volume in this frequency range, and c is the velocity of light, then the radiation density is as shown in curve R–J of Figure 26. This law, derived by Rayleigh and Jeans, is the inevitable consequence of classical wave theory. It predicts that the radiation density increases without limit as the frequency increases, and this implies that, in equilibrium, all the energy in the universe is found in the high-frequency end of the electromagnetic spectrum and none is present in matter. This catastrophe may be avoided by assuming that there is some upper limit to the possible frequency of radiation, but the Rayleigh–Jeans law does not agree with measurements of temperature radiation in the visible and ultraviolet regions of the spectrum (curve R–J, Figure 26). The experimental results are represented by curve P in Figure 26, and Planck suggested a formula, now called Planck's law, that fitted this curve. Originally this formula was probably an inspired guess, but Planck soon showed that it could be derived by assuming that a mode of frequency can change only by $h\nu$ (or by a multiple of $h\nu$) in which h is a new universal constant now called Planck's constant. The minimum "element of energy" (as Planck referred to it) is now called the quantum. For visible light the quantum is about 3.5×10^{-12} erg or 3.5×10^{-19} joule.

The theory in which Planck introduced the quantum was concerned with the interaction between radiation and matter. He expected that only minor modifications of the electromagnetic theory of light in free space and of the electron theory of matter would be needed. This expectation was not realized. Experiments, some of which will be described below, led Einstein to regard light as an assembly of entities, later called photons. A frequency ν, energy $h\nu$, momentum $h\nu/c = h/\lambda$, and angular momentum $h/2\pi$ are associated with a photon of circularly polarized light. In many situations the photons appear to be localized

Properties of a photon

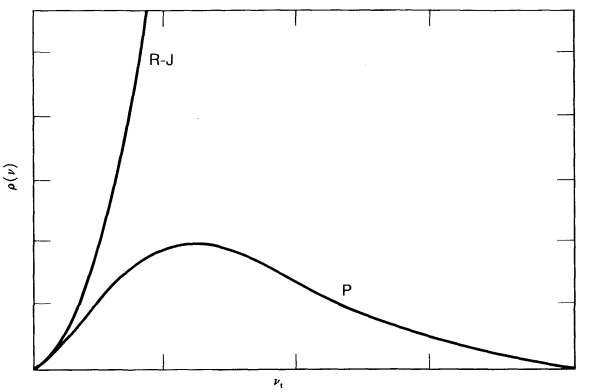


Figure 26: Equilibrium radiation density curve R–J (Rayleigh–Jeans); curve P (Planck; see text).

concentrations of energy and momentum—*i.e.*, to have the properties of particles. Yet the evidence that led to the rejection of a corpuscular theory in favour of a wave theory is still as valid and as compelling as it was in the time of Fresnel.

The theory, which includes logical descriptions both of those experiments that seem to require a corpuscular theory and those that support a wave theory, is known as quantum mechanics. This is a general theory of radiation and matter—not a special hypothesis about their interaction. The account given below of some aspects of quantum mechanics that are important in relation to experiments on light must be related to the general theory (see MECHANICS: *Quantum mechanics*).

PHOTONS

When light of suitable wavelength is incident upon the clean surface of a metal, electrons are emitted. If W_m is the maximum kinetic energy of one of those photoelectrons released by light of frequency ν, then it is found that the maximum kinetic energy is Planck's constant times the difference between two frequencies, or $W_m = h(\nu - \nu_0) = h\nu - W_0$, in which W_0 is written for $h\nu_0$. No electrons are emitted when ν is less than the critical frequency ν_0. It is known from other experiments that a minimum amount of energy called the work function is needed for an electron to escape from a surface. If W_0 is accepted as the work function for emission of a photoelectron, then it is easily understood that no electrons are emitted when frequency ν is less than the critical frequency and that the maximum energy of escape for an electron is $h\nu - W_0$. Some electrons may make collisions after absorption of energy $h\nu$ and either fail to emerge from the surface or else emerge with lower energy. Escape

by gaining energy from two photons of frequency less than ν_0 is extremely improbable because the energy due to one absorption is almost always lost by collision before a second photon is absorbed.

The number of photons absorbed, and therefore the number of electrons released, are proportional to the energy density in the beam of light and to the time. The energy of individual electrons is independent of the energy density—*i.e.,* independent of the amplitude of the waves.

Conservation of energy and momentum. Niels Bohr, a Danish physicist, postulated that an atom can exist only in one of a set of discrete stationary states of internal energy ($W_1, W_2, \ldots$). The frequency $h\nu_{nm}$ of light emitted in passing from state n to m is related to the energy difference by Bohr's frequency condition: that Planck's constant times the frequency is equal to the energy difference between two states; *i.e.*,

Bohr's postulate

$$h\nu_{nm} = W_n - W_m.$$

For some pairs of states ($W_n - W_m$) can be measured by experiments on collisions between electrons and atoms, and Bohr's condition is found to be satisfied within the limits of experimental error. Many other, less direct but more accurate, experiments are also in agreement with this equation.

In Bohr's theory, and subsequently in quantum mechanics, an angular momentum is associated with each atomic state. In the simplest situation, when polarized light is emitted in a magnetic field, the angular momentum about a known axis changes by $h/2\pi$. The changes are such that angular momentum is conserved if a photon of circularly polarized light has an angular momentum of $h/2\pi$ about the direction of propagation.

It was shown by a United States physicist, Arthur Holly Compton, that when X-radiation is scattered by a nearly free electron, there is a wavelength change. A formula that gives the wavelength change as a function of the angle of scattering was derived by assuming that a "particle" of energy $h\nu$ and momentum $h\nu/c$ collides with a particle of mass m that is at rest, and that energy and linear momentum are conserved. The resting electron gains momentum and kinetic energy by changing its speed. The photon cannot change its speed, which is always c. It therefore must change its frequency when it changes energy and momentum. The measured change of frequency agrees with that calculated by Compton.

Compton scattering

An atom that has absorbed a quantum of energy $h\nu_{mn}$ and changed from state m to state n may return to state m with emission of a photon of the same frequency (resonance radiation)—*i.e.,* $\nu_{mn} = \nu_{nm}$. If there is a state s intermediate in energy between states m and n, the return may be in two stages; n to s and s to m, with the emission of two photons. The sum of the two quanta $h\nu_{ns}$ and $h\nu_{sm}$ is then equal to a single quantum $h\nu_{nm}$ so that energy is conserved. In gases, these emissions usually follow the absorption in less than a microsecond, and the term fluorescence is used to include both resonance radiation and multistage processes. In solids and liquids there are sometimes intermediate states that are metastable so that emission occurs over much longer periods (phosphorescence; see below *Luminescence*). In all these processes the frequency of the emitted radiation is not greater than that of the absorbed radiation in accord with an empirical law discovered in the 19th century. In quantum theory this law is a necessary consequence of conservation of energy.

In Rayleigh scattering, the frequency of the scattered light is equal to that of the incident light. The physicists Sir Chandrasekhara Venkata Raman of India and Leonid Isaakovich Mandelstamm of the Soviet Union independently discovered processes in which light is scattered by an atom or molecule that changes its state. There is a corresponding discrete change in the frequency of the scattered radiation. When the scattering atom or molecule is in its lowest state, the frequency of the unscattered radiation is always less than that of the incident light (Stokes lines). When light is scattered by a molecule that is not in its lowest state, photons of higher frequency may be found in the scattered radiation (anti-Stokes lines). The scattered photon then has more energy than the incident photon, but the molecule gives up energy by passing to a lower state. The Raman effect is observed when light is scattered by solids and liquids. There is also an additional kind of scattering in which photons exchange energy and momentum with the thermal vibrations of the solid or liquid (Brillouin scattering, named after a French-U.S. physicist, Léon Brillouin).

Raman effect

The experiments quoted are examples of a large number that are in accord with the view that, within the limits of measurement, energy and momentum are conserved in detail—*i.e.,* in each individual interaction between an atom and radiation.

Spontaneous and stimulated emission. Planck's theory gives the distribution of radiation energy with frequency in an assembly of atoms and radiation in thermodynamic equilibrium at a common temperature in an enclosure. An Austrian physicist, Ludwig Boltzmann, applying the same principles to the assembly of atoms, derived a formula that gives the distribution of atoms between different stationary states—*i.e.,* the ratio of the number N_n in any other state n to the number N_m in any other state m. In the enclosure now considered, Boltzmann's and Planck's laws must both be satisfied though each atom is changing its state frequently by emitting or absorbing a photon or by collision. Equilibrium can be maintained if, and only if, each individual reaction is balanced by its own reverse process. Einstein assumed that the equilibrium was maintained by the following process: (1) absorption of radiation at a rate $B_{mn}N_m\rho(\nu)$; (2) spontaneous emission at a rate $A_{nm}N_n$; and (3) stimulated emission at a rate $B_{nm}N_n\rho(\nu)$, in which A_{nm}, B_{mn}, and B_{nm} are constants, now known as Einstein coefficients. When the rate of absorption was equated to the rate of emission (*i.e.,* to the sum of spontaneous and stimulated emission), an equation of the same form as Planck's law was obtained. This equation became identical with Planck's law if

Einstein coefficients

$$B_{nm} = B_{mn} = \frac{c^3}{8\pi h\nu^3}A_{nm}. \qquad (18)$$

If there were no stimulated emission (*i.e.,* if $B_{nm} = 0$), Planck's law and Boltzmann's law could not both be satisfied. Einstein postulated the existence of stimulated emission in 1905, but for more than 30 years it appeared to be only of theoretical interest. At the densities of radiation then available, the rate of stimulated emission was small compared with the rate of spontaneous emission. Equation (18) shows that in the microwave region of the electromagnetic spectrum (frequencies of a few thousand megahertz) the ratio B_{mn}/A_{nm} is much larger than in the visible spectrum. Accordingly, it was much easier to develop microwave amplification by stimulated emission of radiation (maser) than the corresponding laser effect with radiation in or near the visible spectrum. The existence of stimulated emission can be fitted into a classical wave theory, like many other concepts that originated in relation to theories of photons.

Interference and diffraction of photons. It was at one time suggested that interference and diffraction phenomena are caused by collisions between photons considered as small particles, or at least to some kind of complex interaction between photons. A simple experiment performed by a British physicist, Geoffrey Ingram Taylor, in 1908 excludes this possibility. He photographed the diffraction pattern of a needle and reduced the illumination until long exposures were needed to obtain an image. When the chance of two or more energy quanta passing through the apparatus simultaneously was made extremely small, the diffraction pattern was exactly the same as that obtained with a strong source of light.

This experiment is supported by many later experiments, showing that interference and diffraction are to be associated with single photons. The discussion in the section on coherence above implies that when ordinary sources of light are used each photon can interfere only with itself and not with any other photon. An interference pattern can, of course, be photographed only by recording the effects of many photons because one photon can activate only one grain on the photographic plate, but it does not

matter whether the photons all arrive over a time span of a microsecond or of several weeks.

The photograph of an interference pattern is both a wave phenomenon because it shows the characteristic spatial periodicity and a quantum phenomenon because the whole energy of a photon can be used to activate a single grain. It is not possible to trace the path of one photon (regarded as a particle) through the apparatus and, at the same time, obtain the interference pattern. In Young's experiment there is no way of finding out through which slit a given photon passes—except by covering one slit and thus losing the interference fringes.

THE WAVE–PARTICLE NATURE OF LIGHT

Uncertainty relations. Experiments with the photoelectric effect show that energy can be transferred from one atom to another in a way that suggests that photons are corpuscles—*i.e.,* localized concentrations of energy and momentum. Other experiments imply equally clearly that the light emitted from an atom, when it loses energy, must be represented by wave groups. Both of these sets of experiments are equally valid, and together they require that the photon must have both wave-group and particle properties at the same time.

It will now be assumed that the intensity of the waves gives the relative probability that a suitable detector will detect a photon at a given point. Experimentally, waves and particles are both abstractions, each describing the same physical system. Quantum mechanics does not seek a relationship such that particles are guided by waves according to the laws of classical physics—as in the theory of the English physicist J.J. Thomson, in which photons are pictured as closed tubes of electric force guided by Maxwellian waves. The waves and particles now to be considered do not have all the relevant properties of classical waves and particles. When a photon is absorbed, the whole of a wave group, possibly extending over a large volume, is annihilated. The position of a photon particle cannot be exactly specified. It is known only that it is within the wave group and most likely exists where the waves are most intense. Also the energy and momentum of the photon are determined by the frequency of the waves. Fourier analysis of wave groups shows that, when the wave group is short (so that the position of the photon is fairly well determined), the frequency of the waves and, hence, the energy and momentum of the photon are known only within a wide range. The product of the uncertainty in position and the uncertainty in momentum can never be less than a certain minimum value. It will now be shown how these relations appear in some experimental situations.

In one experiment a parallel beam of light is incident on a slit of width d, producing a diffraction pattern on a screen. The y-coordinate, at the moment when a photon passes through the slit, is known within an uncertainty $\Delta y = \pm d$. The direction of propagation when the photons emerge from the slit is uncertain because of diffraction, but most of the photons are observed within an angular range $\pm\theta$, in which $\sin\theta = \lambda/d$. There is therefore an uncertainty Δp_y in the y-component of the momentum (p) of about $\pm(h \sin\theta)/\lambda$, so that the product of the two uncertainties, Δp and Δy, is

$$\Delta p \Delta y \sim \frac{dh \sin\theta}{\lambda} \sim h ,$$

in which the sign $\sim$ indicates is of the "order of magnitude of." In a second experiment, monochromatic light is passed through an optical shutter. The shutter is opened for a short time Δt so that a wave group of duration Δt and length $\Delta z = c\Delta t$ passes through the shutter. Then $c\Delta t$ is the length l of the wave train. This implies that the frequency ν varies within a range $\Delta\nu = 1/\Delta t$—*i.e.,* that the photon energy varies within a range $\Delta E \sim h/\Delta t$ so that $\Delta E \Delta t \sim h$. More detailed analysis shows that, if the uncertainties Δt and ΔE are the root mean square deviations of statistical theory, then their product is equal to or larger than $h/4\pi$; so that $\Delta E \Delta t \ h/4\pi$.

Uncertainty in energy of photons

Individual examples do not prove that the uncertainty relations are generally true any more than individual examples prove the general validity of the second law of thermodynamics. The fact that every example agrees, and that ingenious attempts to find exceptions all fail, constitutes a kind of proof by default, but the real proof is in the whole range of experimental work that demonstrates the equal status and inescapable association of wave and particle properties.

Maximum observations. The uncertainty relations show that precise measurement of both position and momentum are incompatible. It follows that a precise deterministic theory of particle dynamics is not appropriate because the precise initial condition of a particle is not observable. An extension of the uncertainty relations shows that the number of photons in a light beam and its phase cannot be simultaneously measured with precision, and thus a classical wave theory is also excluded.

In any experimental situation there are, in general, two or more observations that are compatible. Neither affects the other, and their results are independent of the order in which they are made. A complete assembly of such measurements is said to constitute a maximum observation. A state of a system is specified by giving the result of a maximum observation or in an equivalent way. The relative probabilities of future observations on a system can be calculated from the result of a maximum observation without reference to any previous results obtained on the system. For every system there is one certain prediction—that an immediate repetition of the maximum observation will give the same result.

Empirical questions. There is a certain type of question to which the quantum mechanics gives no direct answer because the question is not related to a possible observation. One may ask, for instance, what happens to a photon of right-handed circularly polarized light as it passes through a half-wave plate. It is known that the photon will emerge as a photon of left-handed circularly polarized light and that angular momentum is transferred to the plate. The wave theory gives an account of the progress of a wave through the plate. If, however, one tries to visualize the particle and asks for a mental picture of the way in which this particle alters as it goes through the plate, there is no ready answer. The theory has nothing to say unless an experiment is designed to determine what happens to the photon in passage through the plate. For example, a half-wave plate is split in two and the light is passed through the quarter-wave plate, which forms one half. It may subsequently be passed through a Nicol prism analyzer. There will be some orientation of the Nicol prism such that the photon will certainly pass through. If a Nicol prism is oriented at a to this direction, the probability for a single photon to pass through is $\cos^2 a$. This detailed theoretical prediction agrees with experiment. Thus the quantum mechanics correctly predicts the results of an experiment but leaves unanswered the original question about the way in which the photon, visualized as a particle, changes its angular momentum as it passes through the half-wave plate.

Assemblies of photons. The application of quantum mechanics to the electromagnetic field in an enclosure justifies a representation in terms of photons. These may be regarded as indistinguishable particles. Any number of these particles may exist simultaneously in the same energy state. The statistics of particles that obey these conditions is called Einstein–Bose statistics. The distribution of particles among different states may be calculated by the usual methods of statistical mechanics in which probability (or entropy) is maximized subject to certain limitations. In an assembly of material particles the total energy and also the number of particles are constants. With photons, the second condition is no longer valid because two photons of frequencies ν_{ns} and ν_{sm} may be replaced by one of frequency $\nu_{nm} = \nu_{ns} + \nu_{sm}$, thus decreasing the number of photons while conserving the total energy. The statistical calculation leads straightforwardly to Planck's law. It is found that for ordinary sources of light the probable number of photons in any given state is always much less than unity. For a laser source many photons are found in relatively few states, and, indeed, in an ideal single

mode continuously operated laser, all photons would be in one state.

A considerable amount of experimental and theoretical work has recently been done on coincidences of photon detection in two or more detectors that are observing the same source. This kind of photon statistics measures the coherence of light from a star and hence the size of the star.

Assemblies of photons and atoms. In the Dirac theory of absorption and emission of radiation, quantum mechanics is applied to an enclosure that contains radiation and free atoms or molecules. Processes involving creation and annihilation of photons with corresponding atomic transitions must maintain equilibrium. The theory derives expression for the Einstein A and B coefficients of absorption and emission and for their ratios. It is also shown that, whereas spontaneous emission is random, stimulated emission is coherent with the radiation that excites it. The theory extends to include an account of Rayleigh scattering (and hence of dispersion) and of the Raman effect. All problems concerned with the absorption, emission, and scattering of light are, in principle, soluble, although the mathematical formulation, even of approximate formulas, is usually difficult.

The quantum-mechanical theory is adequate at optical frequencies and at lower frequencies. It is also adequate at the higher frequencies associated with ultraviolet radiation and for X-ray scattering including the Compton effect. It is inadequate for photons of the still higher frequencies involved in the creation and annihilation of fundamental particles. It is to be expected that, in due course, a basic theory of electromagnetic forces and of their relation to gravitation and to the strong forces between nucleons will emerge. The present theory will then have the status of an approximation valid for a wide range of frequencies including the frequencies associated with visible light.

Limitations of quantum mechanics

(R.W.Di.)

Luminescence

SOURCES AND PROCESS

Luminescence is the emission of light by certain materials when they are relatively cool. It is in contrast to light emitted from incandescent bodies, such as burning wood or coal, molten iron, and wire heated by an electric current. Luminescence may be seen in neon and fluorescent lamps; television, radar, and X-ray fluoroscope screens; organic substances such as luminol or the luciferins in fireflies and glowworms; certain pigments used in outdoor advertising; and also natural electrical phenomena such as lightning and the aurora borealis. In all these phenomena, light emission does not result from the material being above room temperature, and so luminescence is often called cold light. The practical value of luminescent materials lies in their capacity to transform invisible forms of energy into visible light.

Luminescence emission occurs after an appropriate material has absorbed energy from a source such as ultraviolet or X-ray radiation, electron beams, chemical reactions, and so on. The energy lifts the atoms of the material into an excited state, and then, because excited states are unstable, the material undergoes another transition, back to its unexcited ground state, and the absorbed energy is liberated in the form of either light or heat or both (all discrete energy states, including the ground state, of an atom are defined as quantum states). The excitation involves only the outermost electrons orbiting around the nuclei of the atoms. Luminescence efficiency depends on the degree of transformation of excitation energy into light, and there are relatively few materials that have sufficient luminescence efficiency to be of practical value.

Luminescence a quantum process

Luminescence and incandescence. As mentioned above, luminescence is characterized by electrons undergoing transitions from excited quantum states. The excitation of the luminescent electrons is not connected with appreciable agitations of the atoms that the electrons belong to. When hot materials become luminous and radiate light, a process called incandescence, the atoms of the material are in a high state of agitation. Of course, the atoms of every material are vibrating at room temperature already, but this vibration is just sufficient to produce temperature radiation in the far infrared region of the spectrum. With increasing temperature this radiation shifts into the visible region. On the other hand, at very high temperatures, such as are generated in shock tubes, the collisions of atoms can be so violent that electrons dissociate from the atoms and recombine with them, emitting light: in this case luminescence and incandescence become indistinguishable.

Luminescent pigments and dyes. Nonluminescent pigments and dyes exhibit colours because they absorb white light and reflect that part of the spectrum that is complementary to the absorbed light. A small fraction of the absorbed light is transformed into heat, but no appreciable radiation is produced. If, however, an appropriate luminescent pigment absorbs daylight in a special region of its spectrum, it can emit light of a colour different from that of the reflected light. This is the result of electronic processes within the molecule of the dye or pigment by which even ultraviolet light can be transformed to visible—*e.g.*, blue—light. These pigments are used in such diverse ways as in outdoor advertising, blacklight displays, and laundering: in the latter case, a residue of the "brightener" is left in the cloth, not only to reflect white light but also to convert ultraviolet light into blue light, thus offsetting any yellowness and reinforcing the white appearance.

EARLY INVESTIGATIONS

Although lightning, the aurora borealis, and the dim light of glowworms and of fungi have always been known to mankind, the first investigations (1603) of luminescence began with a synthetic material, when Vincenzo Cascariolo, an alchemist and cobbler in Bologna, Italy, heated a mixture of barium sulfate (in the form of barite, heavy spar) and coal; the powder obtained after cooling exhibited a bluish glow at night, and Cascariolo observed that this glow could be restored by exposure of the powder to sunlight. The name *lapis solaris,* or "sunstone," was given to the material because alchemists at first hoped it would transform baser metals into gold, the symbol for gold being the Sun. The pronounced afterglow aroused the interest of many learned men of that period, who gave the material other names, including phosphorus, meaning "light bearer," which thereafter was applied to any material that glowed in the dark.

Today, the name phosphorus is used for the chemical element only, whereas certain microcrystalline luminescent materials are called phosphors. Cascariolo's phosphor evidently was a barium sulfide; the first commercially available phosphor (1870) was "Balmain's paint," a calcium sulfide preparation. In 1866 the first stable zinc sulfide phosphor was described. It is one of the most important phosphors in modern technology.

Phosphors

One of the first scientific investigations of the luminescence exhibited by rotting wood or flesh and by glowworms, known from antiquity, was performed in 1672 by Robert Boyle, an English scientist, who, although not aware of the biochemical origin of that light, nevertheless established some of the basic properties of bioluminescent systems: that the light is cold; that it can be inhibited by chemical agents such as alcohol, hydrochloric acid, and ammonia; and that the light emission is dependent on air (as later established, on oxygen).

In 1885–87 it was observed that crude extracts prepared from West Indian fireflies (*Pyrophorus*) and from the boring clam, *Pholas,* gave a light-emitting reaction when mixed together. One of the preparations was a cold-water extract containing a compound relatively unstable to heat, luciferase; the other was a hot-water extract containing a relatively heat-stable compound, luciferin. The luminescent reaction that occurred when solutions of luciferase and luciferin were mixed at room temperature suggested that all bioluminescent reactions are "luciferin–luciferase reactions." In view of the complex nature of bioluminescent reactions, it is not astonishing that this simple concept of bioluminescence has had to be modified. Only a small number of bioluminescent systems have been investigated for their respective luciferin and the corresponding luciferase, the best known being the bioluminescence of

Bioluminescence

fireflies from the United States; a little crustacean living in the Japanese sea (*Cypridina hilgendorfii*); and decaying fish and flesh (bacterial bioluminescence). Although bioluminescent systems have not yet found practical applications, they are interesting because of their high luminescence efficiency.

The first efficient chemiluminescent materials were nonbiological synthetic compounds such as luminol (with the formula 5-amino-2,3-dihydro-1.4-phthalazinedione). The strong blue chemiluminescence resulting from oxidation of this compound was first reported in 1928.

PHOSPHORESCENCE AND FLUORESCENCE

The name luminescence has been accepted for all light phenomena not caused solely by a rise of temperature, but the distinction between the terms phosphorescence and fluorescence is still open to discussion. With respect to organic molecules, the term phosphorescence means light emission caused by electronic transitions between levels of different multiplicity (explained more fully below), whereas the term fluorescence is used for light emission connected with electronic transitions between levels of like multiplicity. The situation is far more complicated in the case of inorganic phosphors.

The term phosphorescence was first used to describe the persistent luminescence (afterglow) of phosphors. The mechanism described above for the phosphorescence of excited organic molecules fits this picture in that it is also responsible for light persistence up to several seconds. Fluorescence, on the other hand, is an almost instantaneous effect, ending within about 10^{-8} second after excitation. The term fluorescence was coined in 1852, when it was experimentally demonstrated that certain substances absorb light of a narrow spectral region (*e.g.*, blue light) and instantaneously emit light in another spectral region not present in the incident light (*e.g.*, yellow light) and that this emission ceases at once when the irradiation of the material comes to an end. The name fluorescence was derived from the mineral fluorspar, which exhibits a violet, short-duration luminescence on irradiation by ultraviolet light.

LUMINESCENCE EXCITATION

Chemiluminescence and bioluminescence. Most of the energy liberated in chemical reactions, especially oxidation reactions, is in the form of heat. In some reactions, however, part of the energy is used to excite electrons to higher energy states, and, for fluorescent molecules, chemiluminescence results. Studies indicate that chemiluminescence is a universal phenomenon, although the light intensities observed are usually so small that sensitive detectors are necessary. There are, however, some compounds that exhibit brilliant chemiluminescence, the best known being luminol, which, when oxidized by hydrogen peroxide, can yield a strong blue or blue-greenish chemiluminescence. Other instances of strong chemiluminescences are lucigenin (an acridinium compound) and lophine (an imidazole derivative). In spite of the brilliance of their chemiluminescence, not all of these compounds are efficient in transforming chemical energy into light energy, because only about 1 percent or less of the reacting molecules emit light. During the 1960s, esters (organic compounds that are products of reactions between organic acids and alcohols) of oxalic acid were found that, when oxidized in nonaqueous solvents in the presence of highly fluorescent aromatic compounds, emit brilliant light with an efficiency up to 23 percent.

Role of enzymes

Bioluminescence is a special type of chemiluminescence catalyzed by enzymes. The light yield of such reactions can reach 100 percent, which means that almost without exception every molecule of the reacting luciferin is transformed into a radiating state. All of the bioluminescent reactions best known today are catalyzed oxidation reactions occurring in the presence of air.

Triboluminescence. When crystals of certain substances—*e.g.*, sugar—are crushed, luminescent sparkles are visible. Similar observations have been made with numerous organic and inorganic substances. Closely related are the faint blue luminescence observable when adhesive tapes are stripped from a roll, and the luminescence exhibited when strontium bromate and some other salts are crystallized from hot solutions. In all of these cases, positive and negative electric charges are produced by the mechanical separation of surfaces and during the crystallization process. Light emission then occurs by discharge, either directly, by molecule fragments, or via excitation of the atmosphere in the neighbourhood of the separated surface: the blue glow coming from adhesive tapes being unrolled is emitted from nitrogen molecules of the air that have been excited by the electric discharge.

Thermoluminescence. Thermoluminescence means not temperature radiation but enhancement of the light emission of materials already excited electronically by the application of heat. The phenomenon is observed with some minerals and, above all, with crystal phosphors after they have been excited by light.

Photoluminescence. Photoluminescence, which occurs by virtue of electromagnetic radiation falling on matter, may range from visible light through ultraviolet, X-ray, and gamma radiation. It has been shown that, in luminescence caused by light, the wavelength of emitted light generally is equal to or longer than that of the exciting light (*i.e.*, of equal or less energy). As explained below, this difference in wavelength is caused by a transformation of the exciting light, to a greater or lesser extent, to nonradiating vibration energy of atoms or ions. In rare instances—*e.g.*, when intense irradiation by laser beams is used or when sufficient thermal energy contributes to the electron excitation process—the emitted light can be of shorter wavelength than the exciting light (anti-Stokes radiation).

The fact that photoluminescence can also be excited by ultraviolet radiation was first observed by a German physicist, Johann Wilhelm Ritter (1801), who investigated the behaviour of phosphors in light of various colours. He found that phosphors luminesce brightly in the invisible region beyond violet and thus discovered ultraviolet radiation. The transformation of ultraviolet light to visible light has much practical importance.

Gamma rays and X-rays excite crystal phosphors and other materials to luminescence by the ionization process (*i.e.*, the detachment of electrons from atoms), followed by a recombination of electrons and ions to produce visible light. Advantage of this is taken in the fluoroscope used in X-ray diagnostics and in the scintillation counter that detects and measures gamma rays directed onto a phosphor disk that is in optical contact with the face of a photomultiplier tube (a device that amplifies light signals).

Electroluminescence. Like thermoluminescence, the term electroluminescence includes several distinct phenomena, a common feature of which is that light is emitted by an electrical discharge in gases, liquids, and solid materials. Benjamin Franklin, in the United States, for example, in 1752 identified the luminescence of lightning as caused by electric discharge through the atmosphere.

Electric discharge

An electric-discharge lamp was first demonstrated in 1860 to the Royal Society of London. It produced a brilliant white light by the discharge of high voltage through carbon dioxide at low pressure. Modern fluorescent lamps are based on a combination of electroluminescence and photoluminescence: mercury atoms in the lamp are excited by electric discharge, and the ultraviolet light emitted by the mercury atoms is transformed into visible light by a phosphor (see LIGHTING AND LIGHTING DEVICES).

The electroluminescence sometimes observed at the electrodes during electrolysis is caused by the recombination of ions (therefore, this is a sort of chemiluminescence). The application of an electric field to thin layers of luminescing zinc sulfide can produce light emission, which is also called electroluminescence.

A great number of materials luminesce under the impact of accelerated electrons (once called cathode rays); *e.g.*, diamond, ruby, crystal phosphors, and certain complex salts of platinum. The first practical application of cathodoluminescence was in the viewing screen of an oscilloscope tube constructed in 1897; similar screens, employing improved crystal phosphors, are used in television, radar, oscilloscopes, and electron microscopes.

The impact of accelerated electrons on molecules can

produce molecular ions, ions of molecule fragments, and atomic ions. In gas-discharge tubes these particles were first detected as "canal rays" or anode rays. They are able to excite phosphors but not as efficiently as electrons can.

Radioluminescence. Radioactive elements can emit alpha particles (helium nuclei), electrons, and gamma rays (high-energy electromagnetic radiation). The term radioluminescence, therefore, means that an appropriate material is excited to luminescence by a radioactive substance. When alpha particles bombard a crystal phosphor, tiny scintillations are visible to microscopic observation. This is the principle of the device used by an English physicist, Ernest Rutherford, to prove that an atom has a central nucleus. Self-luminous paints, such as are used for dial markings for watches and other instruments, owe their behaviour to radioluminescence. These paints consist of a phosphor and a radioactive substance, *e.g.,* tritium or radium. An impressive natural radioluminescence is the aurora borealis: by the radioactive processes of the sun, enormous masses of electrons and ions are emitted into space in the solar wind. When they approach the Earth, they are concentrated by its geomagnetic field near the poles. Discharge processes of the particles in the upper atmosphere yield the famous luminance of the auroras.

LUMINESCENT MATERIALS AND PHOSPHOR CHEMISTRY

The first phosphor synthesized was probably an impure barium sulfide preparation with very low luminance efficiency and with the serious shortcoming that it was rather quickly decomposed in moist air, yielding hydrogen sulfide. A more stable sulfide-type phosphor was produced in 1866 by heating zinc oxide in a stream of hydrogen sulfide. In 1887 it became known that these sulfides do not luminesce in a chemically pure state but only when they contain small quantities of a so-called activator metal. Later, other materials, such as certain metal oxides, silicates, and phosphates, were found to luminesce if they were prepared by special procedures.

Sulfide-type phosphors, activators, fluxes. The sulfides of zinc and of cadmium are the most important basic materials of sulfide-type phosphors. An important condition of getting highly efficient phosphors is that these sulfides must first be prepared to the highest possible chemical purity before the necessary amount of activator can be added precisely. The emission of zinc sulfide can be shifted to longer wavelengths by increasing substitution of the zinc ions by cadmium ions. Zinc sulfide and cadmium sulfide phosphors are especially efficient in electroluminescence.

Sulfide-type phosphors are produced from pure zinc or cadmium sulfide or their mixtures by heating them together with small quantities (0.1–0.001 percent) of copper, silver, gallium, or other salts (activators) and with about 2 percent of sodium or another alkali chloride at about 1,000° C (1,832° F). The role of the alkali halides is to facilitate the melting process and, above all, to serve as coactivators (fluxes). Only small quantities of the alkali halide are integrated into the phosphor, but this small quantity is highly important for its luminescence efficiency. Copper-activated zinc and cadmium sulfides exhibit a rather long afterglow when their irradiation has ceased, and this is favourable for application in radar screens and self-luminous phosphors.

Oxide-type phosphors. Certain oxide-type minerals have been found to luminesce when irradiated. In some of them, activators must first be introduced into the crystal. Examples are ruby (aluminum oxide with chromium activator—bright-red emission) and willemite (zinc orthosilicate with manganese activator—green emission). On the other hand, scheelite (calcium tungstate) emits a blue luminescence without activator. All of these minerals have been made synthetically, with remarkably higher efficiencies than those that occur naturally. Silicates, borates, and phosphates of the second group of the periodic table of elements, such as zinc silicate, zinc beryllium silicate, zinc and cadmium borates, and cadmium phosphates, become efficient phosphors when activated with manganese ions, emitting in the red to green region of the spectrum. They have been incorporated into colour television screens to emit the colours blue (silver-activated zinc sulfide), green (manganese-activated zinc orthosilicate), and red (europium-activated yttrium vanadate).

Synthetic phosphors

Centres, activators, coactivators, poisons. The study of phosphor chemistry has yielded a detailed picture of the role of the above-mentioned activators and fluxes. Philipp Anton Lenard, a physicist in Germany, was the first (1890) to describe activator ions as being distributed in zinc sulfide and other crystalline materials that serve as the host crystal. The activator ions are surrounded by host-crystal ions and form luminescing centres where the excitation–emission process of the phosphor takes place. These centres must not be too close together within the host crystal lest they inactivate each other. For high efficiency, only a trace of the activator may be inserted into the host crystal, and its distribution must be as regular as possible. In high concentration, activators act as "poisons" or "killers" and thus inhibit luminescence. The term killer is used especially for iron, cobalt, and nickel ions, whose presence, even in small quantities, can inhibit the emission of light from phosphors.

Killers

Phosphors, such as calcium tungstate or zinc sulfide, that need no activator appear to have their luminescing centres in special groups of atoms different from the symmetry of their own crystal lattice, such as the group WO_4 in the compound calcium tungstate ($CaWO_4$), or, similarly, the SiO_4 group in zinc orthosilicate, (Zn_2SiO_4). That luminescing properties of a centre are strongly dependent on the symmetry of neighbouring ion groups with respect to the whole phosphor molecule is clearly proved by the spectral shifts of certain phosphors activated with lanthanide ions, which emit in narrow spectral regions. Because of this altering effect on the symmetry of luminescing centres, small quantities (about 0.2 percent) of titania incorporated in zinc orthosilicate give a remarkable increase in luminescence. Titania is called an intensifier activator because it increases the host-crystal luminescence, whereas a substance that produces luminescence not exhibited by the chemically pure host crystal is called an originative activator.

The fluxes (*e.g.,* sodium chloride) act as coactivators by facilitating the incorporation of activator ions. Copper ions, for instance, are used as activators of zinc chloride phosphors and are usually introduced in the copper(II), or cupric, form (the Roman numeral indicates the oxidation state; that is, I means that the element has one electron involved in a chemical bond and II that it has two electrons involved; the larger oxidation state is indicated by the *-ic* ending and the smaller by the *-ous* ending). If a copper(II) compound is incorporated into the zinc sulfide by heating, copper(I) sulfide (or cuprous sulfide, formula Cu_2S) will be produced with crystals that will not fit into the host-crystal zinc chloride because their form is so different, and only a relatively few luminescent centres will be possible. On the other hand, if a coactivator such as sodium chloride is introduced along with the copper(II) salt, the copper(II) ions are reduced to form copper(I) chloride (or cuprous chloride, formula CuCl) crystals with the same structure as the host crystal. Thus, many luminescent centres will be produced, and strong activation will result.

In describing a luminescent phosphor, the following information is pertinent: crystal class and chemical composition of the host crystal, activator (type and percentage), coactivator (intensifier activator), temperature and time of crystallization process, emission spectrum (or at least visual colour), and persistence. A few phosphors and their activators are listed in Table 2.

Organic luminescent materials. Although the inorganic phosphors are industrially produced in far higher quantities (several hundred tons per year) than the organic luminescent materials, some types of the latter are becoming more and more important in special fields of practical application. Paints and dyes for outdoor advertising contain strongly fluorescing organic molecules such as fluorescein, eosin, rhodamine, and stilbene derivatives. Their main shortcoming is their relatively poor stability in light, because of which they are used mostly when durability is not required. Organic phosphors are used as optical brighteners for invisible markers of laundry, banknotes, identity cards, and stamps and for fluorescence microscopy of tis-

Table 2: Visual Properties of Some Luminescent Materials

phosphor (phosphor/activator; coactivator)	emission	
	colour*	persistence
rhombohedral Zn_2SiO_4/Mn 0.3% (rhombohedral zinc orthosilicate/manganese 0.3%); 1,200° C 60 min, slow cooling	green (525 nm)	short (0.01 sec)
β-Zn_2SiO_4/Mn 0.3% (beta zinc orthosilicate/manganese 0.3%); 1,600° C 10 min, quench cooling	yellow (610 nm)	short (0.01 sec)
cubic ZnS/Cu 0.03%; Cl (cubic zinc sulfide/copper 0.03%; chloride); 950° C 10 min, slow cooling	green-blue (516 nm)	long (hours)
hexagonal ZnS/Cu 0.03%; Cl (hexagonal zinc sulfide/copper 0.03%; chloride); 1,250° C 10 min, slow cooling	green (528 nm)	very long (up to 24 hours)

*The wavelengths of the respective emission maxima are given in parentheses. 1 nanometre (nm) = 10^{-9} metre = 1 millimicron = 10 angstroms.

sues in biology and medicine. Their "invisibility" is due to the fact that they absorb practically no visible light. The fluorescence is excited by invisible ultraviolet radiation (black light).

Photoradiation in gases, liquids, and crystals. When describing chemical principles associated with luminescence, it is useful, at first, to neglect interactions between the luminescing atoms, molecules, or centres with their environment. In the gas phase these interactions are smaller than they are in the condensed phase of a liquid or a solid material. The efficiency of luminescence in the gas phase will be far greater than in the condensed phases because in the latter the energy of the electrons excited by photons or by chemical-reaction energy can be dissipated as thermal, nonradiative energy by collision of the atoms or by the rotational and vibrational energy of the molecules. This effect has to be taken into account even more when the radiation of single atoms is compared with that of multiatomic molecules. For molecules, radiative (electronic-excitation) energy is internally converted to vibrational energy; that is, there are radiationless transitions of electrons in atoms. This is the explanation for the fact that only a relatively small number of compounds are able to exhibit efficient luminescence. In crystals, on the other hand, the binding forces between the ions or atoms of the lattice are strong compared with the forces acting between the particles of a liquid, and electron-excitation energy, therefore, is not as easily transformed into vibrational energy, thus leading to a good efficiency for radiative processes.

Effect of binding force

LUMINESCENCE PHYSICS

Mechanism of luminescence. The emission of visible light (that is, light of wavelengths between about 690 nanometres and 400 nanometres, corresponding to the region between deep red and deep violet) requires excitation energies the minimum of which is given by Einstein's law stating that the energy (E) is equal to Planck's constant (h) times the frequency of light (ν), or Planck's constant times the velocity of light (c) in a vacuum divided by its wavelength (λ); that is,

$$E = h\nu = \frac{hc}{\lambda}.$$

The energy required for excitation therefore ranges between 40 kilocalories (for red light), about 60 kilocalories (for yellow light), and about 80 kilocalories (for violet light) per mole of substance. Instead of expressing these energies in kilocalories, electron volt units (one electron volt = 1.6×10^{-12} erg; the erg is an extremely small unit of energy) may be used, and the photon energy thus required in the visible region ranges from 1.8 to 3.1 electron volts.

The excitation energy is transferred to the electrons responsible for luminescence, which jump from their ground-state energy level to a level of higher energy. The energy levels that electrons can assume are specified by quantum mechanical laws. The different excitation mechanisms considered below depend on whether or not the excitation of electrons occurs in single atoms, in single molecules, in combinations of molecules, or in a crystal. They are initiated by the means of excitation described above: impact of accelerated particles such as electrons, positive ions, or photons. Often, the excitation energies are considerably higher than those necessary to lift electrons to a radiative level; for example, the luminescence produced by the phosphor crystals in television screens is excited by cathode-ray electrons with average energies of 25,000 electron volts. Nevertheless, the colour of the luminescent light is nearly independent of the energy of the exciting particles, depending chiefly on the excited-state energy level of the crystal centres.

Electron excitation

Electrons taking part in the luminescence process are the outermost electrons of atoms or molecules. In fluorescent lamps, for example, a mercury atom is excited by the impact of an electron having an energy of 6.7 electron volts or more, raising one of the two outermost electrons of the mercury atom in the ground state to a higher level. Upon the electron's return to the ground state, an energy difference is emitted as ultraviolet light of a wavelength of 185 nanometres. A radiative transition between another excited state and the ground-state level of the mercury atom produces the important ultraviolet emission of 254-nanometre wavelength, which, in turn, can excite other phosphors to emit visible light. (One such phosphor frequently used is a calcium halophosphate incorporating a heavy-metal activator.)

This 254-nanometre mercury radiation is particularly intensive at low mercury vapour pressures (around 10^{-5} atmosphere) used in low-pressure discharge lamps. About 60 percent of the input electron energy may thus be transformed into near-monochromatic ultraviolet light; *i.e.*, ultraviolet light of practically one single wavelength.

Whereas at low pressure there are relatively few collisions of mercury atoms with each other, the collision frequency increases enormously if mercury gas is excited under high pressure (*e.g.*, eight atmospheres or more). Such excitation leads not only to collisional de-excitation of excited atoms but also to additional excitation of excited atoms. As a consequence, the spectrum of the emitted radiation no longer consists of practically one single, sharp spectral line at 254 nanometres, but the radiation energy is distributed over various broadened spectral lines corresponding to different electronic energy levels of the mercury atom, the strongest emissions lying at 303, 313, 334, 366, 405, 436, 546, and 578 nanometres. High-pressure mercury lamps can be used for illumination purposes because the emissions from 405 to 546 nanometres are visible light of bluish-green colour; by transforming a part of the mercury line emission to red light by means of a phosphor, white light is obtained.

When gaseous molecules are excited, their luminescence spectra show broad bands; not only are electrons lifted to levels of higher energy but vibrational and rotational motions of the atoms as a whole are excited simultaneously. This is because vibrational and rotational energies of molecules are only about 10^{-2} and 10^{-4}, respectively, those of the electronic transition energies, and these many energies can be added to the energy of a single electronic transition, which is represented by a multitude of slightly different wavelengths making up one band. In larger molecules, several overlapping bands, one for each kind of electronic transition, can be emitted. Emission from molecules in solution is predominantly bandlike caused by interactions of a relatively great number of excited molecules with molecules of the solvent. In molecules, as in atoms, the excited electrons generally are outermost electrons of the molecular orbitals.

The terms fluorescence and phosphorescence can be used here, on the basis not only of the persistence of luminescence but also of the way in which the luminescence is produced. When an electron is excited to what is called, in spectroscopy, an excited singlet state, the state will have a lifetime of about 10^{-8} second, from which the excited electron can easily return to its ground state (which normally is a singlet state, too), emitting its excitation energy

as fluorescence. During this electronic transition the spin of the electron is not altered; the singlet ground state and the excited singlet state have like multiplicity (number of subdivisions into which a level can be split). An electron, however, may also be lifted, under reversal of its spin, to a higher energy level, called an excited triplet state. Singlet ground states and excited triplet states are levels of different multiplicity. For quantum mechanical reasons, transitions from triplet states to singlet states are "forbidden," and, therefore, the lifetime of triplet states is considerably longer than that of singlet states. This means that luminescence originating in triplet states has a far longer duration than that originating in singlet states: phosphorescence is observed.

Phosphorescence

The interactions of a large number of atoms, ions, or molecules are greater still in solution and in solids; to obtain a narrowing of the spectral band, subzero temperatures (down to that of liquid helium) are applied in order to reduce vibrational motions. The electronic energy levels of crystals such as zinc sulfide and other host crystals used in phosphors form bands: in the ground state practically all electrons are to be found on the valence band, whereas they reach the conduction band after sufficient excitation (see MATTER). The energy difference between the valence band and the conduction band corresponds to photons in the ultraviolet or still shorter wavelength region. Additional energy levels are introduced by activator ions or centres bridging the energy gap between valence band and conduction band, and, when an electron is transferred from the valence band to such an additional energy level by excitation energy, it can produce visible light on return to the ground state. A rather close analogy exists between the forbidden transitions of certain excited molecular electronic states (triplet–singlet, leading to phosphorescence) and the transition of an electron of an inorganic phosphor kept in a trap: traps (certain distortions in the crystal lattice) are places in the crystal lattice where the energy level is lower than that of the conduction band, and from which the direct return of an electron to the ground state is also forbidden.

When a solid is bombarded by photons or particles, the excitation of the centres can occur directly or by energy transfer. In the latter case, excited but nonluminescing states are produced at some distance from the centre, with the energy moving through the crystal in the form of excitons (ion-electron pairs) until it approaches a centre where the excitation process can occur. This energy transfer can also be realized by radiation in inorganic phosphors containing two activators, as well as in solutions of organic molecules.

Spontaneous and stimulated emission. The radiative return of excited electrons to their ground state occurs spontaneously, and when there exists an assembly of excited electrons their individual spontaneous radiative transitions are independent of each other. Therefore, the luminescence light is incoherent (the emitted waves are not in phase with each other) in this case. Sometimes the emission of luminescence can be stimulated by irradiation with photons of the same frequency as that of the emitted light; such stimulated transitions are used in lasers, which produce very intensive beams of coherent monochromatic light.

Exponential decay

The spontaneous luminescent emission follows an exponential law that expresses the rate of intensity decay and is similar to the equation for the decay of radioactivity and some chemical reactions. It states that the intensity of luminescent emission is equal to an exponential value of minus the time of decay divided by the decay time, or $L = L_0 \exp(-t/\tau)$, in which L is the intensity of emission at a time t after an initial intensity L_0, and τ is the decay time of the luminescence; that is, the time in which the assembly of the excited atoms would decrease in luminescence intensity to a value of 0.368 L_0.

When excited atoms of the centres are in contact with other atoms, as is the case in condensed phases (liquids, solids, in gases of not-too-low pressure), part of the excitation energy will be transformed into heat by collisional deactivation (thermal quenching). The decay time, therefore, has to be replaced by an effective excited-state lifetime, resulting in a more complicated exponential decay law that depends on the collision frequency, the energy imparted to the excited atoms of the centre that causes the transfer of excitation energy into heat (activation energy), a constant, and the temperature of the luminescent material. This law describes the actual luminescence decay of a great number of luminescent materials; *e.g.*, calcium tungstate.

Increase of activation energy for nonradiative deactivation of excited-centres luminescence decay can be achieved by changing the host crystal or by electron traps. The traps are imperfections in the crystal lattice where electrons are captured after they have been ejected from a luminescent centre by excitation energy. That the luminescent properties of phosphor centres are strongly dependent on the chemical nature of the host crystal may be seen in Table 3, showing that the same activator ions (manganese ions with two positive charges, indicated as Mn^{2+}, or Mn[II]), in different host crystals yield remarkably different-coloured emissions and decay times (measured in fractions of a second).

Prolonging the emission time of phosphors up to days or even longer (production of phosphorescence of the phosphors) is possible by inserting traps into the host crystal. Trapped electrons cannot return directly to the centre. In order to be released from the traps they must first obtain additional thermal energy—in this case, thermal energy stimulates luminescence—after which they recombine with a centre and undergo radiative transition. Trapping in crystals has its analogy to forbidden transitions in molecules (triplet–singlet transitions) or in radiation processes from metastable atomic-energy levels.

Table 3: Influence of Host Crystal on the Lifetime and Emission Colour of the Excited Phosphors

host crystal	activator	time (second)	emission colour
Tetragonal zinc fluoride, ZnF_2	manganese(II)	0.1	orange
Rhombic cadmium sulfate, $CdSO_4$	manganese(II)	0.05	orange
Rhombic magnesium sulfate, $MgSO_4$	manganese(II)	0.03	red
Rhombic zinc phosphate, $Zn_3(PO_4)_2$	manganese(II)	0.02	red
Cadmium silicate, $CdSiO_3$	manganese(II)	0.019	orange
Zinc orthosilicate, Zn_2SiO_4	manganese(II)	0.018	yellow
Cadmium pyroborate, $Cd_2B_2O_5$	manganese(II)	0.015	red orange
Rhombohedral zinc orthosilicate, Zn_2SiO_4	manganese(II)	0.013	green
Rhombohedral zinc germanate, Zn_2GeO_4	manganese(II)	0.0105	green yellow
Cubic zinc aluminate, $ZnAl_2O_4$	manganese(II)	0.0055	blue green
Cubic zinc gallate, $ZnGa_2O_4$	manganese(II)	0.0043	green blue
Hexagonal zinc sulfide, ZnS	manganese(II)	0.0004	orange

An example of a practical application of stimulated emission of a phosphor with trapped electrons is cubic strontium sulfide/selenide activated with samarium and europium ions, the coactivators being strontium sulfate and calcium fluoride. This phosphor has been used in devices for viewing scenes at night by reflected infrared light emitted by infrared lamps. The traps in this phosphor have been identified as samarium ions, whereas europium ions are the active ions in the centres. The phosphor is first excited by photons of about three electron volts (blue light), which results in an ejection of an electron from a europium ion (Eu^{2+}) centre. This excited electron is trapped by a triply charged samarium ion (Sm^{3+}), which is transferred to a doubly charged samarium ion (Sm^{2+}). Heat or irradiation by infrared photons releases one electron from the doubly charged samarium ion (Sm^{2+}). The electron is then recaptured from a triply charged europium ion (Eu^{3+}), yielding an excited doubly charged europium ion (Eu^{2+}), which returns to its ground state by emitting a photon of 2.2 electron volts energy (yellow light). The trap depth of this phosphor (*i.e.*, the energy required for release of an electron from it) is large compared to the thermal

energy of the lattice of the host crystal, and, therefore, the lifetime of the traps at room temperature is many months long. Bombarding this phosphor with photons of energy higher than that of infrared photons but not sufficient for excitation can lead to photoquenching: the traps are emptied far more rapidly, and thermal deactivation of the centres is enhanced.

When iron, cobalt, or nickel ions are present in a phosphor, an excited electron can be captured by these ions. The excitation energy is then emitted as infrared photons, not as visible light, so that luminescence is quenched. These ions, therefore, are called killers—the killing process being opposite to stimulation.

In chemiluminescence, such as the oxidation of luminol, light emission depends not only on radiative and quenching or intramolecular deactivation processes but also on the efficiency of the chemical reaction leading to molecules in an electronically excited state.

In bioluminescence reactions, the production of electronically excited molecules, as well as their radiative transitions back to their ground state, is efficiently catalyzed by the enzymes acting here, and bioluminescence light output is therefore high.

The luminescence photons emitted by one kind of excited atom, molecule, or phosphor can excite another to emit its specific luminescence: this type of energy transfer is observed with inorganic as well as organic substances. Thus, excited benzene molecules can excite naphthalene molecules by radiative-energy transfer. The radiation produced by the luminol chemiluminescence can produce fluorescence when fluorescein is added to the reaction mixture. In most of these cases the acceptor molecules have luminescent electrons with energy levels lower than those of the primary excited molecules, and emitted secondary luminescence is therefore of longer wavelength than the primary. Practical application of this phenomenon, called cascading, is used in radar kinescopes, which have composite fluorescent screens consisting of a layer of blue-emitting zinc sulfide/silver (chloride) phosphor—the hexagonal crystal, ZnS/Ag(Cl) deposited on a layer of yellow-emitting zinc or cadmium sulfide/copper [chloride] phosphor [the hexagonal crystal, (Zn,Cd)S/Cu (Cl)].

Cascading

The cathode-ray electrons excite the blue-emitting phosphor, whose photons, in turn, excite the yellow-emitting phosphor, which has traps with a decay time of about 10 seconds. Excitation of the blue-emitting phosphor alone would be unfavourable, as the sharply focussed cathode rays are absorbed by the blue phosphor to a small extent only, and its decay time is too short; also, direct excitation of the yellow-emitting phosphor alone would yield poor efficiency because the traps are emptied too rapidly by the heat produced by the relatively high-energetic electron impact.

Another energy-transfer mechanism is referred to as sensitization: a calcium carbonate phosphor (rhombohedral $CaCO_3$/Mn), for example, emits orange light under cathode-ray irradiation but is not excited by the 254-nanometre emission of mercury atoms, whereas this emission produces the same orange light with calcium carbonate (rhombohedral $CaCO_3$) activated by manganese and lead ions. This is not cascade luminescence: a mechanical mixture of a manganese and a lead-activated calcium carbonate exhibits no emission under ultraviolet radiation. In a phosphor containing both activators, the lead ions act as sensitizers in introducing an additional excitation band into the system from which the manganese ions get their excitation energy in a nonradiative energy transfer. Similar sensitization is observed in gases and in liquids.

Solid state energy states. The complicated problems concerning the energy states in solids of a luminescence centre are commonly visualized by adapting the energy-level diagram used in describing energy transitions in an isolated diatomic molecule (Figure 27).

In this diagram, the potential energy of a centre is plotted as a function of the average distance ($\bar{x}$) between the atoms: $\bar{x}^*$ represents the ground state and $\bar{x}_0$ represents the lowest excited state of the centre. In a tetrahedral permanganate-ion centre (MnO_4), for example $\bar{x}$ would be the average distance between the central manganese ion and an oxygen ion in any of the corners of the tetrahedron.

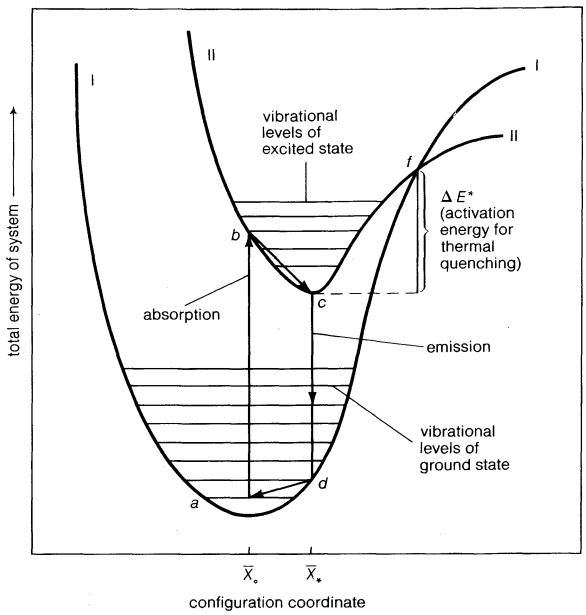


Figure 27: Energy levels of a luminescent centre (see text).

At a temperature of absolute zero the ground-state energy level is near the bottom of curve I at the minimum amplitude of atomic vibration. At room temperature (300 K [81° F]) the ground state lies higher, at *a*, where the centre has considerable vibrational energy. When an electron of the centre is excited, it is lifted to the higher energy level at *b* in curve II. This electronic transition occurs far more rapidly than the readjustment of the atoms of the centre, which then occurs within a time of about 10^{-12} second to reach the minimum vibrational level at *c*. The energy difference ($b - c$) is dissipated as heat in the host-crystal lattice. From the excited-state level *c*, the electron can return to the ground-state level *d* shown in Curve I, the liberated energy being emitted as a photon.

The last step is a readjustment of the centre to *a*, the energy difference ($d - a$) again being dissipated as heat. Nonradiative transition of the excited electron back to its ground state occurs when the electron is excited to an energy level above the intersection point *f* of the ground-state and the excited-state energy curve. This is caused mainly by increasing the vibrations of the lattice by application of higher temperatures. The energy difference ($f - c$) is equal to the activation energy already mentioned, and therefore most centres become increasingly nonradiative at higher temperatures. In trap-type phosphors the temperature must be sufficiently high, of course, to eject the electron from the traps.

In some phosphors—calcium tungstate ($CaWO_4$), for example—absorption and emission of the exciting energy appear to take place mainly in the same centre; the excited electron remains near the centre. Such phosphors do not exhibit photoconductivity because only a few excited electrons succeed in reaching the conduction band where they are freely mobile. The luminescence decay is exponential.

Photoconducting phosphors

Zinc sulfide phosphors, however, are photoconducting, which means that many excited electrons are lifted to the conduction band of the host crystal. The energy levels of different centres and of the host-crystal lattice have to be taken into account simultaneously.

The relative levels of the zinc sulfide valence band (ground state of the host-crystal lattice) and the conduction band (excited state of the host-crystal lattice), of activator levels and of trap levels are shown in Figure 28. Points 1, 2, 3, and 4 represent one situation in a host crystal, and points 5, 6, 7, 8, 9, and 10 represent another situation.

The activator ions introduce additional ground-state levels and excited-state levels of energies between those of the valence and the conduction band of the zinc sulfide. When the excitation energy is sufficiently high, an electron is raised to the conduction band ($1 \rightarrow 2$, $5 \rightarrow 6$, corre-

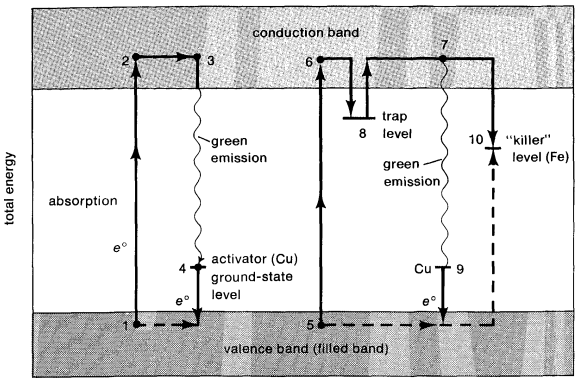


Figure 28: Transition of an electron from the valence band to the conduction band by light absorption (see text).

sponding to the ionization continuum in a gas). It moves away from the centre (2 → 3, 6 → 8) and may either be trapped by an imperfection of the lattice (8) or return to an ionized centre (activator), in which it first occupies an excited level (3 → 4) and then drops to the ground state of the activator centre by emitting a photon. An activator centre that captures such an excited electron has already lost one of its own electrons to a positive hole (electron vacancy) in the host-crystal lattice.

The energetic level of the traps is about 0.25 electron volt beneath the conduction-band level. A trapped electron (8) must be raised to the conduction band by thermal energy before a recombination with an ionized activator centre can occur. The green emission (530 nanometres) of the zinc sulfide phosphor (ZnS/Cu) is explained by the recombination of an electron from the conduction band and a copper ion in an activator centre (7 → 9); the blue emission (463 nanometres) is due to recombination of the excited electron and a copper ion in an interstitial place.

Direct excitation of the activator centres is also possible. When an electron recombines with a killer ion (10), no visible emission occurs.

In solid-state electroluminescence, the radiative processes occurring in a phosphor under irradiation are produced by applying external electric fields of several hundred volts, alternating at several thousand cycles per second. Special preparations of zinc sulfide (hexagonal ZnS), with an iodine coactivator and high concentrations of a copper activator, are embedded in a thin layer of about 0.01 centimetre (0.004 inch) of insulating organic material or glass which is mounted between the electrodes.

High luminescence efficiencies result. Application of a direct-current field yields luminescence in crystals of gallium arsenide (GaAs), silicon carbide (SiC), cadmium sulfide (CdS), and zinc monocrystals of sulfide with copper activator (ZnS/Cu); the cathode injects electrons into the conduction band, whereas the anode removes electrons.

Efficiency of luminescence; luminance. The efficiency of luminescence emission must be regarded on an energy and a quantum basis. When every exciting photon yields an emitted photon of the same energy (as is the case for resonance excitation—*i.e.*, excitation of fluorescence by a monochromatic light of exactly the same wavelengths as the resulting fluorescence—and radiation of isolated atoms in dilute gases), the luminescence efficiency is 100 percent with respect to input energy as well as to the number of quanta. When the number of secondary photons is equal to that of the primary, but their energy is less because some energy is dissipated as heat, the quantum efficiency is 100 percent, but the luminescence efficiency is less than 100 percent. The quantum efficiency of most luminescences is far lower than 100 percent; zinc sulfide phosphors have about 20 percent efficiency, and solid state electroluminescence is less than 10 percent efficient.

Quantum efficiency

In chemiluminescence, the quantum efficiency of "brilliant" reactions, such as luminol, is about one percent and up to 23 percent in the oxalate chemiluminescence. Solid state electroluminescence, or electroluminescence of gases excited by high-frequency electric fields, is usually less than 10 percent.

The light intensity of luminescent processes depends chiefly on the excitation intensity, the density, and the lifetime of the radiative atoms, molecules, or centres. For practical purposes this luminous intensity per unit area is called photometric brightness or luminance of a material and is measured in lambert or millilambert (0.001 lambert) units (one lambert is equal to one candle per square centimetre divided by π). The luminance of some luminescent light sources is listed in Table 4. (K.-D.G.)

Table 4: Luminance of Some Luminescent Light Sources

sources	lamberts
Fluorescent lamps	
20-watt T 12 standard warm white	1.67
40-watt T 12 standard warm white	2.1
96-watt T 12 standard warm white	2.052
Television screens	0.10
Fresh snow in full sunlight	10.
Sun (observed from Earth's surface at meridian)	519,000.
Lasers	exceeding that of Sun

BIBLIOGRAPHY

Comprehensive works: A.C.S. VAN HEEL and C.H.F. VELZEL, *Wat is licht?* (1968; Eng. trans., *What Is Light?,* 1968), is a good book for the general reader; numerous illustrations are accompanied by a simple text. M. MINNAERT, *Licht en kleur in het landschap* (1937; Eng. trans., *The Nature of Light and Colour in the Open Air,* new ed., 1959), also recommended for the general reader, deals with optical effects such as rainbows, mirages, and many other effects of light and colour. MAX BORN and EMIL WOLF, *Principles of Optics,* 3rd rev. ed. (1965), is a comprehensive treatise on the wave theory of light, particularly strong on diffraction, scattering, polarization, and coherence—suitable for the reader with advanced mathematical knowledge. ROBERT W. DITCHBURN, *Light,* 2nd ed. (1963), is a text at the university level that includes both theory and experimental evidence in relation to wave phenomena and quantum optics. K.D. FROOME and LOUIS ESSEN, *The Velocity of Light and Radio Waves* (1969), is a critical review of modern methods of measuring the constant *c.* JOHN H. SANDERS, *The Velocity of Light* (1965), contains reprints of original papers by A.A. Michelson, F.G. Pease, F. Pearson, Louis Essen, A.C. Gordon-Smith, and E. Bergstrand, together with a review of other work on the subject. RICHARD MORRIS, *Light* (1979), a general introduction with good bibliographies. For a history of early theories, see A.I. SABRA, *Theories of Light: From Descartes to Newton* (1981).

Quantum mechanics and electromagnetic theory: LEONARD I. SCHIFF, *Quantum Mechanics,* 2nd ed. (1955), is a systematic exposition of the physical concepts of quantum mechanics, including a description of the electromagnetic field and its interactions with matter. JOSEPH NEEDHAM, *Science and Civilisation in China,* vol. 4 (1962), includes an authoritative account of the study of light in ancient China and a detailed comparison with parallel developments due to Greek and Arab philosophers. EDMUND T. WHITTAKER, *A History of Theories of Aether and Electricity,* rev. ed., vol. 1 (1951), deals with the background to the development of the electromagnetic theory of light in the 19th century. FRANK H. READ, *Electromagnetic Radiation* (1980), includes lucid explanations of the basic concepts.

Luminescence: E.N. HARVEY, *A History of Luminescence: From the Earliest Times Until 1900* (1957), a classical work dealing rather extensively with the historical development of the different types of luminescence, especially bioluminescence and chemiluminescence. E.J. BOWEN (ed.), *Luminescence in Chemistry* (1968), a textbook for students and research workers; C.A. PARKER, *Photoluminescence of Solutions* (1968), a detailed explanation of the basic principles of luminescence as applied to photoluminescence in solutions, kinetics, apparatus, and analytical applications; M. ZANDER, *Phosphorimetry* (1968), the first modern monograph dealing exclusively with phosphorescence of organic materials, with a complete bibliography; G.G. GUILBAULT (ed.), *Fluorescence: Theory, Instrumentation, and Practice* (1967), a fairly technical account written by outstanding specialists in their respective fields (good background knowledge is necessary); D.M. HERCULES (ed.), *Fluorescence and Phosphorescence Analysis* (1966), chapters of different grades of detail covering the luminescence field; G.F.J. GARLICK, "Luminescence," *Handbuch der Physik,* vol. 26, pp. 1–128 (1958), an extended text (written in English) systematically explaining the physical phenomena and theory of luminescence.

Lighting and Lighting Devices

Lighting in modern times derives almost entirely from electric light sources: lamps with a wire filament in an evacuated glass bulb or fluorescent tubes. A building is lighted throughout to levels comparable with daylight, and selectively placed light is focussed on work of special difficulty. Lighting devices of many kinds have been used through the long ages since man learned to control fire. Evidence of torches has been found at archaeological sites; the cave paintings at Altamira in Spain and the Dordogne in France are perhaps the oldest.

From its beginnings, artificial light has had two distinct purposes: to permit visibility in natural darkness and to create visual effects. The development of lighting technology has been characterized by invention of light sources with greater output and efficiency. With the introduction of each light source, technology for its control has advanced. The candle produced the candelabra that enhances light with prismatic glass adornments. The oil lamp brought in its train the development of various reflectors to concentrate the light. The gas mantle gave rise to the first efforts at precise optical control with prisms and reflectors, the principles of which were later applied to the electric light bulb, which, in turn, because of its safety and cleanliness, permitted the first designs of decorative and architectural lighting fittings and lampshades. The last major advance in light sources, the fluorescent tube, gave rise to the luminous ceiling, a characteristic feature of many modern buildings.

Two important aspects of the development of lighting technology in recent years can be identified. First, lighting technologists and specialists in human vision have collaborated, and a science of lighting and seeing has grown up, based on the idea that artificial lighting must be not only decorative but efficient and comfortable in terms of human sight. Second, the lighting technologist has learned from the architect, who has always concerned himself with natural lighting in buildings but rarely with its artificial counterpart. The lighting engineer has learned about the flow of light, modelling of people and objects, and the active creation of beauty by the interplay of light, form, and colour. Perhaps to these could be added a third development, the recognition by building designers and engineers that large amounts of artificial light give off heat, which ignored can cause discomfort but controlled can provide warmth where needed. Understanding the relationship between light and heat has given rise to the concept of total environmental design. (For information on the physics of light, see LIGHT and OPTICS. The principles of human vision are covered in SENSORY RECEPTION)

This article treats the development of lighting technology, current lighting practice, and the measurement and calculation of lighting. It is divided into the following sections:

Development of lighting technology 29
  The beginnings of artificial lighting 29
    Nonelectrical lighting devices
    Electric lighting devices
  Modern advances in lighting technology 32
    Fixtures and voltage
    Safety devices
    Directing light
    The lighting designer
Current lighting practice 33
  Interior lighting 33
    Dwellings
    Schools
    Offices
    Hospitals
    Factories
  Exterior lighting 35
Measurement and calculation of lighting 36
  Fundamentals of light measurement and control 36
    Illuminance, reflectance, and utilance
    Light distribution
    Glare, or visual comfort, index
    Light-measuring instruments
  Standards of lighting 38

Development of lighting technology

THE BEGINNINGS OF ARTIFICIAL LIGHTING

The earliest form of artificial illumination was the fire used by prehistoric man for warmth and for protection against wild beasts, and flaming brands from the campfire were probably the first torches. From this beginning, thousands of centuries before written history, the torch in various forms has continued as a source of illumination throughout the world.

The flambeau of the Middle Ages, carried by pedestrians at night, was a torch consisting of twisted fibre coated with some flammable substance. These torches provided the earliest form of street lighting, being fastened in metal holders to the outside walls of houses. Pieces of resinous wood, stuck between the stones of the fireplace, gave forth bright but smoky flames for lighting the homes of early New England.

Insect illumination

Unusual sources of artificial illumination are the glowworms, beetles, and fireflies supposedly confined in lanterns in the West Indies or stuck in wax in shallow saucers in Java. Oily carcasses of various animals with wicks drawn through them have been used; the penguin has been used in this manner.

Nonelectrical lighting devices. *Oil lamps.* The earliest known prehistoric lamp dates to the Mousterian Age. This was a hollowed piece of stone, very crude, found at Le Moustier, France, in 1928. Terra-cotta lamps of 8000–7000 BC have been found on the Mesopotamian plains, and Egyptian and Persian copper and bronze lamps of about 2700 BC have been discovered at various times.

By 1000 BC increased lighting efficiency was provided by a wick of vegetable fibre burning in a saucer-type vessel holding olive or nut oil, and by 500–400 BC oil lamps had come into general use. Roman lamps of unglazed and glazed terra-cotta had two or more spouts; and later more elaborate bronze and iron lamps were introduced. About 100 BC the Romans developed the first true lantern of horn, with a perforated top. Vegetable-oil lamps were used by the early Jews, Greeks, and Romans.

The first record of mineral-oil lamps was made about AD 50, when Pliny the Elder mentioned this use of oil found on the Adriatic shores. Japanese history in 615 mentions "burning water," and Marco Polo wrote in the 13th century of Baku oil springs furnishing oil satisfactory for lighting purposes.

Early attempts were made to improve the efficiency of oil lamps. About the latter half of the 1st century AD, Hero of Alexandria devised a lamp in which pressure of brine acting through a column of air raised the oil to the wick. In 1490 Leonardo da Vinci added a cylindrical glass chimney fitted into a glass globe filled with water. The light burned steadily, and through lens action the work surface was brightly lighted for night study.

Leonardo's lamp

Various types of oil lamps were used through the 14th to the 17th century but with little improvement or efficiency. The Betty lamp of the New England Pilgrims (1620), a metal receptacle with a wick lying in a slot and protruding

from one side, was equipped for hanging from mantelpieces or shelves. Fish oil was the chief illuminant used.

In 1784 the Swiss physicist Aimé Argand patented a lamp with round burner and tubular wick and a chimney for directing and regulating the flow of air to the flame. An accident led to the discovery that a round glass chimney reduced flickering, and the Argand lamp marked the greatest advancement in artificial illumination to that date. In 1800 Bertrand G. Carcel added a clockwork pump for raising and feeding oil to the wick; and, because of the steadiness of the light produced, the Argand lamp in this form became a photometric standard. Benjamin Franklin discovered that two wicks side by side a small distance apart gave more light than two single-wick lamps.

Distilled coal oil produced by James Young in 1850–51 was an improvement, but within a few years petroleum provided an inexpensive illuminant of high lighting efficiency, and the crude-oil lamp was practically superseded. In the following 20 years an average of 80 patents per year were granted for various types of improved oil lamps. One of the most important was the duplex burner lamp introduced in 1865 by Joseph Hinks. In 1868 the Wells light combined an oil spray with vaporized oil through the use of pressure, giving a highly luminous flame, the principle used in modern flare lamps. The Kitson lamp of 1885 was a pressure-type lamp using platinum mantles, replaced later by Welsbach mantles (see below).

During the latter half of the 19th century, oil lamps underwent many changes and improvements intended to make them more attractive for domestic use and more efficient for industrial purposes. They were mounted on elaborate bases, and the oil reservoirs were decorated; some were designed for suspension from ceilings, and others were portable. The latter designs carried over into the 20th century, and the kerosene lamp with a flat wick in a perforated metallic holder and a plain glass chimney came into extensive use in rural districts.

Candles. The use of candles dates from the beginning of the Christian era, and candle making is probably one of the oldest industries of mankind. In Africa, oily nuts were burned in clay saucers; later they were strung on twigs, thereby providing continuous illumination.

In AD 100 Greek and Roman true candles were made of flax threads coated with pitch or wax, although the Phoenicians are credited with the first use of wax candles about AD 400. The candle was not in as common use as the oil lamp for several centuries, but it returned during the Middle Ages. Throughout the 16th to the 18th century and until the discovery of petroleum, candles provided the only method of artificial illumination available to people of average means. Their use has been closely related to religious ceremonies and customs throughout history.

The earliest candles were rush piths dipped in tallow. Later, splinters of wood, also dipped in tallow or beeswax and called kindle lights, were used. Homemade candles of the 16th century were of tallow. The whaling industry in the 18th century brought spermaceti into use. The

Spermaceti candles

spermaceti candle, because of its clear, steady flame, was used as a standard measure for artificial light, the term "one candlepower" being based on the light given by a pure spemaceti candle weighing one-sixth of a pound and burning at the rate of 120 grams an hour. The isolation of stearine in 1823 and the development of paraffin in the 1850s furnished improved materials for candles.

Candleholders have assumed various forms. The Prickett chandelier introduced in Italy in 1492 was a multiple-spiked holder mounted in groups around a decorative figure on the wall, with mutton-fat candles forced on the spikes. In the 15th century crystal chandeliers with small cups for holding wax candles lighted homes, ballrooms, and public gathering places. Of these candles, 120 would have been required to equal the light of one 100-watt electric lamp, and the number required for a large room produced much heat. Candles in elaborate chandeliers were used for illumination of the House of Commons in England as late as 1834, and their replacement by gas was strongly opposed by members. With the advent of gas, the use of candles was gradually confined to church ceremonies, decorative purposes, and festive occasions.

Gas light. Ancient records of Egypt and Persia provide numerous instances of spontaneous discharge of combustible gases from fissures in the earth. Natural gas was employed as an illuminant by the Chinese centuries before the modern era. The gas was brought to the surface from beds of rock salt 1,500 to 1,600 feet (450 to 490 metres) deep, conveyed through bamboo pipes and used for lighting salt mines and homes in Szechwan Province.

In 1664 near Wigan, Lancashire, England, John Clayton discovered a pool of natural gas near a coal mine and later described satisfactory experiments in extracting coal gas by distillation. In 1726 Stephen Hales reported that 158 grams of Newcastle coal would yield 180 cubic inches of "inflammable air," and in 1733 Sir James Lowther sent specimens of distilled coal gas to the Royal Society for experimentation.

In 1784 Jean Pierre Minckelers first used gas for lighting. Contemporaneously, in Paris, Philippe Lebon experimented with distillation of gas from various materials for domestic use, and in 1799 Lebon patented a *thermolampe,* using gas distilled from wood.

Unquestionably the major development in artificial lighting during the 19th century was the introduction of methods for generating and distributing coal gas to all buildings in an urban community. The subsequent discovery of natural gas, again in the United States and Canada, reduced the cost to the fortunate citizens within reach of the natural-gas fields but did not materially alter the design of lighting equipment. The first development of coal-gas illumination is credited to William Murdock, a Scottish engineer working in Cornwall, who, in the tradition of many British inventors, tried his experiments on his own family first by lighting his own home by gas in full knowledge of the risk of explosion. No explosion occurred, and, a few years later, at the turn of the century, Murdock was commissioned to install gas lighting in London. The gas lighting in Pall Mall, installed in 1820, is reputed to be one of the oldest extant street-lighting installations in the world. For several decades, gas lighting made use of a simple burner in which the yellow light of the flame itself was the source of illumination. During the 1820s a new form of burner was introduced in which a controlled amount of air was admitted to the gas current, producing a high-temperature but nonluminous flame that heated a refractive, noncombustible material to a very high temperature. This became the source of light: the higher the temperature of the material, the whiter the colour and the greater the light output.

The search for an optimum material that would be durable, infusible, and capable of being repeatedly heated to white heat, culminated during the 1880s in the development of the Welsbach mantle, which, in its final form,

Welsbach gas mantle

consisted of a network woven in cotton, of spherical, cylindrical, or linear form, impregnated with the salts of thorium and cerium. By suitable treatment these mantles could be made sufficiently strong to be transported and retailed. Once they were fitted to the lighting appliance, however, the material that gave them stability burnt off, and a slight jar could destroy the mantle. In spite of this disadvantage, the introduction of the Welsbach gas mantle resulted in gas lighting being employed everywhere within reach of the distributive network of coal- or natural-gas supply. The incandescent light from the thorium–cerium salts in the mantle was not truly white but was tinged a greenish colour with the luminous emissions of these salts, giving rise to the somewhat garish and ghostly appearance characteristic of gaslight. This was unfavourably contrasted with the pleasant, warm glow of the candle, and, as a result, the introduction of the gas mantle was also responsible for a revival in the use of candlelight as a luxury to be indulged in by the wealthy. Improvements in gas lighting itself, particularly by the use of pressure boosted above the safe level for the distribution mains, gave a considerable increase in lighting efficiency.

Coal gas was used only in urban areas in which the cost of distribution was low. This gave rise to the development of forms of local gas generation to confer the advantages of gas lighting in rural areas. Acetylene gas, generated by the dripping of water on calcium carbide, was a particu-

larly valuable development, although its unpleasant smell gave rise to ventilation problems when used in domestic interiors. The natural acetylene flame, however, had a high luminous efficiency and a colour closely approximating natural daylight, which rendered it specially valuable in industrial processes requiring a high standard of fidelity to natural colours.

Electric lighting devices. *The electric lamp.* Despite the successful history of gaslight, modern lighting owes very little to anything that happened before the invention of the incandescent electric lamp about 1870. The incandescent lamp was not the first electric lamp; lighting devices employing an electric arc struck between electrodes of carbon had been developed and were in use earlier. Arc lamps were massive and complicated pieces of equipment that could be installed and maintained only by a skilled engineer. Used for street lighting, arc lamps had advantages, including reasonable reliability, high efficiency, and, above all, a pleasant colour, closely approximating natural sunlight. The light of the arc lamp was particularly kind to the colour of the human complexion and the stone of historic buildings. Partly for this reason, these devices were retained in dignified city settings—for example, in the City of London—long after more modern light sources had come into wide use.

Arc lamps

The importance attached to the invention of the incandescent filament lamp can be gauged by the degree of controversy that has arisen over attribution of its invention. Arguments and claims on both a national and individual basis have been made. While it is true that early lamps of a kind were made in France and Russia and that Sir Joseph Swan of England demonstrated and manufactured carbon-filament lamps of a sophisticated kind before anybody else, the true invention of electric lighting should be credited to Thomas Edison in the United States. To Swan, the lamp was an important technical and scientific development with a utilitarian purpose. To Edison, it was part of a grand scheme of electric lighting, which included the generation and distribution of electric power. The first incandescent lamps, both of Swan and Edison, consisted of a filament of carbon wire in an evacuated glass bulb, two ends of the wire being brought out through a sealed cap and thence to the electric supply. When the supply was connected, the filament glowed and by virtue of the vacuum did not oxidize away quickly as it would have done in air. The commercial success of the lamps depended upon the production of a high degree of vacuum and strength great enough for them to be transported without damage. Though these problems were solved by both Swan and Edison, it was Edison who made the major commercial development. Edison's firm arranged a historic demonstration of electric lighting in Menlo Park, New Jersey, in 1879 and received contracts for generating plant, distribution networks, lamp outlets, and bulbs. It is important to realize that the whole of this concept, not merely the invention of the light bulb, was Edison's contribution. In particular, Edison employed the system of parallel rather than series installation, so that when a bulb failed, as often happened in early days, the rest of the installation did not go out as would have been the case with the series operation first used by Swan. In fact, the distribution system of electric lighting in use today is nothing more than a highly sophisticated version of Edison's original concept. Among early electric lighting installations was that at the Savoy Theatre in London, built for the Gilbert and Sullivan operas.

Edison's distribution system

Of the major improvements in electric lighting that followed Edison's first installations, the most important was development of metallic filaments, particularly the tungsten filament, which is now used almost exclusively in filament lamps. Tungsten was known to be particularly suitable for use in electric lamps, because, of all electrically conducting materials suitable for drawing into filament wires, it had the highest melting point. This meant that lamps could run at a higher temperature and therefore emit both whiter light and more light for the same electrical input than was possible with the less durable and less refractory carbon filament. Tungsten could not be used, however, until a process had been invented for drawing it into thin wire filaments; this development took place in the early 1900s.

From this time forward, light-bulb manufacture became a major industry; much money was spent on research in the physics of light production in the incandescent lamp in order to increase its efficiency, sturdiness, and life in use. Some of the leading technical physicists of the day were engaged by manufacturers, who profited considerably as steady improvements were introduced. The early tungsten lamps, like the carbon lamps, suffered from migration of filament molecules to the glass bulb, causing a blackening of the bulb, a loss in light output, and progressive thinning of the filament until it broke. The introduction of a small amount of inert gas (argon) reduced migration and enabled the filament to be run at a higher temperature, giving a whiter light, higher efficiency, and longer life. Further improvements, such as the development of the coiled filament and then the coiled-coil filament, improved tungsten-filament lamps to the point at which the manufacturer had to decide whether to market a lamp with almost infinite life but low light output, running at a low colour temperature, or to sell the consumer a brighter and more efficient lamp with a relatively short life. The manufacturers settled the matter between themselves by agreeing to market a lamp that would last for an average of 1,000 hours, producing a light of colour considerably yellower than that of natural daylight but still whiter than that of an oil lamp or candle. This was judged to be the most acceptable compromise for the consumer, and in spite of occasional demands for lamps of longer life, little change has taken place over the past 40 years. The most recent development is the introduction of the tungsten–halogen filament lamp, in which the presence of a small amount of iodine reduces the migration of tungsten, permitting a higher running temperature. Because of the chemical effect of the iodine on glass, the filament must be placed in a quartz rather than a glass bulb, increasing initial cost. Though these lamps have whiter colour and longer life, they have not yet been used for general domestic lighting. They are rapidly coming into universal use for automobile headlamps and other forms of light projectors calling for maximum light output from bulbs of minimum size.

Vapour-tube devices. During the late 19th century, many physicists, particularly Sir William Crookes in England, experimented with methods of generating radiation by striking an arc between electrodes in an evacuated tube to which small amounts of elemental vapour had been admitted. The physicist Georges Claude in Paris developed such a tube with neon gas as the filling; when a high voltage was applied to the two electrodes at either end of the tube, it emitted a deep red light whose value for decorative and advertising purposes was quickly recognized. Neon signs soon decorated the exteriors of commercial buildings in the cities of the world; experiments with other vapour fillings—particularly mercury vapour, which gave a blue light—enabled a variety of colours to be produced. The discharge tube, however, had little application in interiors until the development in the 1930s of the fluorescent tube, a long tube with a mercury-vapour filling, the inner wall of which was coated with a material that fluoresced white or near white when subjected to the radiation of the mercury discharge and the quantity of light was multiplied a hundredfold. The lamp became a practical possibility for the illumination of interiors.

Neon lights

Somewhat before the fluorescent tube came into use, lamps with much shorter arcs, in which the vapour operated at a higher pressure, were introduced, including the high-pressure mercury-vapour-discharge lamp and the sodium-discharge lamp. Neither of these found much use for interior lighting, however, because of their colour rendering. They made people look like ghosts, and it was only for street lighting, where safety arguments could be advanced, that the mercury- or the sodium-discharge lamp was accepted. Application of these lamps for exterior lighting developed so rapidly that within 20 years almost all principal streets in cities of the world (excepting the United States) were lit by sodium-discharge lamps. The ghostly rendering of the human complexion was balanced

by the pleasant appearance of the brick and stone facades of buildings. At mid-century the pattern of lighting seemed clear: filament and fluorescent lamps for interiors, neon and similar strips for exterior advertising and signs, and sodium-discharge lamps for streets.

Within decades, however, developments suggested that this pattern might change radically. The mercury-discharge lamp, which, after a brilliant start, rapidly gave way to the sodium lamp, had undergone extensive development; new forms becoming commercially available suffered from fewer of the disadvantages of the parent. Such lamps may eventually find both internal and external use. Yet the day of the sodium lamp was not over, because equally new developments, particularly the high-pressure sodium lamp, had overcome to some extent the unpleasant colour rendering, and it was difficult to predict whether the mercury or the sodium lamp or both together would prevail. Lamp manufacturers continued to experiment with new devices, sometimes in the laboratory and sometimes with a commercial product. In this category might be included the xenon-discharge lamp, a high-powered source with colour almost identical to natural sunlight. The electroluminescent lamp, by virtue of which walls and ceilings could be made self-luminous, was once believed to be the lighting of the future, but it has made little recent progress.

Not all lamp development is concerned with general lighting on a massive scale. So much is now known about the physics of incandescent lighting, of discharge lighting, and of other forms of generating light that any particular requirement can almost certainly be met by a lamp tailored to the purpose. Great lamps of massive light output have been made, using the arc principle, the vapour-discharge principle, and the incandescent-filament principle for lighthouses, searchlights, and the floodlighting of large areas, such as airport landing runways. Microscopic filament lamps have been created for insertion in the body to enable the surgeon, aided by suitable optical equipment, to make visual inspection to supplement other methods of diagnosis. The development of flexible optic fibres, thin wires of transparent plastic that "conduct" light by internal reflection from a suitable lamp at one end to the target at the other, have also been used for medical and other purposes. Other lamps for special uses include highly actinic lamps (*i.e.*, radiating energy in the visible and ultraviolet spectral regions and producing chemical changes thereby) for photography, colour television, and cinematography or lamps with emission in selected parts of the spectrum for industrial uses, infrared lamps for paint drying, and both infrared and ultraviolet lamps for therapeutic use. There are probably few branches of applied physics in which progress has been so rapid during the mid-20th century as in the development of artificial light sources.

Optic fibres

MODERN ADVANCES IN LIGHTING TECHNOLOGY

Lighting as an applied science was founded during the 1890s and 1900s when—with development of the incandescent-filament lamp, the carbon arc, and the maintenance for some years of the supremacy of the Welsbach mantle—the extension of the working day into the hours of darkness became feasible. A massive expansion in artificial-lighting technology resulted.

Lighting technology was advanced partly by unskilled and semiskilled inventors and mechanics and partly by trained physicists and engineers who explored basic principles. The first group produced rapid results, while the second developed the theoretical knowledge upon which subsequent technology could be based.

Fixtures and voltage. The first requirement was to accommodate the light bulb where it was required on the ceiling, desk, or wall. The Edison-screw lamp terminal (male) and power outlet (female) probably had as important an effect upon the development of lighting for domestic and industrial use as any other single invention. All that was required to install electric lighting was for an electrician to affix the Edison-screw power outlet to the appropriate wall or ceiling surface with the necessary twin wires back to the supply terminal and then for the householder to screw into it a light bulb with an Edison-screw cap. Installation was safe, simple, and reliable. The various electric-power companies, Edison's being the pioneer, steadily refined and introduced new methods of power distribution that simplified electric lighting and made its use safer and more reliable. Individual light switches for each room and each lamp soon became sufficiently inexpensive to be universal. Controversy arose, not yet settled, over the best method of distribution to avoid accident by electrocution. The same amount of light can be generated by a lamp carrying a small current (amperes) but subjected to a high electrical pressure (voltage) as would be given by a different lamp carrying a much heavier current but subjected to a lower voltage. It is much cheaper to distribute electricity at a high voltage, because the current can be less and hence the amount of expensive copper in the conductors can be reduced. On the other hand, the risk of death by electrocution is greater at the higher voltage, and experience unfortunately showed that, in spite of instructions, people would occasionally touch the two terminals of the electrical supply together and subject themselves to the full electrical pressure. In the United States a decision was taken to supply households at no more than 110 volts. This meant that the distribution network had to be heavier and therefore more expensive. The United Kingdom, on the other hand, decided to use about 200 volts, thereby approximately halving the cost of domestic distribution but considerably increasing the risk. Neither country has seen fit to make a change. After a period of uncertainty, the whole of Europe followed the British rather than the American lead.

Edison screw

Safety devices. All forms of electric lamps are liable to failure, sometimes gradual, sometimes immediate, sometimes catastrophic. Gradual failure only results in loss of light, which may often be imperceptible over a period. Modern discharge lamps, for example, last indefinitely but slowly lose their light output. Immediate failure, such as the breaking of the filament of an incandescent lamp, causes no more than temporary inconvenience. Sometimes, however, a filament lamp or a discharge lamp may fail, probably due to a manufacturing flaw, in such a way as to give rise to a surge of current against which the distribution system must be protected. Every circuit must have some form of fuse or contact breaker to cut off the current if it exceeds safe levels. With a fuse, one need only detect the cause of the fault, remedy it, and replace the fuse. The contact breaker is a switch that is tripped by an excess of current; it can be reset at once without providing another fuse. The past pattern was to use fuses in domestic circuits and contact breakers in larger installations where a professional electrician was available to detect the cause of failure before resetting the contact breaker. Modern usage relies on the circuit breaker. (For power generation and transmission, see ENERGY CONVERSION.)

Directing light. The next stage in the development of lighting devices was to provide means by which the light from the electric lamp could be directed where it was wanted. At this point, the technology of lighting for seeing and lighting for effect parted company. In the first type of lighting, it is not desirable for the light source to be visible. Instead, maximum light is directed onto the work and none allowed to escape elsewhere. In lighting for effect, however, it is sometimes desirable to see the source for its brilliance; in particular, the light is directed to create effects, modelling, shadows, reflections, and high spots without considering whether people can see to read or work. The lighting in a church, for example, may be designed to enhance the appearance of the architecture.

Lighting for seeing and for effect

To achieve lighting for seeing, two principles of optics—reflection and refraction—were used, either separately or in combination. Light, when received at a smooth, polished surface, is reflected in a predictable and calculable way. Likewise, light received at the (incident) surface of a transparent medium, such as glass, is changed in direction (refracted), and, if the other (emergent) surface of the glass is not parallel to the first surface, the emergent ray of light is changed in direction. Thus, by suitable choice of incident and emergent surface, light can be refracted into any desired direction. The principle of lighting devices employing reflection or refraction is therefore to wrap such a reflecting device (mirror) or refracting device (lens

or prism) around as much of the light source as possible, collecting all the light emitted in unwanted directions and redirecting it where desired.

Both forms of device can be used for a multiplicity of purposes. The searchlight reflector and the lighthouse refractor are both designed to concentrate the maximum intensity in the desired angle but for different purposes. The searchlight concentrates the light in order to illuminate and make visible a distant object. The lighthouse, on the other hand, concentrates the beam to permit the maximum visibility of the source itself from the greatest possible distance.

In the home, factory, or office, on the other hand, a narrow, concentrated beam is rarely required. The purpose is to illuminate the whole of a writing desk or an entire working area. Such lighting devices, therefore, must produce a wide but uniform beam, since it is irritating and fatiguing to work where light is bright at one point and dull at another. The lighting engineer aims at uniformity of illumination and uses reflector contours that calculation and experience indicate will give the most satisfactory result.

High skills were soon brought to bear on the design of optical controls for light, particularly in such areas as hospital operating theatres, where the maximum amount of entirely shadowless lighting is directed onto the surgeon's task; or in street lighting, in which light emitted from a small source is directed along a narrow band of street and sidewalk with perhaps 200 feet or 60 metres between adjacent lampposts. Street-lighting equipment in particular has made use of refraction as the controlling principle, in part because of its greater accuracy but chiefly because the refractor units could be made in the form of glass bowls or dishes that could be tightly sealed to the lamp-holding device to keep out weather and insects.

For interior lighting, however, the need for precise optical control progressively diminished as light became cheaper and more plentiful because of the more efficient generation and distribution of electric power, and because lamps became more efficient in producing light rather than heat. Interior lighting is now a matter of lighting the whole environment instead of only the working task. Fixtures have employed new methods of controlling light, such as the use of optical diffusion. Special techniques were developed, first in central Europe and subsequently elsewhere, that permitted a white opal glass to be blown into attractive shapes, such as spheres and cylinders, in which the lamp could be placed as part of the fixture and its light diffused to give a uniform white-light source of large size and pleasant appearance. The Bauhaus School of Architecture and Design in Germany under Walter Gropius combined with the skill of the central European glassmakers revolutionized interior lighting design during the 1930s. Other diffusing materials have since supplanted glass, and hence there exists an immense range of plastic diffusing light fixtures of all possible shapes and sizes. The reflecting light fixture still has its uses in general room lighting. A contemporary fashion among architects is to inset lighting into the ceiling and direct it through small holes, almost invisible during the day; this has led to development of ceiling-inset spotlights, which use all the old skills of the light-reflector designer.

Diffusing light

The lighting designer. The most essential tasks of the lighting designer are to spread adequate light on the work for efficient and comfortable seeing and to light the environment so that people can see well and enjoy what they see. The determination of the amount of light for efficient and comfortable seeing is a matter for the specialist in physiological optics and in the relation between lighting and seeing. Experimental work in England, Germany, and the United States showed that visual performance could be analyzed into three basic factors: sharpness of vision (visual acuity), detection of contrast, and detection of movement. International agreement was eventually reached on a method by which necessary levels of lighting for specific tasks could be prescribed. By breaking down a visual task—for example, that of threading a needle—into its component parts, the amount of light necessary for efficient and comfortable seeing could be prescribed. This prescription forms the basis of the codes of recommended lighting practice that have been adopted in most countries.

The factors governing discomfort from bright light sources are also known, and a glare prescription can be written, called a visual comfort index in the United States and a glare index in most other countries. Codes of lighting practice can therefore lay down the glare index based on the levels of tolerable brightness of lighting devices in their particular environment in order to avoid discomfort from glare.

Although this article is intended to be concerned primarily with the physical aspects of lighting design, in other words, lighting technology or illuminating engineering, it is notable that lighting is concerned with human sensations; that is, human vision. It must be kept in mind that the eye can adapt to the kind of lighting in which it finds itself. The level of lighting by moonlight, for example, is one-1,000,000th of that by sunlight, and yet the human eye can adapt sufficiently to moonlight to perform many tasks.

This adaptation to light not only affects the range of brightness but also has a profound effect upon the sensation given by any particular physical amount of light. A surface of a certain brightness (luminance) seen in a dimly lit street at night, for example, will appear very bright, whereas a surface of the same brightness (luminance) seen in daylight looks like a very dark shadow. Lighting design must take this into account. Thus it is not physical brightness but apparent brightness that determines the subjective effect. Though apparent brightness is the ultimate criterion for the designer, experience and existing technology allow almost the whole of lighting design to be undertaken in terms of physical measures.

Apparent brightness

Current lighting practice

The brief summary of current lighting practice below attempts to strike a balance among practices in different countries and by different schools of design. This is difficult; there is the "illuminating engineering" school of lighting, and there is the "architects or interior decorators" school, quite different in approach and results and almost irreconcilable. They accord most closely in Britain and are farthest apart in the United States.

INTERIOR LIGHTING

Interior lighting in most countries tends to follow similar patterns, dictated by the availability of light sources, by architectural fashion, and by the special lighting requirements of the task or the environment or both. The amounts of light now considered necessary for efficient and comfortable vision are very much higher than they once were. Lighting levels in fact increased 50-fold in 50 years. This is not because human sight has deteriorated. Ample evidence indicates that in spite of the fact that more people wear spectacles than ever before, human sight remains much what it was in the past, and there is certainly no deterioration due to the use of artificial light. It is simply that people demand perfect vision and thus wear correcting glasses for minor visual defects that were previously neglected. In the same way, people want to be able to see excellently everywhere, whereas previously they were satisfied to move closer to the candle or other light source to read or perform other tasks.

Dwellings. Light installed by a skilled lighting designer and interior decorator can do much to create the character of an interior environment. In modest homes a central pendant fitting supplemented by one or two wall fittings or desk lamps, or lamp standards, remained in the late 20th century the best way of achieving good lighting on sound visual principles with the greatest economy.

The eye functions most efficiently (with greatest visual acuity and comfort) when the task is lit to a slightly higher brightness than the immediate surroundings, which in turn should be slightly brighter than the general environment. This standard is usually achieved in a home when a moderate level of general lighting is provided from a central fixture and a high level of working light from a desk or floor lamp. Less satisfactory is the use of a desk

lamp alone in a room, because this leaves the room too dark and gives rise to visual fatigue.

Recognizing this, the Illuminating Engineering Society of the United States, as long ago as the 1930s, designed a study lamp that provided both working light on the desk and general light in the room; this unit is probably still the most efficient lighting fixture for the economic lighting of working rooms in dwellings.

Fluorescent lighting has not yet found its way into homes to a great extent, apart from utility areas such as kitchens and bathrooms. This is partly due to the large size of the lamp, which is out of scale with normal rooms in small houses, and partly to the fact that the light that it emits is less compatible with the social character of dwellings than the warmer, yellower filament lamp. People apparently do not want to prolong the colour of daylight into their social lives after dark.

Lighting for television viewing

Lighting for television viewing has been a source of some controversy, but it has been demonstrated that people with normal sight can light television-viewing rooms as they please without harming their eyes, though eye muscles may become fatigued more quickly in a darkened room. The colours of a colour-television screen are adulterated by too much white light falling upon the screen, and so, for watching colour television, the level of light should preferably be somewhat less than that for watching black-and-white. Lighting for television is more a matter of common sense and setting the light as the viewers want it, avoiding annoying reflections. Persons with subnormal sight, however, may need care in the prescription of lighting for television viewing.

Schools. With changes in education, schools have been made more attractive and welcoming than ever before. This development has been most marked in England, where school building and school lighting corresponds much more closely to that in homes than it does in other countries. In continental Europe and the United States, emphasis is still placed primarily on lighting the task. American schools are almost always lit by fluorescent lighting, inset in the ceiling or in the form of an overall luminous ceiling, providing shadowless lighting to a high level over the whole working space. In England, on the other hand, the emphasis is on the lighting of the environment. Continental Europe, with the exception of Scandinavia, tends to follow the American pattern.

Offices. Ceiling-mounted fluorescent lighting is used in offices to create a uniform level of light over the whole of the working area, allowing maximum flexibility in office arrangement. The levels provided range from about one-fifth to one-tenth of the light from a natural sky. This appears to be about the optimum required for work on white paper. Improvements in the colour of fluorescent lighting are constantly being made, and the distortions in the appearance of the human complexion have been largely overcome.

The concept of office landscaping, created in West Germany, has emphasized the need for a uniform level of lighting over the whole working space. The essence of the office-landscaping principle is to permit every task to be placed anywhere in the office space and planned in relation to the flow of work demanded, and, for this, uniform lighting is essential.

Lighting practice in offices sometimes makes use of the principle of permanent supplementary artificial lighting of interiors (PSALI), originating at the Building Research Station in England, in which the light from windows and from the artificial-lighting system are planned together as part of an integrated whole. Before this, the artificial lighting was planned primarily for operation after dark, while the windows provided the working light during the daytime and had to be very large if light was to penetrate into a deep office. PSALI means that windows no longer must provide all the working light during the daytime but are important principally because they provide a view and an awareness of the world outside. PSALI has been developed primarily in Great Britain and the Scandinavian countries and, to a lesser extent, in continental Europe. In the United States the contrary tradition has developed, of working by artificial lighting entirely, by day as by night, and even drawing blinds over the windows by day to eliminate glare.

Hospitals. Here, again, different approaches have been taken. American hospitals tend to be much more "institutional" in character, making little use of colour on the walls and hangings and providing high levels of uniform fluorescent lighting from ceiling-mounted fixtures and from fixtures placed immediately above the bedheads. British practice, on the other hand, employs much lower levels of light, on the theory that sick people often want to rest; colour is used extensively both on walls and in draperies. There has long been a strong prejudice in England against the use of fluorescent lighting in hospitals, and many modern hospitals are still lighted by filament bulbs. Hospital day spaces are designed to resemble a home both in the use of light and of colour, as are visitors' waiting rooms.

Lighting the operating theatre

The essential requirement in an operating theatre is that high-intensity light should be concentrated upon the operating area so that, no matter where the surgeon places his hands, no shadows are cast. For shadow-free lighting, one very large source is needed, which may either be a suspended lighting unit (the least costly solution) or a luminous ceiling as developed in France. A luminous ceiling may be an enormous elliptical reflector, with the light source at one focus and the operating area at the other, or it may consist of a multiplicity of small projectors concentrated on the operating area, possibly with an array of push buttons at the surgeon's control to change the direction and distribution of the light. It is important to light the whole theatre as well as the operating area, both to enable the assistant staff to work efficiently and to minimize visual fatigue on the part of the surgeon.

Factories. In industrial processes, the prime requirement is to light the task efficiently, since people are concentrating on their work and are less concerned with the environment. Factory lighting is provided almost universally by ceiling-mounted fluorescent equipment to a high level of illumination, ranging from one-tenth to one-fourth that given by the natural sky on a bright day. It is important to bear in mind that the factory-lighting problem is different from that of offices. Most office work is concerned with white paper, whereas most factory processes are concerned with objects that reflect much less light than white paper; therefore, the same degree of visual comfort results from a very much higher level of illumination. The higher level of illumination assists acuity and contrast and colour discrimination, thereby improving visual performance. Experience over the years has shown that higher levels of lighting give rise to higher productivity and that good lighting pays for itself.

Some critical visual tasks demand additional lighting, both to reduce visual fatigue and to give added visual efficiency. Assembly of fine instruments, watches, and cameras demands light on the work approximating that given by the bright sky, which is better achieved by selective lighting. Similarly, in drawing offices, many draftsmen like to have an individual desk lamp for selective lighting on the work, plus a high level of general illumination in the office.

Factory light fixtures

The factory-lighting designer often must choose between using totally enclosed and open fixtures. Though the totally enclosed fixture may be readily cleaned and maintained, open fittings can be so designed that the heat of the lamp creates an upward draft through the fitting, which carries dirt with it and so reduces the amount that settles on the reflecting and transmitting surface of the fixture. Thus, open fixtures are said to require less maintenance than enclosed ones. With the increasing tendency to build factories as totally enclosed artificial environments—with heat supplied from high levels of lighting, from machinery and industrial processes, and from the workers' bodies—the use of the open-type fixture is indicated, coupled with efficient filtering of the recirculated air. A further disadvantage of the enclosed fixture is that in a factory it may overheat, and, since fluorescent lamps rapidly drop in light output with increase of temperature, the light output from the fixture will fall.

The building lighting in a factory is usually arranged

either as high-bay lighting, in which the fixtures are symmetrically positioned at maximum height from the floor, or as continuous-run lighting, in which fixtures are placed end to end across the width of the factory. The high-bay lighting system is better suited to spaces like aircraft hangars because it gives a uniform distribution of illumination with minimum interference with the volume of the space. The continuous-run system can be related directly to an assembly line or row of benches. Furthermore, it can be closely integrated with the daylighting, the continuous run of fluorescent lighting fittings being set along the same lines as the continuous run of glazing in the roof, either to replace it after dark or to act as a permanent supplement during the daytime. Luminous ceilings are also employed in factories, particularly in those that demand a very close control over the interior climate. The luminous ceiling may consist of an uninterrupted plastic surface behind which the fluorescent lamps are arranged at regular intervals to give a uniform distribution of brightness and to help maintain a uniform temperature in the space below. In certain specialized manufacturing operations, such as the assembly of delicate watch mechanisms or submarine repeaters, in which avoiding dust is important, the overall sealed ceiling lighting system has obvious advantages.

Few factories are lighted today by uncorrected mercury-vapour or sodium-discharge lamps. During the 1930s and '40s many such installations were made, and some still remain, but they are being rapidly replaced by fluorescent lighting. The two main complaints against the use of mercury vapour and sodium were the unpleasant colour and the fact that flicker and stroboscopic effects were marked. Such effects could cause dangerous misjudgments of speed with rotating machinery. Modern fluorescent lighting gives less trouble, but on alternating current some residual flicker, greater than that from filament lighting, is present. In poorly engineered lighting installations, lamp flicker may be acutely annoying. Fluorescent lamps near the end of their useful life also tend to flicker. Flicker can be reduced by use of special circuits or special lamps with a long "afterglow."

Flicker in fluorescent lighting

EXTERIOR LIGHTING

While the design of much interior lighting, particularly in nonindustrial buildings, can be undertaken with some technical assistance by the architect or interior designer, most exterior lighting demands the services of the specialized lighting engineer. Lighting an airfield, for example, is recognized as a highly specialized and skilled job that would be dangerous to leave to any but an experienced professional. Not only must the airstrip itself be illuminated so the incoming pilot can judge position and distance accurately, but the approach lighting must be designed so that the pilot can orient the aircraft onto the right flight path with great accuracy from a great distance. The principles of central and peripheral vision have been studied in order to understand how people orient themselves in space and time with no other visual clues than points of light arranged in patterns in space. Despite great advances in instrumented landing equipment, airfield lighting promises to continue indefinitely as an important field for the lighting engineer.

The lighting engineer also has a special role in such public places as streets, sports arenas, and parks and in the decorative floodlighting of buildings. All these forms of exterior lighting call upon the use of specialized equipment more or less tailored to the job.

Streets. Artificial lighting was perceived at an early date as a potential crime deterrent. The first installations of gas street lighting were made in London and Paris in the early 19th century with this purpose in mind. Toward the end of the century, large cities began to recognize the prestige value of street lighting in the centre of their cities and developed more sophisticated street lighting with ornate cast-iron lampposts housing fixtures mounted 25 to 30 feet (eight to nine metres) above the roadway with powerful filament lamps or arc lamps lighting not only the roadway but the prestigious buildings around. This phase of street lighting lasted approximately up to World War I.

With the development of the internal-combustion engine and the increase of traffic on the roads at night, street lighting acquired new importance. Levels of illumination were considerably increased with the aid of mercury- and sodium-vapour-discharge lamps. At this stage practice in the United States, in Great Britain, and in continental Europe diverged. The United States was most backward in the development of fixed street lighting, relying more upon the light from automobile headlamps and from shop windows, so that by the end of the 1930s the United States was at least 10 years behind Europe. In Great Britain silhouette vision formed the basis for street-lighting design, the principle being that light from the street lamps was reflected off the "sheen" of the road surface to provide a background brightness against which objects were seen black in silhouette. Objections that this was an unnatural way to see things were largely overruled by the fact that far better revealing power could be achieved this way for the same expenditure of electrical power than by any other method. Continental Europe, on the other hand, adhered to the principle of natural vision, lighting the object preferentially and keeping the background dark. Revealing power by the continental system, however, was far less. The disadvantage of the British system was that it required by its very nature a high level of light to reach the observer's eye because the optical system of the lantern could not be controlled with sufficient precision to confine the light to the road surface alone. The continental system, on the other hand, because it did not depend upon reflection, was far more comfortable and less glaring.

Silhouette vision

After World War II, work undertaken in a number of laboratories showed that one of the major causes of street accidents was skidding and that much could be done to minimize this by the better design of road surfaces combined with tire-tread improvements. The road surfaces with best skid-resistance characteristics had exactly the wrong reflection characteristics for the silhouette principle of street lighting. Attention was therefore directed everywhere to the further development of the continental street-lighting system of illuminating the objects directly. With greater affluence, the much higher levels of lighting necessary for the proper operation of the continental system could be afforded. The silhouette principle, of immense value at its time, has now been abandoned, although in practice much vision on the streets at night is still by silhouette.

Street lighting today employs sodium-discharge equipment almost universally for traffic routes because of its high luminous efficiency and relative ease of optical control, and fluorescent-tube lighting or colour-corrected mercury-vapour-discharge lamps in the centres of cities. Filament lighting, like gas-lighting, has been almost entirely abandoned. The use of fluorescent lighting necessarily gives rise to very large lighting fixtures, which may be mounted on standards or hung over the roadway, of a size and proportion that would not have been tolerated in the early part of the century.

Sports fields and arenas. Although experiments in lighting outdoor sports areas were reported in the 1880s, the first commercial successes came in the 1930s. Lighting of these installations has developed rapidly since World War II. As soon as lighting engineers could demonstrate that the cost of lighting a large arena to the levels necessary for games or athletic performances was economically feasible, sportsground lighting developed rapidly. The basic principle is almost always that of employing massive banks of light projectors around the periphery of the ground, floodlighting the arena, rather than suspending lighting units above the arena itself. This has led to the development of powerful projectors based on the searchlight principle, and, for this, the development of the tungsten–halogen projector lamp, with its high light output and relatively long life, has been crucial to success. Arc lamps, the basis of earlier searchlight design, have been almost entirely supplanted, though the development of new forms of colour-corrected mercury-vapour-discharge lamps or the further development of the xenon-discharge lamp may prove eventually to be of value in sports-arena lighting.

Floodlighting of buildings. This field has also been

transformed by the availability of filament lamps of high light output and long life. As with the lighting of sports arenas, the principle is to project light onto the facade of the building from powerful projectors located sufficiently far from the building not to impede circulation around it. The aim in floodlight design is to distribute the light, in amount and direction, so that all parts of the facade are adequately illuminated and the sculpture and form of the building are modelled either as they are in daylight or in such a way that other aspects of the design make themselves apparent. These principles have also been applied to the lighting of the interior of monumental buildings, such as cathedrals.

Measurement and calculation of lighting

FUNDAMENTALS OF LIGHT MEASUREMENT AND CONTROL

Lighting engineering is based on a few simple principles of the physics of radiation. A light source is said to radiate light because part of its emitted radiation causes a particular sensation when it strikes the retina of the human eye. Outside the visible spectrum, radiation that does not give rise to a sensation of light may be of short wavelengths (ultraviolet radiation) or longer wavelengths (infrared radiation).

The name light flux is given to the rate of emission from the source; the unit of light flux is the lumen. The light output from any lighting fixture is expressed in lumens. A wax candle emits about 13 lumens, a 100-watt filament lamp about 1,200 lumens, and a five-foot (1.5-metre) fluorescent tube about 5,000 lumens.

The lumen

Illuminance, reflectance, and utilance. The lighting level—that is, the amount of light spread over the area receiving the light—is called the illuminance (sometimes the illumination) and is expressed in lumens per unit area. One lumen per square foot is the amount of illuminance of a surface one foot (30 centimetres) from the light of one ordinary wax candle. (Hence, in the United States the unit of illumination is the footcandle, identical with the lumen per square foot.) Unobstructed daylight from a bright, overcast sky gives about 1,000 lumens per square foot. The unit of illuminance in metric units is the lux—one lumen per square metre. One lumen per square foot equals 10.76 lux. Unobstructed daylight from a bright, overcast sky gives about 10,000 lux. Overall fluorescent lighting to the standard required for factory work is about 1,000 lux.

Light falling on a surface is reflected off it according to the ability of the surface to reflect light. This reflection factor, or reflectance, relates the light falling on the surface as seen by the observer. A white surface has a reflectance of nearly 100 percent, while a black surface has a reflectance of about 2 percent and a medium-gray surface about 40 percent. If the surface diffuses light equally in all directions (*e.g.,* a matte surface paint or a dull velvet), the brightness of the surface will be related directly to the illumination falling upon it. The physicist and engineer use the word luminance to express the physical measure of brightness. The unit of luminance used in the United States is the footlambert. Luminance, illuminance, and reflectance are related by the equation: Luminance (footlamberts) equals illuminance (footcandles) times reflectance. Thus a diffusing off-white surface of 50 percent reflectance receiving 10 footcandles will have a luminance of five footlamberts. Illuminance to a level of 100 footcandles falling on a desk top of reflectance of 30 percent on which is a white paper of reflectance of 80 percent would thus give a luminance of 80 footlamberts to the paper and 30 footlamberts to the desk top.

In order to calculate the amount of light necessary to produce a certain level of illuminance in a room, the concept of utilance, or coefficient of utilization, must be introduced. This is the proportion of light that is actually received on the surface under consideration. If in the room the total light from all the lighting fixtures in the installation is directed onto the working reference surface and nowhere else, the illuminance would obviously be given by the density of light flux in lumens over the area of the reference surface; that is, illuminance equals total flux (lumens) divided by area of reference plane (square feet or metres).

Much of this light flux from the fixtures is, however, lost in various ways. Some light is absorbed in the fixtures themselves, while more light falls on the ceilings, walls, furniture, and elsewhere and is absorbed there and never reaches the reference surface. The coefficient of utilization, or utilance, is the measure of the ratio of the light flux actually reaching the reference plane to the light flux initially emitted by the lamps of the installation. It is determined by the loss of light by absorption at the surfaces of the room, the absorption of light in the fixtures, and the effect of the proportions of the room. Calculation of the utilance is a complicated process, but it is made easy by the publication of tables of utilance. Computer programs for its calculation are also available. In practice, the utilance varies from about 0.15 for totally indirect lighting (which hits the ceiling first and is reflected downward onto the working surface) through about 0.3 or 0.4 for general lighting (emitted in all directions) to about 0.6 for direct lighting (directed by means of reflectors or refractors directly on to the working surface).

Calculating the utilance

If, for example, the required average level of illuminance in a factory 100×30 feet (30×9 metres) is to be 100 footcandles, and the coefficient of utilization has been found from tables for the particular space, the reflection characteristics of the walls, floor, and ceiling to be 0.3, then the total amount of light flux in lumens required from the lighting installation will be: required light flux (lumens) equals illuminance times area utilance, equals

$$\frac{100 \times 3000}{0.3}$$

equal 1,000,000 lumens.

If the installation is to use five-foot (1.5-metre) fluorescent tubes, each giving 5,000 lumens, 200 tubes will be needed. These might be arranged in continuous trunking along the length of the factory, 20 tubes per line, requiring 10 lines across the width of 30 feet (nine metres), or, more economically, the tubes might be twinned giving 40 tubes per line and hence five lines spaced, say, five feet (1.5 metres) apart, across the width of the factory.

This calculation does not allow for the fact that the light output will drop during the life of the installation, due to the inevitable accumulation of dirt and to the fall-off in light output as the lamps get older. It is usual in lighting calculations to step up the light output by about 25 percent to allow for depreciation losses.

The above calculation is very simple and does not require the skills of the lighting engineer. It is no more than a first attempt to indicate how much lighting is required. More often the lighting designer must direct light specifically where it is wanted for efficient and comfortable seeing. The light from bare lamps must be controlled according to his plans. In order to design a fixture that will, when correctly mounted, direct the light to where it is wanted, he must first know the light output of the source (the bare lamp) in all directions. This directional light output must be expressed in terms of the amount of light flux (lumens) being emitted by the source within a given cone (solid angle). Light flux per unit solid angle is expressed in candelas (formerly candlepower) and is called the intensity of the light source. An ordinary wax candle has an intensity of approximately one candela.

The intensity distribution must be measured in order to determine where the light from the lamp is going and, thence, to redirect it where it is wanted. The intensity distribution is measured by placing a light receptor at a fixed distance from the lamp and in all necessary directions. The amount of light received on the receptor surface enables the intensity of the source to be calculated by means of the inverse-square law.

Inverse-square law

The inverse-square law of radiation states that the illuminance, E, at a point varies directly with the intensity, I, of the source and inversely as the square of the distance, D, between the source and the point. If the surface at the point is normal to the direction of the incident light, the inverse-square law gives: $E = I/D^2$.

This is true if the source is a point and is approximately true if the distance, D, is great compared with the dimen-

sions of the source. If, therefore, in measuring the intensity distribution of a point source of light, the illuminance at a given point at a given distance from the source is obtained, the light receptor being perpendicular to the direction of the source, the intensity I is given by: $I = E \times D^2$.

If the light receptor is moved around the light source on, for example, a rotating arm so that it is always at the same distance from the light source and always normal to it, the intensity in any given direction will therefore be proportional to the illuminance of the receptor, which can then be calibrated directly in units of intensity (candelas). Such an instrument employed in this way is called a light-distribution photometer.

Light distribution. The light distribution from a source can be expressed in different ways. If it is known that the light source is symmetrical about an axis (an ordinary general service pearl bulb, for example, is to all intents and purposes symmetrical about the vertical axis through the cap of the lamp, and a fluorescent tube is to all intents and purposes symmetrical about an axis along the centre line of the tube), the light distribution can be expressed in two dimensions in the form of a polar curve, a graph on circular coordinates. The light output can equally be expressed on a graph in rectangular coordinates, but the polar distribution curve gives a better visualization of the light-output characteristics of the lamp.

Light output of a complete fixture

The same procedure is used for measuring the light output of the complete fixture. After the fixture has been designed to direct light where required, the light output from the complete fixture must be measured in order to proceed with the next stage of lighting design. Here, again, if the light output from the fixture is symmetrical, a simple polar curve in two dimensions will give all the necessary information. If, however, the light output is not symmetrical—as would be the case, for example, in a street-lighting fixture where the purpose is to direct light in two directions only—it is then necessary to express the light output of the fixture in three dimensions. This is often done by means of a special graph based on principles of trigonometry and called an isocandle diagram. The information in the polar-curve intensity diagram, for a symmetrical fixture, or in the isocandle intensity diagram, for a nonsymmetrical fixture, gives all that is necessary to proceed on lighting design.

If, for example, the requirement is to produce an illuminance of 10 footcandles on a surface normal to the fixture and distant 100 feet (30 metres) from it, the inverse-square law will give the necessary intensity as being: $10 \times (100 \times 100) = 100{,}000$ candelas in the direction of the surface. The intensities in other directions can be found in the same way. Hence, the required light distribution of the fixture can be determined from the illuminance requirements. Equally, in reverse, if a given fixture is available for the job, the illuminance values in significant directions and at given distances from the fixture can be found and compared with requirements.

This type of point-by-point lighting design is necessary when it is required to find or to provide specific illuminance at a specific point. A great deal of lighting design, however, particularly of interiors and also of large areas out-of-doors—such as a building facade—is more concerned with average illumination over a whole area. For this determination, it is better to know the light distribution from the fixture in broad zones. Thus the average intensity may be tabulated for a symmetrical fixture in zones. Even more simply, the light-output distribution of the fixture may be classified into three distribution categories—direct, general, and indirect distributions. In a direct distribution, all the light is concentrated downward onto the working surface. In indirect distribution, the light is directed upward onto the ceiling and reflected back to the working surface. A general distribution is one with equal intensities in all directions, so that some light reaches the working surface directly and the rest reaches it only after reflection.

Light distribution categories

These different distributions not only determine the calculation of the lighting design but also in large measure the character of the lighting. Totally direct lighting on the working surface, while concentrating the attention on the job, gives a dark ceiling, often considered unpleasant and believed to lead to visual fatigue. It is a type of lighting formerly characteristic of industrial areas when light was expensive and had to be directed exclusively onto the work.

Indirect lighting

Indirect lighting is pleasant and restful and has often been advocated as the ideal form of lighting since it essentially imitates the diffuse light from the sky. It is not entirely satisfactory for work lighting, however, because the eyes tend to be drawn upward to the bright ceiling. It also lacks life and tends to be soporific. Indirect lighting, provided with a little sparkle from, for example, small reflecting surfaces and supplemented by local lighting on the work, can provide a very pleasant illumination. The best lighting, however, is usually provided by some form of general lighting combined with some direct lighting on special areas.

Glare, or visual comfort, index. Glare control, a recent development in lighting technology, takes two forms, one very simple, one very complex. The simplest means of glare control limits the perceived luminance of the fixture in the directions of view most likely to cause trouble. A typical example of such control is in the regulations for school buildings in England and Wales, where it is required that the luminance over an angle up to 45° above the line of sight should not on the average exceed 1,000 footlamberts. Otherwise it is necessary to determine the glare index for the whole environment.

Precalculated tables of glare index are published in Great Britain, similar to the tables of utilance, which enable the lighting engineer to find the glare index for his installation, provided he knows the light output and light distribution of his fixtures, as well as their number, position and mounting height, and the dimensions and reflectance characteristics of the room surfaces. Computer programs also exist.

Light-measuring instruments. The commonest photometric devices are based on the changes that occur in the electrical properties of certain substances when light strikes them. Simple measurements of illuminance made by the lighting engineer usually employ a photoelectric photometer with a light receptor such as a selenium photocell, connected to a suitable instrument that measures the current generated by the cell when light falls upon it.

Another light-sensitive material, cadmium sulfide, has been used in photometers: the change in its electrical conductivity upon illumination causes a change in the current produced by a battery. The instrument is calibrated at a standardizing laboratory in terms of the international standard of light.

Very simple light meters, which can be carried in the pocket, are of use to the lighting engineer as a rough

Recommended Lighting Values*

class	examples	recommended illuminance		limiting glare index
		lux	footcandles	
Visual task				
Little difficulty	stores	200	20	—
Ordinary tasks	general offices	400	40	—
Some difficulty	business machines	600	60	—
Fine detail	tailoring	900	90	—
Very fine detail	gem cutting	local, up to 3,000	300	—
Environment				
Glare of little consequence	boiler houses	—	—	28
Secondary consequence	machine shops	—	—	25
Limited glare	shops	—	—	22
Close control	general offices	—	—	19
Closer control	school rooms	—	—	16
Minimum	hospital wards	—	—	13
No glare tolerable	art galleries	—	—	10

*Values taken from current lighting codes in Great Britain.

guide. They cannot, however, match in accuracy a precision instrument.

STANDARDS OF LIGHTING

Standards of lighting as laid down in regulations are in terms of illuminance (footcandles) and glare index, and these standards vary from one visual task to another and from one type of environment to another. The Table gives a brief selection of recommended values taken from current lighting codes in Great Britain.

American recommendations are at present higher than those for other countries. Differences in actual practice in the different countries are narrowing, however. America sets a high target, while European countries tend to set a lower target, which is often surpassed.

So far as the control of glare is concerned, there has not yet been international agreement on the method of specifying glare, and it is only in the United Kingdom that a universal and comprehensive generally used system of glare control is available. Typical values of recommended maximum, limiting glare index on the British system only, are therefore given.

BIBLIOGRAPHY. There are two quite different kinds of books written about lighting, those written by architects and interior decorators who use light as a tool to create atmosphere and character, and those written by illuminating engineers. Occasionally, as in DEREK PHILLIPS, *Lighting in Architectural Design* (1964), and R.G. HOPKINSON and J.D. KAY, *The Lighting of Buildings* (1969), the attempt has been made to marry the two different aspects of lighting, the former by an architect who practices as a lighting consultant and the latter a collaborative effort by an illuminating engineer and an architect working together. Generally, however, the architectural lighting books say relatively little about the hardware, and the illuminating engineering books say very little about the effects of light. Probably the two most useful reference books are the ILLUMINATING ENGINEERING SOCIETY, *Lighting Handbook* (issued every five years); and the ILLUMINATING ENGINEERING SOCIETY OF GREAT BRITAIN, *Recommendations for Good Interior Lighting.* The former is a compendium of every aspect of lighting and also includes a mass of useful trade material, whereas the latter is a brief summary and probably the best introduction to the subject for both the specialist and nonspecialist.

The best books on illuminating engineering were written some time ago, when the subject was a major discipline in schools of electrical engineering. The *Theory and Design of Illuminating Engineering Equipment,* by L.W. JOLLEY, J.M. WALDRAM, and G.H. WILSON (1930), is a "lighting engineer's bible" so far as principles and design are concerned. So are H.H. HIGBIE, *Lighting Calculations* (1934); and D.H. JACOBS, *Fundamentals of Optical Engineering* (1943). More recent books on the hardware of lighting are S.T. HENDERSON and A.M. MARSDEN (eds.), *Lamps and Lighting,* 2nd ed. (1972), and J.M. WALDRAM, *Street Lighting* (1952); J.W.T. WALSH, *Photometry* (1958), gives not only the fundamentals of photometry but also the basis of lighting engineering insofar as this relates to the measurement of light.

Later books on lighting and seeing are H.C. WESTON, *Sight, Light and Work* (1962); and a short book written primarily for the informed layman, R.G. HOPKINSON, *Lighting and Seeing* (1969). There are a great many specialist books about particular aspects of lighting—the NUFFIELD PROVINCIAL HOSPITALS TRUST, *Studies in the Function and Design of Hospitals* (1965), is of particular value to architects in showing the relation between natural and artificial lighting and building design. Much recent work on lighting has not been gathered between hard covers yet, and so it is necessary to refer to the original papers. These appear particularly in *Illuminating Engineering,* published by the Illuminating Engineering Society of New York; and *Lighting Research and Technology,* published by the Illuminating Engineering Society of Great Britain.

(R.G.H.)

Lima

Lima, the national capital of the Republic of Peru, is also the country's commercial and industrial centre. Central Lima is located at an elevation of 512 feet (156 metres) on the south bank of the Río Rímac, about eight miles (13 kilometres) inland from the Pacific Ocean port of Callao, and has an area of 27 square miles (70 square kilometres). Its name is a corruption of the Quechua Indian name Rímac, meaning "Talker." The city forms a modern oasis, surrounded by the Peruvian coastal desert and overshadowed by the neighbouring Andes mountains.

This article is divided into the following sections:

Physical and human geography 38
The character of the city 38
The landscape 38
The city site
Climate
The city layout
The people 39
The economy 40
Industry and commerce
Transportation
Administration and social conditions 40
Government
Services
Cultural life 41
History 41
Pre-Columbian and colonial periods 41
The modern city 41
Bibliography 41

Physical and human geography

THE CHARACTER OF THE CITY

Perhaps the best clue to the significance of Lima to the nation of Peru can be found in its most popular nickname: El Pulpo ("The Octopus"). Metropolitan Lima's huge size—it accounts for about one-third of the total population of Peru—has both resulted from and stimulated the concentration of people, capital, political influence, and social innovations. Lima's unique status is but one of the more important consequences of a highly centralized, unitary state that from its inception in the early 19th century solved interregional conflicts by focusing power and prestige on the city. With its port of Callao and its location at the centre of Peru's Pacific coast, Lima was long the only point of contact between the nation and the outside world.

As with many sprawling and rapidly growing metropolitan centres, Lima has its detractors as well as its promoters. Those who remember the more tranquil, traditional days, before the arrival of millions of migrants and before the many buses and automobiles brought pollution and congestion, are prone to use another nickname for the capital: Lima la Horrible. This is the noisy, dirty, gloomy, damp, and depressing Lima, perceptions shared by both short-term visitors and longtime residents. Even though sunshine does break through the dense coastal cloud banks in the summer, Lima then becomes unbearably hot as well as humid, and the sunshine seems to emphasize even more clearly the grimy buildings and lack of greenery in the central city.

THE LANDSCAPE

The city site. Lima sprawls well beyond its original Spanish site at a bridgeable point on the Río Rímac. Disgorging precipitously from the high Andes, the Rímac

has formed a flat-topped alluvial cone, on which the early Spanish colonists established their settlement. Since almost the entire coastal plain in central Peru consists of unconsolidated fluvioglacial deposits, cliff erosion and earthquakes are continual threats. In expanding from its original site, the city has incorporated within its fabric various hills and valleys that are also prone to earth tremors and flash floods. One of the most notable characteristics of Lima is the barren, unvegetated desert that surrounds it on all sides; the grayish-yellow sands support almost no plant or animal life, save where water has been artificially provided.

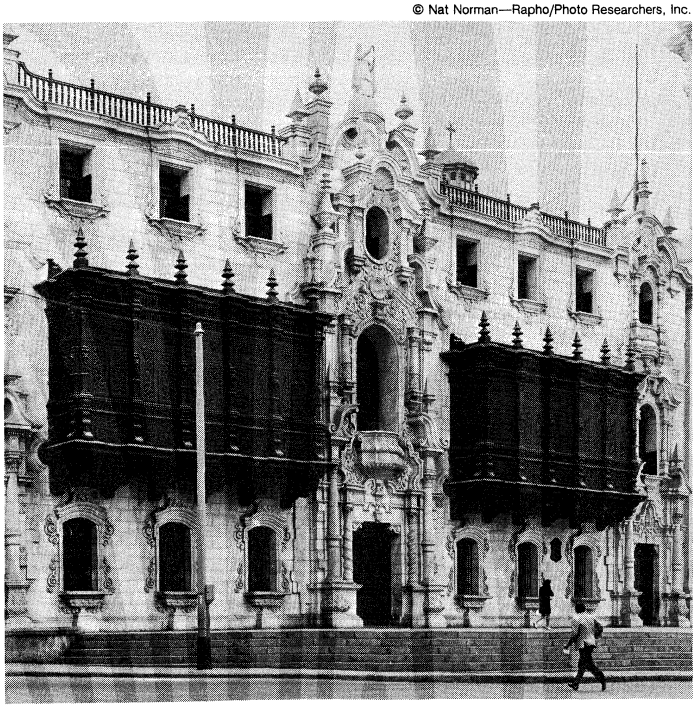
Climate. Though Lima is located at a tropical latitude, the cool offshore Peru (also called Humboldt) Current helps produce a year-round temperate climate. Average temperature ranges from 60° to 64° F (16° to 18° C) in the winter months of May to November and 70° to 80° F (21° to 27° C) in the summer months of December to April. The cooling of the coastal air mass produces thick cloud cover throughout the winter, and the *garúa* (dense sea mist) often rolls in to blanket areas of the city. Precipitation, which rarely exceeds two inches (50 millimetres) per annum, usually results from the condensation of the *garúa*. Lima is perhaps best described as cold and damp in winter and hot and humid in summer.

The *garúa*

Because clouds tend to trap airborne pollutants, Limeños can often taste the air. A permanent problem resulting from the high humidity is oxidation, rust being a common sight. Many of the wealthier citizens established winter homes on the coast north or south of the city proper or in such localities as La Molina, a short distance to the east of Lima, where the climate is free of fog and cloud.

The city layout. Lima contains a series of townscapes well-defined by its long history. The core of old Lima, delineated by Spanish colonists in the 16th century and partly enclosed by defensive walls in the 17th, retains its checkerboard street pattern. Bounded on the north by the Rímac and on the east, south, and west by broad avenues, old Lima contains a few restored colonial buildings (Torre Tagle Palace, the cathedral, and the Archbishop's Palace) interspersed among buildings of the 19th and 20th centuries, many of which were built upon the sites of former colonial residences that had collapsed during the major earthquakes that have struck the city. The old walls, however, were demolished in the mid-19th century. The two principal squares (Plaza de Armas and Plaza Bolívar) still provide the foci of architectural interest within central Lima, and the enclosed wooden balconies so typical of the colonial city have now become features to be preserved or restored. The Presidential Palace (built on the site of Pizarro's house) and many other buildings reflect the past popularity of the French Empire style. On the north side of the Rímac, the old colonial suburb of the same name conserves relics of its past in its curved, narrow streets, tightly packed with single-story houses, and its Alameda de los Descalzos (Boulevard of the Barefoot Monks).

© Nat Norman—Rapho/Photo Researchers, Inc.



The Archbishop's Palace, with colonial-style wooden balconies, on the Plaza de Armas.

The former residential zone of central Lima has undergone several radical modifications, especially since the 1930s. Most of the old spacious mansions have been subdivided so that they now accommodate as many as 50 families. These inner-city slums (variously called *tugurios, corralones,* and *callejones*) have been occupied by immigrants from the countryside striving to gain a foothold in the urban economy and society. Sanitary conditions in such zones are often very poor.

Other parts of old Lima have experienced demolition and reconstruction. Housing has given way to banks, insurance offices, law firms, and government offices. Though there have been repeated attempts to stimulate pride in El Cercado (the formerly walled enclosure), most Limeños regard it as a place to pass through rather than to preserve and enhance. One finds little evidence of gentrification in Lima; unlike other Latin-American capitals and even other cities within Peru, central Lima contains relatively few outstanding architectural features.

Growth of the city

Lima did not expand much beyond the walls of the old city until railways and tramlines were constructed in the mid-19th century. For the next 75 years growth was steady, the axes of urban development from old Lima assuming distinctive characters: the area west to Callao became the industrial corridor; the sweeping bay frontage to the south from Barranco to Magdalena was transformed into the choice residential zone; and eastward, toward Vitarte, a mix of industrial and lower-class suburbs sprang up. As the pace of urban expansion increased in the 1930s, small communities formed in the open country between Lima and the coast. These gradually coalesced into such urban districts as La Victoria, Lince, San Isidro, and Breña. The numerous farms and small tracts of cultivated land between suburbs and barren, dry land also became urbanized as immigrants from the interior occupied these areas. In the 1950s Lima became noted for these *barriadas* (squatter camps of shanties), which as they became more permanently established were renamed *pueblos jóvenes* ("young towns"). These communities have come to contain one-third of the population of metropolitan Lima. The older *pueblo jóvenes,* such as Comas, are now difficult to distinguish from the "established" sections of the city, since the early constructions of cardboard, tin cans, and wicker matting have long since given way to bricks, cement blocks, and neat gardens.

Lima's contemporary townscapes provide such contrasts that it is easy to forget that the rich and the poor belong to the same society. Within a few blocks one can move from luxury to abject poverty. With downtown Lima often heavily congested with traffic, suburban locations were chosen for many new businesses, factories, and shopping centres. The classic corner stores run by Chinese and Japanese immigrants and their descendants are fighting a losing battle against the competition of large, hygienic supermarkets, which in turn are increasingly encircled by *ambulantes* (street vendors).

THE PEOPLE

Just as the physical fabric of Lima has been transformed since the 1930s, so too has its population. It is now difficult to identify what might be called a true Limeño, for in a very real sense Lima has become the most Peruvian of cities; everywhere one can hear different accents, reflecting the myriad origins of the *provincianos* who have made the city a microcosm of the nation. Before the arrival of the highland migrants (commonly called *serranos* or, if demonstrating what are perceived to be Indian characteristics, *cholos*), it was relatively easy to mark the difference between the white elite and other socioracial mixtures. Race, ethnicity, and class, however, now present

a complexity that defies easy classification. The greatest difference that persists, and perhaps even increases, is that which divides the rich and influential from the poor and powerless. One has only to compare the elegance of those who stroll through Kennedy Park in Miraflores on a Saturday night with the squalor of those who beg in central Lima to realize that, in growing, the city has not developed. For the great majority of people access to piped water, sewage systems, inexpensive food, and steady employment are still dreams for the future.

Religion

The vast majority of Limeños are Roman Catholics, which gives the city a traditional, conservative atmosphere; this is evidenced by the enormous crowds of people who gather for such annual religious processions as El Señor de los Milagros ("the Lord of Miracles"), Santa Rosa de Lima, and San Martín de Porres. Growing numbers of residents from the slums and poor suburbs, however, have begun to question the church's positions on social and political issues, encouraged by priests who advocate what has become known as liberation theology.

THE ECONOMY

Whatever indicator is used to measure economic performance, Lima maintains a dominant position within Peru, accounting for three-fifths of the country's industrial output and nearly all of the volume of its financial transactions. The size of Lima's population makes it the premier market for all domestic and imported goods; Limeños make some four-fifths of the nation's consumer purchases each year.

Industry and commerce. Industry in Lima is located primarily in the old Callao–Lima–Vitarte corridor, with more recent additions in zones fringing the Pan-American Highway north and south of the city. Industrial activity is diverse, ranging from shipbuilding, automobile manufacturing, and oil refining to food processing and the manufacture of cement, chemicals, pharmaceuticals, plastics, textiles and clothing, and furniture. Much of this capital-intensive, heavily unionized industrial base, however, operates well below capacity, in most part because of the dire economic situation of Peru.

There has thus been a gradual de-emphasis of the more established industries, and since about 1970 a new type of informal, artisan-based industrial structure has developed. These small-scale, labour-intensive enterprises, which often are family controlled, have been better able to meet the demands of consumers by having goods more readily available (in part by avoiding bureaucratic red tape) and by offering goods for lower prices (in part by avoiding the payment of taxes).

Many industries have located within metropolitan Lima because of its pool of skilled labour, personal access to government officials, and the benefits of well-established networks of marketing and services such as banking. Manufacturing has not provided an adequate solution to the demands of the vast numbers who seek employment. One result has been the rapid rise in service jobs, the majority of which are informal in character. This type of employment has been estimated to account for at least 40 percent of total economic activity in the metropolitan area. The thousands of *ambulantes* who work the streets of central Lima have become a visual reminder of the lack of steady employment in the formal sector. One of the largest employers in Lima—directly and indirectly—is the national government. Its ministries, institutes, and other agencies provide jobs not only for an extensive bureaucracy but also for the hundreds of thousands of people who in various ways serve the needs of those fully employed.

Transportation. The railway line from Callao to Lima is the oldest in South America, while the line that climbs east past Vitarte and into the Andes reaches the highest point of any standard-gauge railway in the world. Other short lines radiate north to Paramonga and south to Lurín. The growth of automobile transportation has given rise to the heavily congested traffic conditions that exist in contemporary Lima. Although there is now a well-developed highway system in the metropolitan area, including an expressway between central Lima and Miraflores, the vast majority of Limeños must cope with an outdated street network and rely on three basic modes of transport: *colectivos,* shared taxis that can hold up to 10 passengers; small buses that can carry up to 20 persons; and municipal buses, of which more than half normally are broken down and the other half operate in bad repair.

Because transport in Lima is at best highly inefficient and at worst chaotic, hundreds of amateur taxi drivers (*piratas*), unlicensed and often ignorant of all but the most obvious locations within the city, offer their services to the harried or unwary pedestrian at peak traffic hours. Several plans for a subway system have been proposed for Lima, in part to overcome the obvious problems of the heavily congested and polluted centre but also to interconnect the peripheral suburbs more effectively and thus divert much traffic from the central city.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The problems of controlling Lima's growth have proved difficult, but those of municipal administration have become almost insoluble. Metropolitan Lima consists of the provinces of Lima and Callao, the latter having departmental status, which are further divided into a total of 45 political districts. Each province and each district is administratively autonomous, so that citywide planning and development can be undertaken only by means of negotiated decisions. The capital district of Lima, with its long-established expertise in urban administration, has repeatedly called for the creation of a metropolitan authority that could more efficiently confront the many issues facing the region. Local district autonomy, however, which was won only after great political effort, has become a major obstacle to any unified approach, although a municipal law enacted in 1984 created a Metropolitan Council for Greater Lima (an assembly of district mayors) as well as agencies for improving cooperation between district councils and sharing technical assistance.

Structure

The system of generating and spending revenues in metropolitan Lima provides an example of the problems of interdistrict coordination. Since 1983 each district has been able both to generate its own revenues and to utilize them as it sees fit. Thus, there has been a growing disparity in the quality of services between the wealthy districts, which can generate adequate revenues for their needs, and the poor districts, which not only generate inadequate revenues but also are in most need of such services as water, sewers, electricity, and paved streets.

The differences in income and expenditures between rich and poor districts are, to some extent, paralleled by distinctive party affiliations and voting behaviour. The poorer districts have generally supported candidates from left-wing parties, while the more affluent suburbs have supported centre-right candidates. This interparty rivalry has hampered efforts at improving cooperation between districts as well as between the municipal and national government.

Services. The rapidity and scale of Lima's growth have placed great strains upon the provision of public services. Potable water, which in the past was obtained from the Rímac and from shallow local wells, now must be brought in via lakes and diverted rivers in the high Andes. Even so, it has been predicted that rising demand in Lima will make it necessary to find new water sources by the close of the 20th century. Equally difficult has been the provision of electricity. Only with the completion in the early 1970s of the expensive hydroelectric project on the Río Mantaro in Huancavelica department has affordable power been available for Lima's industry and residential population. These new sources of water and power, however, have been at the expense of the impoverished Andean departments that have provided them; the demand for payment for their resources by these departments has threatened to further strain Lima's municipal finances.

Within the capital itself the problems of providing services have been legion. Most municipalities have had barely enough income to finance their routine operations, with nothing left over to finance new projects. In addition, municipalities that have been able to allocate money for improved services often have been unable to adequately plan and execute what usually have been complex and

highly technical projects. Finally, even when these projects have been built it has seldom been possible, given the penurious state of the majority of the population, to require payment for the actual cost of the services.

Caught between the need for inner-city renewal and suburban expansion, most municipalities have turned to the national government and such international agencies as the World Bank for assistance. Their argument has been that Lima's problems have become national problems and, as such, require national solutions. The response of the national government has been to point out that if the problems of the rural poor are not solved, especially in the departments of the southern Andes, more and more of those people will move to Lima, further compounding the city's difficulties.

CULTURAL LIFE

In spite of the many and complex problems that confront those who live in Lima, it is still the dominant and most vibrant cultural centre of Peru. Those who want to study almost any subject at the advanced level must go there. Lima contains the oldest and most distinguished universities in the country—including Universidad Nacional Mayor de San Marcos de Lima (National University of San Marcos), founded in 1551, and the Pontificia Universidad Católica del Perú (Pontific Catholic University of Peru), founded in 1917—as well as numerous other schools. Nearly all of the major academies, learned societies, and research institutes are located in metropolitan Lima, as are the national cultural institutions.

The numerous museums in the metropolitan area display the richness of Peru's pre-Columbian and colonial past. Within Lima itself are the well-restored burial sites (*huacas*) of the pre-Inca coastal cultures, and south of the city stand the remains of Pachacamac, one of Peru's largest pre-Hispanic cities. Dozens of other historic sites await funds for excavation and investigation, but almost all are threatened by urban construction.

Media

Lima has several daily newspapers—*El Comercio,* founded in 1839, is the country's oldest—and numerous weekly periodicals, among which the magazine *Caretas* has become established as the newsweekly of Peru. Several television stations broadcast in colour, and some two dozen radio stations provide Lima with excellent coverage. Lima is not a city of bookstores or of book readers: the electronic media and a continual shortage of paper have combined to limit the circulation of the printed word. For the majority of lower-class Limeños the most popular reading materials are the comic books and dime novels that can be rented from street-corner stalls.

Recreation in Lima takes many forms, but perhaps no sports are more important than association football (soccer) for men and volleyball for women. Local soccer clubs have large and devoted followings. Other popular sports include horse racing, cockfighting, bullfighting, swimming, surfing, golf, tennis, and polo. Dozens of cinemas, theatre clubs, and discotheques provide nightlife, and there are scores of *peñas,* nightclubs featuring folk music. The music of Lima, as symbolized in the works of Chabuca Granda and Alicia Maguiño, is always popular, and it has enjoyed a renewed interest on the part of the public at large.

Both in the fashionable international-quality restaurants of central Lima and the bay area and in the hundreds of lesser cafés, *cevicherías,* and *picanterías* can be found a delicious variety of food. Fortunately for Lima, the migrants from other areas of Peru carried with them their highly flavoured regional dishes, making the city a gastronome's delight. Added to these foods are the excellent local beers, grape brandy (*pisco*), and inexpensive wines that are available.

One of the consequences of the massive migration to Lima has been the reinforcement of cultural ties between the capital's new urban communities and their localities of origin. Provincial and district clubs and associations celebrate weekly with songs, dances, and foods typical of the distinctive regions. Much of Peru's folklore can be learned in the heart of Lima itself. Yet *cholo* and *huachafa* (a term in Lima for those who are at the crest of fashion) alike share the pleasures of the beach and the ever-popular social equalizers, beer and *ceviche,* or *cebiche* (a typical coastal dish of marinated fish).

History

PRE-COLUMBIAN AND COLONIAL PERIODS

The area around Lima has been inhabited for thousands of years. Urban communities of significant size date from the pre-Inca Early Intermediate Period (*c.* 200 BC–AD 600), the most important being Pachacamac, which was an important religious site in both pre-Inca and Inca times. Much of the ransom demanded by the conquistador Francisco Pizarro for the Inca chief Atahuallpa (Atahualpa) was obtained from Pachacamac.

Founding of Lima

The Spanish city of Lima was founded by Pizarro on Jan. 6, 1535, which, being Epiphany, prompted the name Ciudad de los Reyes ("City of Kings"). Although the name never stuck, Lima soon became the capital of the new viceroyalty of Peru, chosen over the old Inca capital of Cuzco to the southeast because the coastal location facilitated communication with Spain.

Lima developed into the centre of wealth and power for the entire viceroyalty: as the seat of the *audiencia* (high court), it administered royal justice; and being the headquarters in the viceroyalty of the Inquisition, it pronounced on religious and moral matters. From the late 17th to the mid-19th century, however, Lima grew very slowly in area and population. The city was devastated by a powerful earthquake in 1746. Although the city was rebuilt in grandiose fashion, influenced by the European Enlightenment, it remained politically conservative and socially stratified. Lima maintained its loyalty during the struggles for Latin-American independence in the early 19th century, Peru becoming the last mainland colony to declare its independence from Spain (July 1821).

THE MODERN CITY

Lima's development into a modern city began after the completion of the Lima–Callao railroad in 1851. Interurban railway links to Miraflores, Ancón, and Chosica followed in the next 20 years and provided the opportunity for suburban growth. The small, compact, pedestrian city gradually lost its wealthier residents, who physically distanced themselves from the lower classes by building mansions in and around Miraflores.

A new wave of urban expansion in the 1920s and '30s was set off by the automobile and the subsequent road-building program that improved transportation not only within the capital but also between Lima and other parts of the country. For the first time, migrants could reach Lima relatively easily from every corner of the country, and this rich, powerful, and modernizing centre became a national magnet. The consequences for Lima were drastic. Between 1940 and 1980 some 2,000,000 people moved into the capital. Hundreds of thousands of shanties were constructed on the bare, unoccupied slopes that rose above the red-tiled roofs of the inner suburbs and on the flat desert benches that encircled Lima on all sides. Individual acts of occupying unused and unclaimed pieces of land gave way to well-planned "invasions" involving hundreds of new city residents. So enormous became the numbers of the self-help housing units that the government finally capitalized on the residents' initiatives, awarding titles to the land and trying to provide basic services. The problems of overcrowding, inadequate services, and poverty have become so severe, however, that the suggestion has been made to move the capital from Lima to Cuzco.

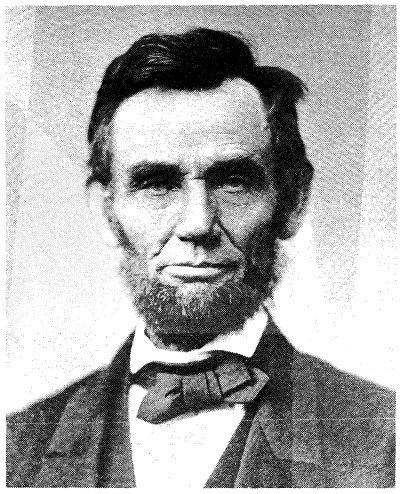
BIBLIOGRAPHY. JUAN BROMLEY, *La fundación de la ciudad de los reyes* (1935), and JUAN BROMLEY and JOSÉ BARBAGELATA, *Evolución urbana de Lima* (1945), provide the best introductions to the city's evolution up to the 20th century. More recent analyses of Lima's growing pains are HENRY A. DIETZ, *Poverty and Problem-Solving Under Military Rule: The Urban Poor in Lima, Peru* (1980); SUSAN LOBO, *A House of My Own: Social Organization in the Squatter Settlements of Lima, Peru* (1982); and PETER LLOYD, *The "Young Towns" of Lima: Aspects of Urbanization in Peru* (1980). An excellent overview of the political dimensions of Lima's development is provided in *Qué hacemos con Lima? Que dicen los candidatos?* (1983).

(D.J.Ro.)

Lincoln

The 16th president of the United States, Abraham Lincoln preserved the Union during the Civil War and brought about the emancipation of the slaves. Among American heroes, Lincoln continues to have a unique appeal for his fellow countrymen and also for people of other lands. This charm derives from his remarkable life story—the rise from humble origins, the dramatic death—and from his distinctively human and humane personality as well as from his historical role as saviour of the Union and emancipator of the slaves. His relevance endures and grows especially because of his eloquence as a spokesman for democracy. In his view, the Union was worth saving not only for its own sake but also because it embodied an ideal, the ideal of self-government, which was of interest to the people of the entire world. Hence the universality of his continuing appeal.

By courtesy of the Library of Congress, Washington, D.C.



Lincoln, 1863.

LIFE

Born on February 12, 1809, in a backwoods cabin three miles south of Hodgenville, Kentucky, Lincoln was two years old when he was taken to a farm in the neighbouring valley of Knob Creek. His earliest memories were of this home and, in particular, of a flash flood that once washed away the corn and pumpkin seeds he had helped his father plant. The father, Thomas Lincoln, was descended from a weaver's apprentice who had migrated from England to Massachusetts in 1637. Though much less prosperous than some of his Lincoln forebears, Thomas was a sturdy pioneer. On June 12, 1806, he married Nancy Hanks. The Hanks genealogy is difficult to trace, but Nancy appears to have been of illegitimate birth. She has been described as "stoop-shouldered, thin-breasted, sad," and fervently religious. Thomas and Nancy Lincoln had three children: Sarah, Abraham, and Thomas (died in infancy).

Childhood and youth. In December 1816, faced with a lawsuit challenging the title to his Kentucky farm, Thomas Lincoln moved with his family to southwestern Indiana. There, as a "squatter" on public land, he hastily put up a "half-faced camp"—a crude structure of logs and boughs with one side open to the weather—in which the family took shelter behind a blazing fire. Soon he built a permanent cabin, and later he bought the land on which it stood. Abraham helped to clear the fields and to take care of the crops but early acquired a dislike for hunting and fishing. In afteryears he recalled the "panther's scream," the bears that "preyed on the swine," and the poverty of Indiana frontier life, which was "pretty pinching at times." The unhappiest period of his boyhood followed the death of his mother in the autumn of 1818. As a ragged nine year old, he saw her buried in the forest, then faced a winter without the warmth of a mother's love. Fortunately, before the onset of a second winter, Thomas Lincoln brought home from Kentucky a new wife for himself, a new mother for the children. Sarah Bush Johnston Lincoln, a widow with two girls and a boy of her own, had energy and affection to spare. She ran the household with an even hand, treating both sets of children as if she had borne them all; but she became especially fond of Abraham, and he of her. He afterward referred to her as his "angel mother."

Mother's death

This stepmother doubtless encouraged Lincoln's taste for reading, yet the original source of his desire to learn remains something of a mystery. Both of his parents were almost completely illiterate, and he himself received little formal education. He once said that, as a boy, he had gone to school "by littles"—a little now and a little then—and his entire schooling amounted to no more than one year's attendance. His neighbours later recalled how he used to trudge for miles to borrow a book. According to his own statement, however, his early surroundings provided "absolutely nothing to excite ambition for education. Of course, when I came of age I did not know much. Still, somehow, I could read, write, and cipher to the rule of three; but that was all." Apparently the young Lincoln did not read a large number of books but thoroughly absorbed the few that he did read. These included Parson Weems's *Life and Memorable Actions of George Washington* (with its story of the little hatchet and the cherry tree), *Robinson Crusoe, Pilgrim's Progress,* and Aesop's *Fables.* From his earliest days he must have had some familiarity with the Bible, for it doubtless was the only book his family owned.

In March 1830 the Lincoln family undertook a second migration, this one to Illinois, with Lincoln himself driving the team of oxen. Having just reached the age of 21, he was about to begin life on his own. Six feet four inches tall, he was rawboned and lanky but muscular and physically powerful. He was especially noted for the skill and strength with which he could wield an ax. He spoke with a backwoods twang and walked in the long-striding, flatfooted, cautious manner of a plowman. Good-natured though somewhat moody, talented as a mimic and storyteller, he readily attracted friends. He was yet to demonstrate whatever other abilities he possessed.

Migration to Illinois

After his arrival in Illinois, having no desire to be a farmer, Lincoln tried his hand at a variety of occupations. As a "rail splitter" he helped to clear and fence his father's new farm. As a flatboatman, he made a voyage down the Mississippi River to New Orleans. (This was his second visit to that city, his first having been made in 1828, while he still lived in Indiana.) On his return he settled in New Salem, a village of about 25 families on the Sangamon River. There he worked from time to time as storekeeper, postmaster, and surveyor. With the coming of the Black Hawk War (1832), he enlisted as a volunteer and was elected captain of his company. Afterward he joked that he had seen no "live, fighting Indians" during the war but had had "a good many bloody struggles with the mosquitoes." Meanwhile, aspiring to be a legislator, he was defeated in his first try and then repeatedly reelected to the state assembly. He considered blacksmithing as a trade but finally decided in favour of the law. Already he had taught himself grammar and mathematics, and now he began to study lawbooks. In 1836, having passed the bar examination, he began to practice law.

Prairie lawyer. The next year he moved to Springfield, Illinois, the new state capital, which offered many more opportunities for a lawyer than New Salem did. At first he was a partner of John T. Stuart; then of Stephen T. Logan; and finally, from 1844 on, of William H. Herndon. Nearly 10 years younger than Lincoln, Herndon was

more widely read, more emotional at the bar, and generally more extreme in his views. Yet this partnership seems to have been as nearly perfect as such human arrangements ever are. Lincoln and Herndon kept few records of their law business, and they split the cash between them whenever either of them was paid. It seems they had no money quarrels.

Within a few years after his removal to Springfield, Lincoln was earning from $1,200 to $1,500 annually, at a time when the governor of the state received a salary of $1,200 and circuit judges only $750. He had to work hard.

Early legal career

To keep himself busy he found it necessary not only to practice in the capital but also to follow the court as it made the rounds of its circuit. Each spring and fall he would set out by horseback or buggy to travel hundreds of miles over the thinly settled prairie, from one little county seat to another. Most of the cases were petty and the fees small.

The coming of the railroads, especially after 1850, made travel easier and practice more remunerative. Lincoln served as a lobbyist for the Illinois Central Railroad to assist it in getting a charter from the state, and thereafter he was retained as a regular attorney for that railroad. After successfully defending the company against the efforts of McLean County to tax its property, he received the largest single fee of his legal career—$5,000. (He had to sue the Illinois Central in order to collect the fee.) He also handled cases for other railroads and for banks, insurance companies, mercantile and manufacturing firms. In one of his finest performances before the bar, he saved the Rock Island Bridge, the first to span the Mississippi River, from the threat of the river transportation interests that demanded the bridge's removal. His business included a number of patent suits and criminal trials. One of his most effective and famous pleas had to do with a murder case. A witness claimed that, by the light of the moon, he had seen Duff Armstrong, an acquaintance of Lincoln's, take part in a killing. Referring to an almanac for proof, Lincoln argued that the night had been too dark for the witness to have seen anything clearly, and with a sincere and moving appeal he won an acquittal.

By the time he began to be prominent in national politics, about 20 years after launching upon his legal career, Lincoln had made himself one of the most distinguished and successful lawyers in Illinois. He was noted not only for his shrewdness and practical common sense, which enabled him always to see to the heart of any legal case, but also for his invariable fairness and utter honesty.

Private life. While residing in New Salem, Lincoln was acquainted with Ann Rutledge. Apparently he was fond of her, and certainly he grieved with the entire community at her untimely death, in 1835, at the age of 19. Afterward stories were told of a grand romance between Abraham and Ann, but these stories lack the support of sound historical evidence. A year after the death of Miss Rutledge, Lincoln was carrying on a halfhearted courtship with Mary Owens. Miss Owens concluded that Lincoln was "deficient in those little links which make up the chain of woman's happiness." She turned down his proposal.

Marriage to Mary Todd

So far as can be known, the first and only real love of Lincoln's life was Mary Todd. High-spirited, quick-witted, and well-educated, Miss Todd came from a rather distinguished Kentucky family, and her Springfield relatives belonged to the social aristocracy of the town. Some of them frowned upon her association with Lincoln, and from time to time he too had doubts whether he ever could make her happy. Nevertheless, they became engaged. Then, on a day in 1841 that Lincoln recalled as the "fatal first of January," they broke the engagement, apparently on his initiative. For some time after that, he was overwhelmed by a mood of terrible depression and despondency. Finally the two were reconciled and on November 4, 1842, were married.

Four children, all boys, were born to the Lincolns. Robert Todd, the eldest and only one to survive to adulthood, was never very close to his father. Edward Baker was nearly four when he died, and William Wallace was 11. Thomas, affectionately known as "Tad," outlived his father; Tad, who had a cleft palate and a lisp, was Lincoln's favourite. Lincoln left the upbringing of his sons largely to their mother, who was alternately strict and lenient in her treatment of them.

The Lincolns had a mutual affectionate interest in the doings and welfare of their boys, were fond of one another's company, and missed each other when apart, as existing letters show. Like most married couples, the Lincolns also had their domestic quarrels, which sometimes were hectic but which undoubtedly were exaggerated by contemporary gossips. Mrs. Lincoln suffered from recurring headaches, fits of temper, and a sense of insecurity and loneliness that was intensified by her husband's long absences on the lawyer's circuit. After his election to the presidency, she was afflicted in spirit by the death of her son Willie, by the ironies of war that made enemies of Kentucky relatives and friends, and by the unfair public criticisms of her as mistress of the White House. She lost all money sense and ran up embarrassing bills. She also put on some painful scenes of wifely jealousy. At last, in 1875, she was officially declared insane, but that was after she had undergone the further shock of seeing her husband murdered at her side. During their earlier married life, Mrs. Lincoln unquestionably encouraged her husband and served as a prod to his own ambition. During their later years together she probably strengthened and tested his innate qualities of tolerance and patience.

Religious views

With his wife, Lincoln attended Presbyterian services in Springfield and in Washington but never joined any church. He once explained:

> When any church will inscribe over its altar, as its sole qualification for membership, the Saviour's condensed statement of the substance of both Law and Gospel, "Thou shalt love the Lord thy God with all thy heart, and with all thy soul, and with all thy mind, and thy neighbor, as thyself," that church will I join with all my heart and all my soul.

Early in life he had been something of a skeptic and freethinker. His reputation had been such that, as he once complained, the "church influence" was used against him in politics. When running for Congress in 1846, he issued a handbill to deny that he ever had "spoken with intentional disrespect of religion." He went on to explain that he had believed in the doctrine of necessity—"that is, that the human mind is impelled to action, or held in rest by some power over which the mind itself has no control." Throughout his life he also believed in dreams and other enigmatic signs and portents. As he grew older, and especially after he became president and faced the soul-troubling responsibilities of the Civil War, he developed a profound religious sense, and he increasingly personified necessity as God. He came to look upon himself quite humbly as an "instrument of Providence" and to view all history as God's enterprise. "In the present civil war," he wrote in 1862, "it is quite possible that God's purpose is something different from the purpose of either party—and yet the human instrumentalities, working just as they do, are of the best adaptation to effect His purpose."

Lincoln was fond of the Bible and knew it well. He also was fond of Shakespeare. In private conversation he used many Shakespearean allusions, discussed problems of dramatic interpretation with considerable insight, and from memory recited long passages with rare feeling and understanding. He liked the essays of John Stuart Mill, particularly the famous one on liberty, but disliked heavy or metaphysical works.

Though he enjoyed the poems of Lord Byron and Robert Burns, his favourite piece of verse was the work of an obscure Scottish poet, William Knox. Lincoln often quoted Knox's lines beginning: "Oh! why should the spirit of mortal be proud?" He liked to relax with the comic writings of Petroleum V. Nasby, Orpheus C. Kerr, and Artemus Ward, or with a visit to the popular theatre.

Early politics. When Lincoln first entered politics, Andrew Jackson was president. Lincoln shared the sympathies that the Jacksonians professed for the common man, but he disagreed with the Jacksonian view that the government should be divorced from economic enterprise. "The legitimate object of government," he was later to say, "is to do for a community of people whatever they need to have done, but cannot do at all, or cannot do so well, for

themselves, in their separate and individual capacities." He most admired Henry Clay and Daniel Webster among the prominent politicians of the time. Clay and Webster advocated using the powers of the federal government to encourage business and develop the country's resources by means of a national bank, a protective tariff, and a program of internal improvements for facilitating transportation. In Lincoln's view, Illinois and the West as a whole desperately needed such aid to economic development. From the outset, he associated himself with the Clay and Webster party, the Whigs.

Member of Illinois State Legislature

As a Whig member of the Illinois State Legislature, to which he was elected four times from 1834 to 1840, he devoted himself to a grandiose project for constructing with state funds a network of railroads, highways, and canals. Whigs and Democrats joined in passing an omnibus bill for these undertakings, but the Panic of 1837 and the ensuing business depression brought about the abandonment of most of them. While in the legislature he demonstrated that, though opposed to slavery, he was no abolitionist. Resolutions were introduced, in 1837, in response to the mob murder of Elijah Lovejoy, an antislavery newspaperman of Alton. Instead of denouncing lynch law, these resolutions condemned abolitionist societies and upheld slavery within the Southern states as "sacred" by virtue of the federal Constitution. Lincoln refused to vote for the resolutions. Together with a fellow member he drew up a protest against them. This maintained, on the one hand, that slavery was "founded on both injustice and bad policy" and, on the other, that "the promulgation of abolition doctrines tends rather to increase than to abate its evils."

During his single term in Congress (1847–49), Lincoln, as the lone Whig from Illinois, gave little attention to legislative matters as such. He proposed a bill for the gradual and compensated emancipation of slaves in the District of Columbia, but the bill was to take effect only with the approval of the "free white citizens" of the district. It displeased abolitionists as well as slaveholders and never was seriously considered.

Involvement in presidential politics

Much of his time Lincoln devoted to presidential politics, to unmaking one president, a Democrat, and making another, a Whig. He found an issue and a candidate in the Mexican War. With his "spot resolutions" he challenged the statement of Pres. James K. Polk that Mexico had started the war by shedding American blood upon American soil. Along with other members of his party, Lincoln voted to condemn Polk and the war while voting supplies for carrying it on. At the same time he laboured for the nomination and election of the war hero Zachary Taylor. After Taylor's success at the polls, Lincoln expected to be named commissioner of the general land office as a reward for his campaign services, and he was bitterly disappointed when he failed to get the job. His criticisms of the war, meanwhile, had not been popular among the voters in his own congressional district. At the age of 40, frustrated in politics, he seemed to be at the end of his public career.

For about five years he took little part in politics, and then a new sectional crisis gave him a chance to reemerge and rise to statesmanship. In 1854 his political rival Stephen A. Douglas manoeuvred through Congress a bill for reopening the entire Louisiana Purchase to slavery and allowing the settlers of Kansas and Nebraska (with "popular sovereignty") to decide for themselves whether to permit slaveholding in those territories. The Kansas–Nebraska Act provoked violent opposition in Illinois and the other states of the old Northwest. It gave rise to the Republican Party while speeding the Whig Party on the way to disintegration. Along with many thousands of other homeless Whigs, Lincoln soon became a Republican (1856). Before long, some prominent Republicans in the East talked of attracting Douglas to the Republican fold, and with him his Democratic following in the West. Lincoln would have none of it. He was determined that he, not Douglas, should be the Republican leader of his state and section. In their basic views, he and Douglas were not quite so far apart as they seemed in the heat of political argument. Neither was an abolitionist, neither a proslavery man. But Lincoln, unlike Douglas, insisted that Congress must exclude slavery from the territories. He disagreed with Douglas' belief that the territories were by nature unsuited to the slave economy and that no congressional legislation was needed to prevent the spread of slavery into them. He declared (1858): "*A house divided against itself cannot stand.* I believe the government cannot endure permanently half slave and half free." He predicted that the country eventually would become "all one thing, or all the other." Again and again he insisted that the civil liberties of every U.S. citizen, white as well as black, were at stake. The territories must be kept free, he further said, because "new free states" were "places for poor people to go and better their condition." He agreed with Thomas Jefferson and other founding fathers, however, that slavery should be merely contained, not directly attacked. In the Lincoln–Douglas debates of 1858, while contesting for Douglas' seat in the United States Senate, he drove home the inconsistency between Douglas' "popular sovereignty" principle and the Dred Scott decision (1857), in which the U.S. Supreme Court held that Congress could not constitutionally exclude slavery from the territories. Though he failed to obtain the Senate seat, Lincoln gained national recognition and soon began to be mentioned as a presidential prospect for 1860.

Lincoln–Douglas debates

On May 18, 1860, after Lincoln and his friends had made skillful preparations, he was nominated on the third ballot at the Republican Convention in Chicago. He then put aside his law practice and, though making no stump speeches, gave his full time to the direction of his campaign. His "main object," he had written, was to "hedge against divisions in the Republican ranks," and he counselled party workers to "say nothing on points where it is probable we shall disagree." With the Republicans united, the Democrats divided, and a total of four candidates in the field, he carried the election on November 6. No one in the Deep South voted for him and no more than 40 out of 100 in the country as a whole. Still, the popular votes were so distributed that he won a clear and decisive majority in the electoral college.

President Lincoln. After Lincoln's election and before his inauguration, the state of South Carolina proclaimed its withdrawal from the Union. To forestall similar action by other Southern states, various compromises were proposed in Congress. The most important, the Crittenden Compromise, included constitutional amendments (1) guaranteeing slavery forever in the states where it already existed and (2) dividing the territories between slavery and freedom. Though Lincoln had no objection to the first of these amendments, he was unalterably opposed to the second and indeed to any scheme infringing in the slightest upon the free-soil plank of his party's platform. "I am inflexible," he privately wrote. He feared that a territorial division, by sanctioning the principle of slavery extension, would only encourage planter imperialists to seek new slave territory south of the American border and thus would "put us again on the highroad to a slave empire." From his home in Springfield he advised Republicans in Congress to vote against a division of the territories. The proposal was killed in committee. Six additional states then seceded and, with South Carolina, combined to form the Confederate States of America.

Secession of the Confederate states

So, before Lincoln took office, a disunion crisis was upon the country. Attention, North and South, focussed in particular upon Ft. Sumter, in Charleston Harbor. This fort, still under construction, was garrisoned by U.S. troops under Maj. Robert Anderson. The Confederacy claimed it and, from other harbour fortifications, threatened it. Foreseeing trouble, Lincoln, while still in Springfield, confidentially requested Winfield Scott, general in chief of the U.S. Army, to be prepared "to either *hold,* or *retake,* the forts, as the case may require, at, and after the inauguration." In his inaugural address (March 4, 1861), besides upholding the Union's indestructibility and appealing for sectional harmony, Lincoln restated his Sumter policy as follows:

> The power confided to me, will be used to hold, occupy, and possess the property, and places belonging to the government, and to collect the duties and imposts; but beyond what may be necessary for these objects, there will be no invasion—no using of force against, or among the people anywhere.

Then, near the end, addressing the absent Southerners:

"You can have no conflict, without being yourselves the aggressors."

No sooner was he in office than Lincoln received word that the Sumter garrison, unless supplied or withdrawn, would shortly be starved out. Still, for about a month, Lincoln delayed to act. He was beset by contradictory advice. On the one hand, General Scott, Secretary of State William H. Seward, and others urged him to abandon the fort; and Seward, through a go-between, gave a group of Confederate commissioners to understand that the fort would in fact be abandoned. On the other hand, many Republicans insisted that any show of weakness would bring disaster to their party and to the Union. Finally Lincoln ordered the preparation of two relief expeditions, one for Ft. Sumter and the other for Ft. Pickens, in Florida. (He afterward said he would have been willing to withdraw from Sumter if he could have been sure of holding Pickens.) Before the Sumter expedition, he sent a messenger to tell the South Carolina governor:

> I am directed by the President of the United States to notify you to expect an attempt will be made to supply Fort-Sumpter with provisions only; and that, if such attempt be not resisted, no effort to throw in men, arms, or ammunition, will be made, without further notice, or in case of an attack upon the Fort.

Without waiting for the arrival of Lincoln's expedition, the Confederate authorities presented to Major Anderson a demand for Sumter's prompt evacuation, which he refused. On April 12, 1861, at dawn, the Confederate batteries in the harbour opened fire.

Attack on Ft. Sumter: Civil War

"Then, and thereby," Lincoln informed Congress when it met on July 4, "the assailants of the Government, began the conflict of arms." The Confederates, however, accused him of being the real aggressor. They said he had cleverly manoeuvred them into firing the first shot so as to put upon them the onus of war guilt. Though some historians have repeated this charge, it appears to be a gross distortion of the facts. Lincoln was determined to preserve the Union; to do so he thought he must take a stand against the Confederacy, and he concluded he might as well take this stand at Sumter.

Lincoln's primary aim was neither to provoke war nor to maintain peace. In preserving the Union, he would have been glad to preserve the peace also, but he was ready to risk a war that he thought would be short.

After the firing on Ft. Sumter, Lincoln called upon the state governors for troops (Virginia and three other states of the upper South responded by joining the Confederacy). He then proclaimed a blockade of the Southern ports. These steps—the Sumter expedition, the call for volunteers, and the blockade—were the first important decisions of Lincoln as commander in chief of the army and navy. He still needed a strategic plan and a command system for carrying it out.

General Scott advised him to avoid battle with the Confederate forces in Virginia, to get control of the Mississippi River, and by tightening the blockade to hold the South in a gigantic squeeze. Lincoln had little confidence in Scott's comparatively passive and bloodless "Anaconda" plan. He believed the war must be actively fought if it ever was to be won. Overruling Scott, he ordered a direct advance on the Virginia front, which resulted in defeat and rout for the Federal forces at Bull Run (July 21, 1861). After a succession of more or less sleepless nights, Lincoln produced a set of memorandums on military policy. His basic thought was this: the armies should advance concurrently on several fronts and should move in such directions as to hold and use the support of Unionists in Missouri, Kentucky, western Virginia, and eastern Tennessee. He later explained:

Lincoln's strategy as commander in chief

> I state my general idea of this war to be that we have the *greater* numbers, and the enemy has the *greater* facility of concentrating forces upon points of collision; that we must fail, unless we can find some way of making *our* advantage an over-match for *his;* and that this can only be done by menacing him with superior forces at *different* points, at the *same* time.

This, with the naval blockade, comprised the essence of Lincoln's strategy.

From 1861 to 1864, while hesitating to impose his ideas upon his generals, Lincoln experimented with command personnel and organization. Accepting the resignation of Scott (November 1861), he put George B. McClellan in charge of the armies as a whole. After a few months, disgusted by the slowness of McClellan, he demoted him to the command of the Army of the Potomac alone. He questioned the soundness of McClellan's plans for the peninsular campaign, repeatedly compelled McClellan to alter them, and, after the Seven Days' Battles before Richmond (June–July 1862), ordered him to give them up. Then he tried a succession of commanders for the army in Virginia—John Pope, McClellan again, Ambrose E. Burnside, Joseph Hooker, and George Gordon Meade—but was disappointed with each of them in turn. Meanwhile, he had in Henry W. Halleck a general in chief who gave advice and served as a liaison with field officers but who shrank from making important decisions. For nearly two years the Federal armies had no very effective unity of command. President Lincoln, General Halleck, and War Secretary Edwin M. Stanton acted as an informal council of war. Lincoln, besides transmitting official orders through Halleck, also communicated directly with the generals, sending personal suggestions in his own name. To generals opposing Robert E. Lee, he suggested that the object was to destroy Lee's army, not to capture Richmond or to drive the invader from Northern soil.

Finally Lincoln looked to the West for a top general. He admired the Vicksburg Campaign of Ulysses S. Grant. Nine days after the Vicksburg surrender (which occurred on July 4, 1863), he sent Grant a "grateful acknowledgment for the almost inestimable service" he had done the country. Lincoln sent also an admission of his own error. He said he had expected Grant to bypass Vicksburg and go on down the Mississippi, instead of crossing the river and turning back to approach Vicksburg from the rear. "I feared it was a mistake," he wrote in his letter of congratulations. "I now wish to make the personal acknowledgment that you were right, and I was wrong."

Choice of Grant as commander of Federal armies

In March 1864 Lincoln promoted Grant to lieutenant general and gave him command of all the Federal armies. At last Lincoln had found a man who, with such able subordinates as William T. Sherman, Philip Sheridan, and George H. Thomas, could put into effect those parts of Lincoln's concept of a large-scale, coordinated offensive that still remained to be carried out. Grant was only a member, though an important one, of a top-command arrangement that Lincoln eventually had devised. Overseeing everything was Lincoln himself, the commander in chief. Taking the responsibility for men and supplies was Stanton, the secretary of war. Serving as a presidential adviser and as a liaison with military men was Halleck, the chief of staff. And directing all the armies, while accompanying Meade's Army of the Potomac, was Grant, the general in chief. Thus Lincoln pioneered in the creation of a high command, an organization for amassing all the energies and resources of a people in the grand strategy of total war. He combined statecraft and the overall direction of armies with an effectiveness that year by year increased. His achievement is all the more remarkable in view of his lack of training and experience in the art of warfare. This lack may have been an advantage as well as a handicap. Unhampered by outworn military dogma, Lincoln could all the better apply his practical insight and common sense—some would say his military genius—to the winning of the Civil War.

Lincoln and the slavery question

There can be no doubt of Lincoln's deep and sincere devotion to the cause of personal freedom. Before his election to the presidency he had spoken often and eloquently on the subject. In 1854, for example, he said he hated the Douglas attitude of indifference toward the possible spread of slavery to new areas. "I hate it because of the monstrous injustice of slavery itself," he declared. "I hate it because it deprives our republican example of its just influence in the world; enables the enemies of free institutions with plausibility to taunt us as hypocrites. . . ." In 1855, writing to his friend Joshua Speed, he recalled a steamboat trip the two had taken on the Ohio River 14 years earlier. "You may remember, as I well do," he

said, "that from Louisville to the mouth of the Ohio there were, on board, ten or a dozen slaves, shackled together with irons. That sight was a continual torment to me; and I see something like it every time I touch the Ohio, or any other slave-border."

Yet, as president, Lincoln was at first reluctant to adopt an abolitionist policy. There were several reasons for his hesitancy. He had been elected on a platform pledging no interference with slavery within the states, and in any case he doubted the constitutionality of federal action under the circumstances. He was concerned about the possible difficulties of incorporating nearly 4,000,000 Negroes, once they had been freed, into the nation's social and political life. Above all, he felt that he must hold the border slave states in the Union, and he feared that an abolitionist program might impel them, in particular his native Kentucky, toward the Confederacy. So he held back while others went ahead. When Gen. John C. Frémont and Gen. David Hunter, within their respective military departments, proclaimed freedom for the slaves of disloyal masters, Lincoln revoked the proclamations. When Congress passed confiscation acts (in 1861 and 1862), he refrained from a full enforcement of the provisions authorizing him to seize slave property. And when Horace Greeley in the *New York Tribune* appealed to him to enforce these laws, Lincoln patiently replied (August 22, 1862):

> My paramount object in this struggle is to save the Union, and is not either to save or to destroy slavery. If I could save the Union without freeing any slave I would do it; and if I could save it by freeing all the slaves I would do it; and if I could save it by freeing some and leaving others alone, I would also do that.

Meanwhile, in response to the rising antislavery sentiment, Lincoln came forth with an emancipation plan of his own. According to his proposal, the slaves were to be freed by state action, the slaveowners were to be compensated, the federal government was to share the financial burden, the emancipation process was to be gradual, and the freedmen were to be colonized abroad. Congress indicated its willingness to vote the necessary funds for the Lincoln plan, but none of the border slave states were willing to launch it, and in any case few Negro leaders desired to see their people sent abroad.

Emancipation Proclamation and the antislavery amendment

While still hoping for the eventual success of his gradual plan, Lincoln took a quite different step by issuing his preliminary (September 22, 1862) and his final (January 1, 1863) Emancipation Proclamation. This famous decree, which he justified as an exercise of the president's war powers, applied only to those parts of the country actually under Confederate control, not to the loyal slave states nor to the Federally occupied areas of the Confederacy. Directly or indirectly the proclamation brought freedom during the war to fewer than 200,000 slaves. Yet it had great significance as a symbol. It indicated that the Lincoln government had added freedom to reunion as a war aim, and it attracted liberal opinion in England and Europe to increased support of the Union cause.

Lincoln himself doubted the constitutionality of his step, except as a temporary war measure. After the war the slaves freed by the proclamation would have risked re-enslavement, had nothing else been done to confirm their liberty. Something else was done: the Thirteenth Amendment was added to the Constitution, and Lincoln played a large part in bringing about this change in the fundamental law. Through the chairman of the Republican National Committee he urged the party to include a plank for such an amendment in its platform of 1864. The plank, as adopted, stated that slavery was the cause of the rebellion, that the President's proclamation had aimed "a death blow at this gigantic evil," and that a constitutional amendment was necessary to "terminate and forever prohibit" it. When Lincoln was re-elected on this platform and the Republican majority in Congress was increased, he was justified in feeling, as he apparently did, that he had a mandate from the people for the Thirteenth Amendment. The newly chosen Congress, with its overwhelming Republican majority, was not to meet until after the lame duck session of the old Congress during the winter of 1864–65. Lincoln did not wait. Using his resources of patronage and persuasion upon certain of the Democrats, he managed to get the necessary two-thirds vote before the session's end. He rejoiced as the amendment went out to the states for ratification, and he rejoiced again and again as his own Illinois led off and other states followed one by one in acting favourably upon it. (He did not live to rejoice in its ultimate adoption.)

Lincoln deserves his reputation as the Great Emancipator. His claim to that honour, if it rests uncertainly upon his famous proclamation, has a sound basis in the support he gave to the antislavery amendment. It is well founded also in his greatness as the war leader who carried the nation safely through the four-year struggle that brought freedom in its train. And, finally, it is strengthened by the practical demonstrations he gave of respect for human worth and dignity, regardless of colour. During the last two years of his life he welcomed Negroes as visitors and friends in a way no president had done before. One of his friends was the distinguished former slave Frederick Douglass. Afterward Douglass wrote: "In all my interviews with Mr. Lincoln I was impressed with his entire freedom from prejudice against the colored race."

Wartime politics

To win the war, President Lincoln had to have popular support. The reunion of North and South required, first of all, a certain degree of unity in the North. But the North contained various groups with special interests of their own. Lincoln faced the task of attracting to his administration the support of as many divergent groups and individuals as possible. So he gave much of his time and attention to politics, which in one of its aspects is the art of attracting such support. Fortunately for the Union cause, he was a president with rare political skill. He had the knack of appealing to fellow politicians and talking to them in their own language. He had a talent for smoothing over personal differences and holding the loyalty of men antagonistic to one another. Inheriting the spoils system, he made good use of it, disposing of government jobs in such a way as to strengthen his administration and further its official aims.

The opposition party remained alive and strong. Its membership included war Democrats and peace Democrats, often called "Copperheads," a few of whom collaborated with the enemy. Lincoln did what he could to cultivate the assistance of the war Democrats, as in securing from Congress the timely approval of the Thirteenth Amendment. So far as feasible, he conciliated the peace Democrats. He gave heed to the complaints of one of them, Gov. Horatio Seymour of New York, in regard to the draft quota for that state. He commuted the prison sentence of another, Congressman Clement L. Vallandigham of Ohio, to banishment within the Confederate lines. In dealing with persons suspected of treasonable intent, Lincoln at times authorized his generals to make arbitrary arrests. He justified this action on the ground that he had to allow some temporary sacrifice of parts of the Constitution in order to maintain the Union and thus preserve the Constitution as a whole. He let his generals suspend several newspapers, but only for short periods, and he promptly revoked a military order suppressing the hostile *Chicago Times*. In a letter to one of his generals he expressed his policy thus:

> You will only arrest individuals and suppress assemblies or newspapers when they may be working palpable injury to the military in your charge, and in no other case will you interfere with the expression of opinion in any form or allow it to be interfered with violently by others. In this you have a discretion to exercise with great caution, calmness, and forbearance.

Considering the dangers and provocations of the time, Lincoln was quite liberal in his treatment of political opponents and the opposition press. He was by no means the dictator critics often accused him of being.

Within his own party he confronted factional divisions and personal rivalries that caused him as much trouble as did the activities of the Democrats. True, he and most of his fellow partisans agreed fairly well upon their principal economic aims. With his approval, the Republicans enacted into law the essentials of the program he had advocated from his early Whig days—a protective tariff; a national banking system; and federal aid for internal im-

provements, in particular for the construction of a railroad to the Pacific Coast. The Republicans disagreed among themselves, however, on many matters regarding the conduct and purposes of the war. Two main factions arose: the "radicals" and the "conservatives." Lincoln himself inclined in spirit toward the conservatives, but he had friends among the radicals as well, and he strove to maintain his leadership over both. In appointing his cabinet, he chose his several rivals for the 1860 nomination and, all together, gave representation to every important party group. Wisely he included the outstanding conservative, Seward, and the outstanding radical, Salmon P. Chase. Cleverly he overcame cabinet crises and kept these two opposites among his official advisers until Chase's resignation in 1864.

Reconstruction of the South

He had to deal with even more serious factional uprisings in Congress. The big issue was the "reconstruction" of the South. The seceded states of Louisiana, Arkansas, and Tennessee having been largely recovered by the Federal armies, Lincoln late in 1863 proposed his "ten percent plan," according to which new state governments might be formed when 10 percent of the qualified voters had taken an oath of future loyalty to the United States. The radicals rejected Lincoln's proposal as too lenient, and they carried through Congress the Wade-Davis Bill, which would have permitted the remaking and re-admission of states only after a majority had taken the loyalty oath. When Lincoln pocket vetoed that bill, its authors published a "manifesto" denouncing him.

Already he was the candidate of the "Union" (that is, the Republican) party for re-election to the presidency, and the Wade-Davis manifesto signalized a movement within the party to displace him as the party's nominee. He waited quietly and patiently for the movement to collapse, but even after it had done so, the party remained badly divided. A rival Republican candidate, John C. Frémont, nominated much earlier by a splinter group, was still in the field. Leading radicals promised to procure Frémont's withdrawal if Lincoln would obtain the resignation of his conservative postmaster general, Montgomery Blair. Eventually Frémont withdrew and Blair resigned. The party was reunited in time for the election of 1864.

In 1864, as in 1860, Lincoln was the chief strategist of his own electoral campaign. He took a hand in the management of the Republican Speakers' Bureau, advised state committees on campaign tactics, hired and fired government employees to strengthen party support, and did his best to enable as many soldiers and sailors as possible to vote. Most of the citizens in uniform voted Republican. He was reelected with a large popular majority (55 percent) over his Democratic opponent, General McClellan.

Lincoln's second term

In 1864 the Democratic platform called for an armistice and a peace conference, and prominent Republicans as well as Democrats demanded that Lincoln give heed to Confederate peace offers, irregular and illusory though they were. In a public letter, he stated his own conditions:

> Any proposition which embraces the restoration of peace, the integrity of the whole Union, and the abandonment of slavery, and which comes by and with an authority that can control the armies now at war against the United states will be received and considered by the Executive government of the United States, and will be met by liberal terms on other substantial and collateral points.

When conservatives protested to him against the implication that the war must go on to free the slaves, even after reunion had been won, he explained: "To me it seems plain that saying reunion and abandonment of slavery would be considered, if offered, is not saying that nothing *else* or *less* would be considered, if offered." After his re-election, in his annual message to Congress, he said: "In stating a single condition of peace, I mean simply to say that the war will cease on the part of the government, whenever it shall have ceased on the part of those who began it." On February 3, 1865, he met personally with Confederate commissioners on a steamship in Hampton Roads. He promised to be liberal with pardons if the South would quit the war, but he insisted on reunion as a precondition for any peace arrangement. In his second inaugural address he embodied the spirit of his policy in the famous words "with malice toward none; with charity for all." His terms satisfied neither the Confederate leaders nor the radical Republicans, and so no peace was possible until the final defeat of the Confederacy.

Policy toward the South

At the end of the war, Lincoln's policy for the defeated South was not clear in all its details, though he continued to believe that the main object should be to restore the "seceded States, so-called," to their "proper practical relation" with the Union as soon as possible. He possessed no fixed and uniform program for the region as a whole. As he said in the last public speech of his life (April 11, 1865), "so great peculiarities" pertained to each of the states, and "such important and sudden changes" occurred from time to time, and "so new and unprecedented" was the whole problem that "no exclusive and inflexible plan" could "safely be prescribed." With respect to states like Louisiana and Tennessee, he continued to urge acceptance of new governments set up under his 10 percent plan during the war. With respect to states like Virginia and North Carolina, he seemed willing to use the old rebel governments temporarily as a means of transition from war to peace. He was on record as opposing the appointment of "strangers" (carpetbaggers) to govern the South. He hoped that the Southerners themselves, in forming new state governments, would find some way by which whites and Negroes "could gradually live themselves out of their old relation to each other, and both come out better prepared for the new." A program of education for the freedmen, he thought, was essential for preparing them for their new status. He also suggested that the vote be given immediately to some Negroes—"as, for instance, the very intelligent, and especially those who have fought gallantly in our ranks."

On the question of reconstruction, however, Lincoln and the extremists of his own party stood even farther apart in early 1865 than a year before. Some of the radicals were beginning to demand a period of military occupation for the South, the confiscation of planter estates and their division among the freedmen, and the transfer of political power from the planters to their former slaves. In April 1865 Lincoln began to modify his own stand in some respects and thus to narrow the gap between himself and the radicals. He recalled the permission he had given for the assembling of the rebel legislature of Virginia, and he approved in principle—or at least did not disapprove—Stanton's scheme for the military occupation of southern states. After the cabinet meeting of April 14, Attorney General James Speed inferred that Lincoln was moving toward the radical position. "He never seemed so near our views," Speed believed. What Lincoln's reconstruction policy would have been, if he had lived to complete his second term, can only be guessed at.

Assassination

On the evening of April 14, 1865, John Wilkes Booth shot Lincoln as he sat in Ford's Theatre in Washington, and early the next morning Lincoln died.

REPUTATION AND CHARACTER

"Now he belongs to the ages," Stanton is supposed to have said as Lincoln breathed his last. Many thought of him as a martyr. The assassination had occurred on Good Friday, and on the following Sunday, memorable as "Black Easter," hundreds of speakers found a sermon in the event. Some of them saw more than mere chance in the fact that assassination day was also crucifixion day. One declared: "Jesus Christ died for the world; Abraham Lincoln died for his country." Thus the posthumous growth of his reputation was influenced by the timing and circumstances of his death, which won for him a kind of sainthood.

Among the many who remembered Lincoln from personal acquaintance, one was sure he had known him more intimately than any of the rest and influenced the world's conception of him more than all the others put together. That one was his former law partner Herndon. When Lincoln died, Herndon began a new career as Lincoln authority, collecting reminiscences wherever he could find them and adding his own store of memories. Though admiring Lincoln, he objected to the trend toward sanctifying the man. He saw, as the main feature of Lincoln's life, the far more than ordinary rise of a self-made man,

a rise from the lowest depths to the greatest heights—"from a stagnant, putrid pool, like the gas which, set on fire by its own energy and self-combustible nature, rises in jets, blazing, clear, and bright." To emphasize this point, Herndon gave his most eager attention to evidences of the dismal and sordid in Lincoln's background. An extremely significant event in Lincoln's development, as Herndon viewed it, was a "romance of much reality" with Ann Rutledge. Lincoln loved no one but Ann and, after her death, never ceased to grieve for her. His memory of her both saddened and inspired him. As for his wife, Mary Todd, she married him out of spite, then devoted herself to making him miserable. So Herndon would have it, and after him countless biographers and novelists and playwrights elaborated upon his views, which persist as accepted knowledge about Lincoln despite their refutation by historical scholarship.

Lincoln has become a myth as well as a man. The legendary is to be found in imaginative literature and in folklore, in poems, plays, novels, anecdotes, and the like. It is also to be found in ostensibly factual productions, including footnoted biographies and histories.

The Lincoln of legend has grown into a protean god who can assume a shape to please almost any worshipper. He is Old Abe and at the same time a natural gentleman. He is Honest Abe and yet a being of superhuman shrewdness and cunning. He is also Father Abraham, the wielder of authority, the support of the weak; and he is an equal, a neighbour, and a friend.

The greatness of Lincoln

Lincoln the man has a reputation that may be considered apart from that of Lincoln the myth. While he was still alive, his reputation began to grow, and before his death his qualities of greatness already were widely recognized. In the midst of the Civil War, for instance, the *Washington Chronicle* found a resemblance between him and George Washington in their "sure judgment"; "perfect balance of thoroughly sound faculties"; and "great calmness of temper, great firmness of purpose, supreme moral principle, and intense patriotism." The *Buffalo Express* referred to his "remarkable moderation and freedom from passionate bitterness," then added: "We do not believe that Washington himself was less indifferent to the exercise of power for power's sake." An English newspaper, the *Liverpool Post,* suggested that "no leader in a great contest ever stood so little chance of being the subject of hero worship as Abraham Lincoln," if one were to judge only by the way he looked. His long arms and legs, his grotesque figure, made him too easy to caricature and ridicule. "Yet," this newspaper concluded, "a worshiper of human heroes might possibly travel a great deal farther and fare much worse for an idol than selecting this same lanky American." His inner qualities—his faithfulness, honesty, resolution, insight, humour, and courage—would "go a long way to make up a hero," whatever the man's personal appearance.

Lincoln's best ideas and finest phrases were considered and written and rewritten with meticulous revisions. Some resulted from a slow gestation of thought and phrase through many years. One of his recurring themes—probably his central theme—was the promise and the problem of self-government. As early as 1838, speaking to the Young Men's Lyceum of Springfield on "The Perpetuation of Our Political Institutions," he recalled the devotion of his Revolutionary forefathers to the cause, and went on to say:

> Their ambition aspired to display before an admiring world, a practical demonstration of the truth of a proposition, which had hitherto been considered, at best no better, than problematical; namely, *the capability of a people to govern themselves.*

Again and again he returned to this idea, especially after the coming of the Civil War, and he steadily improved his phrasing. In his first message to Congress after the fall of Ft. Sumter, he declared that the issue between North and South involved more than the future of the United States.

> It presents to the whole family of man, the question, whether a constitutional republic, or a democracy—a government of the people, by the same people—can, or cannot, maintain its territorial integrity, against its own domestic foes.

And finally at Gettysburg he made the culminating, the supreme statement, concluding with the words:

> . . . that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion—that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

(R.N.C.)

BIBLIOGRAPHY. For the general reader, the most usable and valuable guide is PAUL M. ANGLE, *A Shelf of Lincoln Books* (1946, reissued 1972). Practically all the known writings of Lincoln himself are available in *The Collected Works of Abraham Lincoln,* ed. by ROY P. BASLER, 9 vol. (1953–55; with suppl., 1974). A judicious selection from these volumes is reprinted in *The Living Lincoln,* ed. by PAUL M. ANGLE and EARL S. MIERS (1955).

Classic multivolume biographies are JOHN G. NICOLAY and JOHN HAY, *Abraham Lincoln: A History,* 10 vol. (1890; abridged ed., 1 vol., 1966); ALBERT J. BEVERIDGE, *Abraham Lincoln, 1809–1858,* 2 vol. (1928, reissued 1971); CARL SANDBURG, *Abraham Lincoln: The Prairie Years,* 2 vol. (1926), and *The War Years,* 4 vol. (1939), both issued together in 3 vol., 1960; and JAMES G. RANDALL, *Lincoln the President,* 4 vol. (1945–55). The best of the one-volume lives are BENJAMIN P. THOMAS, *Abraham Lincoln* (1952, reissued 1973); and STEPHEN B. OATES, *With Malice Toward None* (1977). A biography for the general reader is OSCAR HANDLIN and LILIAN HANDLIN, *Abraham Lincoln and the Union* (1980).

Matters of controversy and conflicting interpretation are discussed by DAVID DONALD in *Lincoln Reconsidered,* 2nd ed. (1961); and by RICHARD N. CURRENT in *The Lincoln Nobody Knows* (1958, reprinted 1980). Important works dealing with aspects of Lincoln's statesmanship are DON E. FEHRENBACHER, *Prelude to Greatness: Lincoln in the 1850's* (1962); DAVID M. POTTER, *Lincoln and His Party in the Secession Crisis* (1942, reissued 1979); WILLIAM B. HESSELTINE, *Lincoln and the War Governors* (1948, reissued 1972); KENNETH M. STAMPP, *And the War Came* (1950, reprinted 1980); T. HARRY WILLIAMS, *Lincoln and the Radicals* (1941, reissued 1969), and *Lincoln and His Generals* (1952, reprinted 1981); CHRISTOPHER DELL, *Lincoln and the War Democrats* (1975); DAVID A. NICHOLS, *Lincoln and the Indians* (1978); and GABOR S. BORITT, *Lincoln and the Economics of the American Dream* (1978). RUTH PAINTER RANDALL writes authoritatively of Lincoln's family life in *Mary Lincoln: Biography of a Marriage* (1953), and *Lincoln's Sons* (1955). Randall effectively refutes the views of William H. Herndon, which may most conveniently be found in *Herndon's Life of Lincoln,* ed. by PAUL M. ANGLE (1930, reissued 1965). An able and realistic treatment of Lincoln's legal career is JOHN J. DUFF, *A. Lincoln, Prairie Lawyer* (1960).

Linguistics

Linguistics is the scientific study of language. The word was first used in the middle of the 19th century to emphasize the difference between a newer approach to the study of language that was then developing and the more traditional approach of philology. The differences were and are largely matters of attitude, emphasis, and purpose. The philologist is concerned primarily with the historical development of languages as it is manifest in written texts and in the context of the associated literature and culture. The linguist, though he may be interested in written texts and in the development of languages through time, tends to give priority to spoken languages and to the problems of analyzing them as they operate at a given point in time.

The field of linguistics may be conveniently divided in terms of three dichotomies: synchronic versus diachronic, theoretical versus applied, microlinguistics versus macrolinguistics. A synchronic description of a language describes the language as it is at a given time; a diachronic description is concerned with the historical development of the language and the structural changes that have taken place in it. The goal of theoretical linguistics is the construction of a general theory of the structure of language or of a general theoretical framework for the description of languages (there is some controversy as to the difference between these two conceptions of the subject); the aim of applied linguistics is the application of the findings and techniques of the scientific study of language to a variety of practical tasks, especially to the elaboration of improved methods of language teaching. The terms microlinguistics and macrolinguistics are not yet well established, and they are, in fact, used here purely for convenience. The former refers to a narrower and the latter to a much broader view of the scope of linguistics. According to the microlinguistic view, languages should be analyzed for their own sake and without reference to their social function, to the manner in which they are acquired by children, to the psychological mechanisms that underlie the production and reception of speech, to the literary and the aesthetic or communicative function of language, and so on. In contrast, macrolinguistics embraces all of these aspects of language. A number of areas within macrolinguistics have been given terminological recognition: psycholinguistics, sociolinguistics, anthropological linguistics, dialectology, mathematical and computational linguistics, and stylistics.

A large portion of this article is devoted to theoretical, synchronic microlinguistics, which is generally acknowledged as the central part of the subject; it will be abbreviated henceforth as theoretical linguistics. (J.Lyo.)

The article is divided into the following sections:

History of linguistics 50
- Earlier history 50
  - Non-Western traditions
  - Greek and Roman antiquity
  - The European Middle Ages
  - The Renaissance
- The 19th century 51
  - Development of the comparative method
  - The role of analogy
  - Other 19th-century theories and development
- The 20th century 53
  - Structuralism
  - Transformational grammar
  - Tagmemic, stratificational, and other approaches

Methods of synchronic linguistic analysis 54
- Structural linguistics 54
  - Phonology
  - Morphology
  - Syntax
  - Semantics
- Transformational-generative grammar 59
  - Harris's grammar
  - Chomsky's grammar
  - Modifications in Chomsky's grammar
- Tagmemics 62
  - Modes of language
  - Hierarchy of levels
- Stratificational grammar 62
  - Technical terminology
  - Interstratal relationships
- The Prague school 63
  - Combination of structuralism and functionalism
  - Phonological contributions
  - Theory of markedness
  - Recent contributions

Historical (diachronic) linguistics 64
- Linguistic change 64
  - Sound change
  - Grammatical change
  - Semantic change
  - Borrowing
- The comparative method 65
  - Grimm's law
  - Proto-Indo-European reconstruction
  - Steps in the comparative method
  - Criticisms of the comparative method
  - Internal reconstruction
- Language classification 67

Linguistics and other disciplines 68
- Psycholinguistics 68
  - Language acquisition by children
  - Speech perception
  - Other areas of research
- Sociolinguistics 68
  - Delineation of the field
  - Social dimensions
- Other relationships 69
  - Anthropological linguistics
  - Computational linguistics
  - Mathematical linguistics
  - Stylistics
  - Philosophy of language
  - Applied linguistics

Dialectology and linguistic geography 70
- Dialect geography 70
  - Early dialect studies
  - Dialect atlases
  - The value and applications of dialectology
- Social dialectology 71

Semantics 72
- Modern development of semantics 72
  - Positivist theory
  - Whorfian views
  - School of natural language
  - Modern grammatical influences
- Philosophical views on meaning 73
  - Meaning and reference
  - Meaning and truth
  - Meaning and use
  - Meaning and thought
- Meaning in linguistics 75
  - Semantics in the theory of language
  - Meaning, structure, and context
  - Lexical entries
  - Generative semantics

The study of writing 77
- The scope of the study of writing 77
  - Sources of information
  - Definitions of terms
- Grammatology 78
  - Three main approaches
  - Subdivisions of grammatology
- Epigraphy and paleography 79
- History of the study of writing 79
  - Studies prior to the 18th century
  - Modern Western studies

History of linguistics

EARLIER HISTORY

Non-Western traditions. Linguistic speculation and investigation, insofar as is known, has gone on in only a small number of societies. To the extent that Mesopotamian, Chinese, and Arabic learning dealt with grammar, their treatments were so enmeshed in the particularities of those languages and so little known to the European world until recently that they have had virtually no impact on Western linguistic tradition. Chinese linguistic and philological scholarship stretches back for more than two millennia, but the interest of those scholars was concentrated largely on phonetics, writing, and lexicography; their consideration of grammatical problems was bound up closely with the study of logic.

Certainly the most interesting non-Western grammatical tradition—and the most original and independent—is that of India, which dates back at least two and one-half millennia and which culminates with the grammar of Pāṇini, of the 5th century BC. There are three major ways in which the Sanskrit tradition has had an impact on modern linguistic scholarship. As soon as Sanskrit became known to the Western learned world the unravelling of comparative Indo-European grammar ensued and the foundations were laid for the whole 19th-century edifice of comparative philology and historical linguistics. But, for this, Sanskrit was simply a part of the data; Indian grammatical learning played almost no direct part. Nineteenth-century workers, however, recognized that the native tradition of phonetics in ancient India was vastly superior to Western knowledge; and this had important consequences for the growth of the science of phonetics in the West. Thirdly, there is in the rules or definitions (sutras) of Pāṇini a remarkably subtle and penetrating account of Sanskrit grammar. The construction of sentences, compound nouns, and the like is explained through ordered rules operating on underlying structures in a manner strikingly similar in part to modes of contemporary theory. As might be imagined, this perceptive Indian grammatical work has held great fascination for 20th-century theoretical linguists. A study of Indian logic in relation to Pāṇinian grammar alongside Aristotelian and Western logic in relation to Greek grammar and its successors could bring illuminating insights.

Sanskrit grammar

Whereas in ancient Chinese learning a separate field of study that might be called grammar scarcely took root, in ancient India a sophisticated version of this discipline developed early alongside the other sciences. Even though the study of Sanskrit grammar may originally have had the practical aim of keeping the sacred Vedic texts and their commentaries pure and intact, the study of grammar in India in the 1st millennium BC had already become an intellectual end in itself.

Greek and Roman antiquity. The emergence of grammatical learning in Greece is less clearly known than is sometimes implied, and the subject is more complex than is often supposed; here only the main strands can be sampled. The term *hē grammatikē technē* ("the art of letters") had two senses. It meant the study of the values of the letters and of accentuation and prosody and, in this sense, was an abstract intellectual discipline; and it also meant the skill of literacy and thus embraced applied pedagogy. This side of what was to become "grammatical" learning was distinctly applied, particular, and less exalted by comparison with other pursuits. Most of the developments associated with theoretical grammar grew out of philosophy and criticism; and in these developments a repeated duality of themes crosses and intertwines.

Much of Greek philosophy was occupied with the distinction between that which exists "by nature" and that which exists "by convention." So in language it was natural to account for words and forms as ordained by nature (by onomatopoeia—*i.e.*, by imitation of natural sounds) or as arrived at arbitrarily by a social convention. This dispute regarding the origin of language and meanings paved the way for the development of divergences between the views of the "analogists," who looked on language as possessing an essential regularity as a result of the symmetries that convention can provide, and the views of the "anomalists," who pointed to language's lack of regularity as one facet of the inescapable irregularities of nature. The situation was more complex, however, than this statement would suggest. For example, it seems that the anomalists among the Stoics credited the irrational quality of language precisely to the claim that language did not exactly mirror nature. In any event, the anomalist tradition in the hands of the Stoics brought grammar the benefit of their work in logic and rhetoric. This led to the distinction that, in modern theory, is made with the terms *signifiant* ("what signifies") and *signifié* ("what is signified") or, somewhat differently and more elaborately, with "expression" and "content"; and it laid the groundwork of modern theories of inflection, though by no means with the exhaustiveness and fine-grained analysis reached by the Sanskrit grammarians.

"Analogists" and "anomalists"

The Alexandrians, who were analogists working largely on literary criticism and text philology, completed the development of the classical Greek grammatical tradition. Dionysius Thrax, in the 2nd century BC, produced the first systematic grammar of Western tradition; it dealt only with word morphology. The study of sentence syntax was to wait for Apollonius Dyscolus, of the 2nd century AD. Dionysius called grammar "the acquaintance with [or observation of] what is uttered by poets and writers," using a word meaning a less general form of knowledge than what might be called "science." His typically Alexandrian literary goal is suggested by the headings in his work: pronunciation, poetic figurative language, difficult words, true and inner meanings of words, exposition of form-classes, literary criticism. Dionysius defined a sentence as a unit of sense or thought, but it is difficult to be sure of his precise meaning.

The Romans, who largely took over, with mild adaptations to their highly similar language, the total work of the Greeks, are important not as originators but as transmitters. Aelius Donatus, of the 4th century AD, and Priscian, an African of the 6th century, and their colleagues were slightly more systematic than their Greek models but were essentially retrospective rather than original. Up to this point a field that was at times called *ars grammatica* was a congeries of investigations, both theoretical and practical, drawn from the work and interests of literacy, scribeship, logic, epistemology, rhetoric, textual philosophy, poetics, and literary criticism. Yet modern specialists in the field still share their concerns and interests. The anomalists, who concentrated on surface irregularity and who looked then for regularities deeper down (as the Stoics sought them in logic) bear a resemblance to contemporary scholars of the transformationalist school. And the philological analogists with their regularizing surface segmentation show striking kinship of spirit with the modern school of structural (or taxonomic or glossematic) grammatical theorists.

The European Middle Ages. It is possible that developments in grammar during the Middle Ages constitute one of the most misunderstood areas of the field of linguistics. It is difficult to relate this period coherently to other periods and to modern concerns because surprisingly little is accessible and certain, let alone analyzed with sophistication. In the early 1970s the majority of the known grammatical treatises had not yet been made available in full to modern scholarship, so that not even their true extent could be classified with confidence. These works must be analyzed and studied in the light of medieval learning, especially the learning of the schools of philosophy then current, in order to understand their true value and place.

The field of linguistics has almost completely neglected the achievements of this period. Students of grammar have tended to see as high points in their field the achievements of the Greeks, the Renaissance growth and "rediscovery" of learning (which led directly to modern school traditions), the contemporary flowering of theoretical study (men usually find their own age important and fascinating), and, in recent decades, the astonishing monument of Pāṇini. Many linguists have found uncongenial the combination of medieval Latin learning and premodern philosophy. Yet medieval scholars might reasonably be expected to have bequeathed to modern scholarship the fruits of more than ordinarily refined perceptions of a certain order. These scholars used, wrote in, and studied

Use of Latin by medieval scholars

Latin, a language that, though not their native tongue, was one in which they were very much at home; such scholars in groups must often have represented a highly varied linguistic background.

Some of the medieval treatises continue the tradition of grammars of late antiquity; so there are versions based on Donatus and Priscian, often with less incorporation of the classical poets and writers. Another genre of writing involves simultaneous consideration of grammatical distinctions and scholastic logic; modern linguists are probably inadequately trained to deal with these writings.

Certainly the most obviously interesting theorizing to be found in this period is contained in the "speculative grammar" of the *modistae,* who were so called because the titles of their works were often phrased *De modis significandi tractatus* ("Treatise Concerning the Modes of Signifying"). For the development of the Western grammatical tradition, work of this genre was the second great milestone after the crystallization of Greek thought with the Stoics and Alexandrians. The scholastic philosophers were occupied with relating words and things—*i.e.,* the structure of sentences with the nature of the real world—hence their preoccupation with signification. The aim of the grammarians was to explore how a word (an element of language) matched things apprehended by the mind and how it signified reality. Since a word cannot signify the nature of reality directly, it must stand for the thing signified in one of its modes or properties; it is this discrimination of modes that the study of categories and parts of speech is all about. Thus the study of sentences should lead one to the nature of reality by way of the modes of signifying.

The *modistae*

The *modistae* did not innovate in discriminating categories and parts of speech; they accepted those that had come down from the Greeks through Donatus and Priscian. The great contribution of these grammarians, who flourished between the mid-13th and mid-14th century, was their insistence on a grammar to explicate the distinctions found by their forerunners in the languages known to them. Whether they made the best choice in selecting logic, metaphysics, and epistemology (as they knew them) as the fields to be included with grammar as a basis for the grand account of universal knowledge is less important than the breadth of their conception of the place of grammar. Before the *modistae,* grammar had not been viewed as a separate discipline but had been considered in conjunction with other studies or skills (such as criticism, preservation of valued texts, foreign-language learning). The Greek view of grammar was rather narrow and fragmented; the Roman view was largely technical. The speculative medieval grammarians (who dealt with language as a *speculum,* "mirror" of reality) inquired into the fundamentals underlying language and grammar. They wondered whether grammarians or philosophers discovered grammar, whether grammar was the same for all languages, what the fundamental topic of grammar was, and what the basic and irreducible grammatical primes are. Signification was reached by imposition of words on things; *i.e.,* the sign was arbitrary. Those questions sound remarkably like current issues of linguistics, which serves to illustrate how slow and repetitious progress in the field is. While the *modistae* accepted, by modern standards, a restrictive set of categories, the acumen and sweep they brought to their task resulted in numerous subtle and fresh syntactic observations. A thorough study of the medieval period would greatly enrich the discussion of current questions.

The Renaissance. It is customary to think of the Renaissance as a time of great flowering. There is no doubt that linguistic and philological developments of this period are interesting and significant. Two new sets of data that modern linguists tend to take for granted became available to grammarians during this period: (1) the newly recognized vernacular languages of Europe, for the protection and cultivation of which there subsequently arose national academies and learned institutions that live down to the present day; and (2) the exotic languages of Africa, the Orient, the New World, and, later, of Siberia, Inner Asia, Papua, Oceania, the Arctic, and Australia, which the voyages of discovery opened up. Earlier, the only non-Indo-European grammar at all widely accessible was that of the Hebrews (and to some extent Arabic); and Semitic in fact shares many categories with Indo-European in its grammar. Indeed, for many of the exotic languages scholarship barely passed beyond the most rudimentary initial collection of word lists; grammatical analysis was scarcely approached.

New sources of data

In the field of grammar, the Renaissance did not produce notable innovation or advance. Generally speaking, there was a strong rejection of speculative grammar and a relatively uncritical resumption of late Roman views (as stated by Priscian). This was somewhat understandable in the case of Latin or Greek grammars, since here the task was less evidently that of intellectual inquiry and more that of the schools, with the practical aim of gaining access to the newly discovered ancients. But, aside from the fact that, beginning in the 15th century, serious grammars of European vernaculars were actually written, it is only in particular cases and for specific details (*e.g.,* a mild alteration in the number of parts of speech or cases of nouns) that real departures from Roman grammar can be noted. Likewise, until the end of the 19th century, grammars of the exotic languages, written largely by missionaries and traders, were cast almost entirely in the Roman model, to which the Renaissance had added a limited medieval syntactic ingredient.

From time to time a degree of boldness may be seen in France: Petrus Ramus, a 16th-century logician, worked within a taxonomic framework of the surface shapes of words and inflections, such work entailing some of the attendant trivialities that modern linguistics has experienced (*e.g.,* by dividing up Latin nouns on the basis of equivalence of syllable count among their case forms). In the 17th century, members of Solitaires (a group of hermits who lived in the deserted abbey of Port-Royal in France) produced a grammar that has exerted noteworthy continuing influence, even in contemporary theoretical discussion. Drawing their basic view from scholastic logic as modified by rationalism, these people aimed to produce a philosophical grammar that would capture what was common to the grammars of languages—a general grammar, but not aprioristically universalist. This grammar has attracted recent attention because it employs certain syntactic formulations that resemble in detail contemporary transformational rules, which formulate the relationship between the various elements of a sentence.

The Port-Royal school

Roughly from the 15th century to World War II, however, the version of grammar available to the Western public (together with its colonial expansion) remained basically that of Priscian with only occasional and subsidiary modifications, and the knowledge of new languages brought only minor adjustments to the serious study of grammar. As education has become more broadly disseminated throughout society by the schools, attention has shifted from theoretical or technical grammar as an intellectual preoccupation to prescriptive grammar suited to pedagogical purposes, which started with Renaissance vernacular nationalism. Grammar increasingly parted company with its older fellow disciplines within philosophy as they moved over to the domain known as natural science, and technical academic grammatical study has increasingly become involved with issues represented by empiricism versus rationalism and their successor manifestations on the academic scene.

Nearly down to the present day, the grammar of the schools has had only tangential connections with the studies pursued by professional linguists; for most people prescriptive grammar has become synonymous with "grammar," and the prevailing view held by educated people regards grammar as an item of folk knowledge open to speculation by all, and in nowise a formal science requiring adequate preparation such as is assumed for chemistry. (E.P.H./Ed.)

THE 19TH CENTURY

Development of the comparative method. It is generally agreed that the most outstanding achievement of linguistic scholarship in the 19th century was the development of

the comparative method, which comprised a set of principles whereby languages could be systematically compared with respect to their sound systems, grammatical structure, and vocabulary and shown to be "genealogically" related. As French, Italian, Portuguese, Romanian, Spanish, and the other Romance languages had evolved from Latin, so Latin, Greek, and Sanskrit as well as the Celtic, Germanic, and Slavic languages and many other languages of Europe and Asia had evolved from some earlier language, to which the name Indo-European or Proto-Indo-European is now customarily applied. That all the Romance languages were descended from Latin and thus constituted one "family" had been known for centuries; but the existence of the Indo-European family of languages and the nature of their genealogical relationship was first demonstrated by the 19th-century comparative philologists. (The term philology in this context is not restricted to the study of literary languages.)

Discovery of the Indo-European language family

The main impetus for the development of comparative philology came toward the end of the 18th century, when it was discovered that Sanskrit bore a number of striking resemblances to Greek and Latin. An English orientalist, Sir William Jones, though he was not the first to observe these resemblances, is generally given the credit for bringing them to the attention of the scholarly world and putting forward the hypothesis, in 1786, that all three languages must have "sprung from some common source, which perhaps no longer exists." By this time, a number of texts and glossaries of the older Germanic languages (Gothic, Old High German, and Old Norse) had been published, and Jones realized that Germanic as well as Old Persian and perhaps Celtic had evolved from the same "common source." The next important step came in 1822, when the German scholar Jacob Grimm, following the Danish linguist Rasmus Rask (whose work, being written in Danish, was less accessible to most European scholars), pointed out in the second edition of his comparative grammar of Germanic that there were a number of systematic correspondences between the sounds of Germanic and the sounds of Greek, Latin, and Sanskrit in related words. Grimm noted, for example, that where Gothic (the oldest surviving Germanic language) had an *f,* Latin, Greek, and Sanskrit frequently had a *p* (*e.g.,* Gothic *fotus,* Latin *pedis,* Greek *podós,* Sanskrit *padás,* all meaning "foot"); when Gothic had a *p,* the non-Germanic languages had a *b;* when Gothic had a *b,* the non-Germanic languages had what Grimm called an "aspirate" (Latin *f,* Greek *ph,* Sanskrit *bh*). In order to account for these correspondences he postulated a cyclical "soundshift" (*Lautverschiebung*) in the prehistory of Germanic, in which the original "aspirates" became voiced unaspirated stops (*bh* became *b,* etc.), the original voiced unaspirated stops became voiceless (*b* became *p,* etc.), and the original voiceless (unaspirated) stops became "aspirates" (*p* became *f*). Grimm's term, "aspirate," it will be noted, covered such phonetically distinct categories as aspirated stops (*bh, ph*), produced with an accompanying audible puff of breath, and fricatives (*f*), produced with audible friction as a result of incomplete closure in the vocal tract.

The Neogrammarians and sound laws

In the work of the next 50 years the idea of sound change was made more precise, and, in the 1870s, a group of scholars known collectively as the *Junggrammatiker* ("young grammarians," or Neogrammarians) put forward the thesis that all changes in the sound system of a language as it developed through time were subject to the operation of regular sound laws. Though the thesis that sound laws were absolutely regular in their operation (unless they were inhibited in particular instances by the influence of analogy) was at first regarded as most controversial, by the end of the 19th century it was quite generally accepted and had become the cornerstone of the comparative method. Using the principle of regular sound change, scholars were able to reconstruct "ancestral" common forms from which the later forms found in particular languages could be derived. By convention, such reconstructed forms are marked in the literature with an asterisk. Thus, from the reconstructed Proto-Indo-European word for "ten," **dekm,* it was possible to derive Sanskrit *daśa,* Greek *déka,* Latin *decem,* and Gothic *taihun* by postulating a number of different sound laws that operated independently in the different branches of the Indo-European family. The question of sound change is dealt with in greater detail in the section entitled *Historical linguistics.*

The role of analogy. Analogy has been mentioned in connection with its inhibition of the regular operation of sound laws in particular word forms. This was how the Neogrammarians thought of it. In the course of the 20th century, however, it has come to be recognized that analogy, taken in its most general sense, plays a far more important role in the development of languages than simply that of sporadically preventing what would otherwise be a completely regular transformation of the sound system of a language. When a child learns to speak he tends to regularize the anomalous, or irregular, forms by analogy with the more regular and productive patterns of formation in the language; *e.g.,* he will tend to say "comed" rather than "came," "dived" rather than "dove," and so on, just as he will say "talked," "loved," and so forth. The fact that the child does this is evidence that he has learned or is learning the regularities or rules of his language. He will go on to "unlearn" some of the analogical forms and substitute for them the anomalous forms current in the speech of the previous generation. But in some cases, he will keep a "new" analogical form (*e.g.,* "dived" rather than "dove"), and this may then become the recognized and accepted form.

Other 19th-century theories and development. *Inner and outer form.* One of the most original, if not one of the most immediately influential, linguists of the 19th century was the learned Prussian statesman, Wilhelm von Humboldt (died 1835). His interests, unlike those of most of his contemporaries, were not exclusively historical. Following the German philosopher Johann Gottfried von Herder (1744–1803), he stressed the connection between national languages and national character: this was but a commonplace of romanticism. More original was Humboldt's theory of "inner" and "outer" form in language. The outer form of language was the raw material (the sounds) from which different languages were fashioned; the inner form was the pattern, or structure, of grammar and meaning that was imposed upon this raw material and differentiated one language from another. This "structural" conception of language was to become dominant, for a time at least, in many of the major centres of linguistics by the middle of the 20th century. Another of Humboldt's ideas was that language was something dynamic, rather than static, and was an activity itself rather than the product of activity. A language was not a set of actual utterances produced by speakers but the underlying principles or rules that made it possible for speakers to produce such utterances and, moreover, an unlimited number of them. This idea was taken up by a German philologist, Heymann Steinthal, and, what is more important, by the physiologist and psychologist Wilhelm Wundt, and thus influenced late 19th- and early 20th-century theories of the psychology of language. Its influence, like that of the distinction of inner and outer form, can also be seen in the thought of Ferdinand de Saussure, a Swiss linguist. But its full implications were probably not perceived and made precise until the middle of the 20th century, when the U.S. linguist Noam Chomsky re-emphasized it and made it one of the basic notions of generative grammar (see below).

Ideas of Wilhelm von Humboldt

Phonetics and dialectology. Many other interesting and important developments occurred in 19th-century linguistic research, among them work in the areas of phonetics and dialectology. Research in both these fields was promoted by the Neogrammarians' concern with sound change and by their insistence that prehistoric developments in languages were of the same kind as developments taking place in the languages and dialects currently spoken. The development of phonetics in the West was also strongly influenced at this period, as were many of the details of the more philological analysis of the Indo-European languages, by the discovery of the works of the Indian grammarians who, from the time of the Sanskrit grammarian Pāṇini (5th or 6th century BC), if not before, had arrived at a much more comprehensive and scientific

Development of phonetics

theory of phonetics, phonology, and morphology than anything achieved in the West until the modern period.

THE 20TH CENTURY

Structuralism. The term structuralism has been used as a slogan and rallying cry by a number of different schools of linguistics, and it is necessary to realize that it has somewhat different implications according to the context in which it is employed. It is convenient to draw first a broad distinction between European and American structuralism and, then, to treat them separately.

Structural linguistics in Europe. Structural linguistics in Europe is generally said to have begun in 1916 with the posthumous publication of the *Cours de Linguistique Générale (Course in General Linguistics)* of Ferdinand de Saussure. Much of what is now considered as Saussurean can be seen, though less clearly, in the earlier work of Humboldt, and the general structural principles that Saussure was to develop with respect to synchronic linguistics in the *Cours* had been applied almost 40 years before (1879) by Saussure himself in a reconstruction of the Indo-European vowel system. The full significance of the work was not appreciated at the time. Saussure's structuralism can be summed up in two dichotomies (which jointly cover what Humboldt referred to in terms of his own distinction of inner and outer form): (1) *langue* versus *parole* and (2) form versus substance. By *langue,* best translated in its technical Saussurean sense as language system, is meant the totality of regularities and patterns of formation that underlie the utterances of a language; by *parole,* which can be translated as language behaviour, is meant the actual utterances themselves. Just as two performances of a piece of music given by different orchestras on different occasions will differ in a variety of details and yet be identifiable as performances of the same piece, so two utterances may differ in various ways and yet be recognized as instances, in some sense, of the same utterance. What the two musical performances and the two utterances have in common is an identity of form, and this form, or structure, or pattern, is in principle independent of the substance, or "raw material," upon which it is imposed.

European sense of structuralism

"Structuralism," in the European sense then, refers to the view that there is an abstract relational structure that underlies and is to be distinguished from actual utterances—a system underlying actual behaviour—and that this is the primary object of study for the linguist.

Two important points arise here: first, that the structural approach is not in principle restricted to synchronic linguistics; second, that the study of meaning, as well as the study of phonology and grammar, can be structural in orientation. In both cases "structuralism" is opposed to "atomism" in the European literature. It was Saussure who drew the terminological distinction between synchronic and diachronic linguistics in the *Cours;* despite the undoubtedly structural orientation of his own early work in the historical and comparative field, he maintained that, whereas synchronic linguistics should deal with the structure of a language system at a given point in time, diachronic linguistics should be concerned with the historical development of isolated elements—it should be atomistic. Whatever the reasons that led Saussure to take this rather paradoxical view, his teaching on this point was not generally accepted, and scholars soon began to apply structural concepts to the diachronic study of languages. The most important of the various schools of structural linguistics to be found in Europe in the first half of the 20th century have included the Prague school, most notably represented by Nikolay Sergeyevich Trubetskoy (died 1938) and Roman Jakobson (born 1896), both Russian émigrés, and the Copenhagen (or glossematic) school, centred around Louis Hjelmslev (died 1965). John Rupert Firth (died 1960) and his followers, sometimes referred to as the London school, were less Saussurean in their approach, but, in a general sense of the term, their approach may also be described appropriately as structural linguistics.

Structural linguistics in America. American and European structuralism shared a number of features. In insisting upon the necessity of treating each language as a more or less coherent and integrated system, both European and American linguists of this period tended to emphasize, if not to exaggerate, the structural uniqueness of individual languages. There was especially good reason to take this point of view given the conditions in which American linguistics developed from the end of the 19th century. There were hundreds of indigenous American Indian languages that had never been previously described. Many of these were spoken by only a handful of speakers and, if they were not recorded before they became extinct, would be permanently inaccessible. Under these circumstances, such linguists as Franz Boas (died 1942) were less concerned with the construction of a general theory of the structure of human language than they were with prescribing sound methodological principles for the analysis of unfamiliar languages. They were also fearful that the description of these languages would be distorted by analyzing them in terms of categories derived from the analysis of the more familiar Indo-European languages.

The work of Sapir, Bloomfield, and Whorf

After Boas, the two most influential American linguists were Edward Sapir (died 1939) and Leonard Bloomfield (died 1949). Like his teacher Boas, Sapir was equally at home in anthropology and linguistics, the alliance of which disciplines has endured to the present day in many American universities. Boas and Sapir were both attracted by the Humboldtian view of the relationship between language and thought, but it was left to one of Sapir's pupils, Benjamin Lee Whorf, to present it in a sufficiently challenging form to attract widespread scholarly attention. Since the republication of Whorf's more important papers in 1956, the thesis that language determines perception and thought has come to be known as the Whorfian hypothesis.

Sapir's work has always held an attraction for the more anthropologically inclined American linguists. But it was Bloomfield who prepared the way for the later phase of what is now thought of as the most distinctive manifestation of American "structuralism." When he published his first book in 1914, Bloomfield was strongly influenced by Wundt's psychology of language. In 1933, however, he published a drastically revised and expanded version with the new title *Language;* this book dominated the field for the next 30 years. In it Bloomfield explicitly adopted a behaviouristic approach to the study of language, eschewing in the name of scientific objectivity all reference to mental or conceptual categories. Of particular consequence was his adoption of the behaviouristic theory of semantics according to which meaning is simply the relationship between a stimulus and a verbal response. Because science was still a long way from being able to give a comprehensive account of most stimuli, no significant or interesting results could be expected from the study of meaning for some considerable time, and it was preferable, as far as possible, to avoid basing the grammatical analysis of a language on semantic considerations. Bloomfield's followers pushed even further the attempt to develop methods of linguistic analysis that were not based on meaning. One of the most characteristic features of "post-Bloomfieldian" American structuralism, then, was its almost complete neglect of semantics.

Another characteristic feature, one that was to be much criticized by Chomsky, was its attempt to formulate a set of "discovery procedures"—procedures that could be applied more or less mechanically to texts and could be guaranteed to yield an appropriate phonological and grammatical description of the language of the texts. Structuralism, in this narrower sense of the term, is represented, with differences of emphasis or detail, in the major American textbooks published during the 1950s.

Transformational grammar. The most significant development in linguistic theory and research in recent years was the rise of generative grammar, and, more especially, of transformational-generative grammar, or transformational grammar, as it came to be known. Two versions of transformational grammar were put forward in the mid-1950s, the first by Zellig S. Harris and the second by Noam Chomsky, his pupil. It is Chomsky's system that has attracted the most attention so far. As first presented by Chomsky in *Syntactic Structures* (1957), transformational grammar can be seen partly as a reaction against

The work of Harris and Chomsky

post-Bloomfieldian structuralism and partly as a continuation of it. What Chomsky reacted against most strongly was the post-Bloomfieldian concern with discovery procedures. In his opinion, linguistics should set itself the more modest and more realistic goal of formulating criteria for evaluating alternative descriptions of a language without regard to the question of how these descriptions had been arrived at. The statements made by linguists in describing a language should, however, be cast within the framework of a far more precise theory of grammar than had hitherto been the case, and this theory should be formalized in terms of modern mathematical notions. Within a few years, Chomsky had broken with the post-Bloomfieldians on a number of other points also. He had adopted what he called a "mentalistic" theory of language, by which term he implied that the linguist should be concerned with the speaker's creative linguistic competence and not his performance, the actual utterances produced. He had challenged the post-Bloomfieldian concept of the phoneme (see below), which many scholars regarded as the most solid and enduring result of the previous generation's work. And he had challenged the structuralists' insistence upon the uniqueness of every language, claiming instead that all languages were, to a considerable degree, cut to the same pattern—they shared a certain number of formal and substantive universals.

Tagmemic, stratificational, and other approaches. The effect of Chomsky's ideas has been phenomenal. It is hardly an exaggeration to say that there is no major theoretical issue in linguistics today that is debated in terms other than those in which he has chosen to define it, and every school of linguistics tends to define its position in relation to his. Among the rival schools are tagmemics, stratificational grammar, and the Prague school. Tagmemics is the system of linguistic analysis developed by the U.S. linguist Kenneth L. Pike and his associates in connection with their work as Bible translators. Its foundations were laid during the 1950s, when Pike differed from the post-Bloomfieldian structuralists on a number of principles, and it has been further elaborated since then. Tagmemic analysis has been used for analyzing a great many previously unrecorded languages, especially in Central and South America and in West Africa. Stratificational grammar, developed by a U.S. linguist, Sydney M. Lamb, has been seen by some linguists as an alternative to transformational grammar. Not yet fully expounded or widely exemplified in the analysis of different languages, stratificational grammar is perhaps best characterized as a radical modification of post-Bloomfieldian linguistics, but it has many features that link it with European structuralism. The Prague school has been mentioned above for its importance in the period immediately following the publication of Saussure's *Cours.* Many of its characteristic ideas (in particular, the notion of distinctive features in phonology) have been taken up by other schools. But there has been further development in Prague of the functional approach to syntax (see below). The work of M.A.K. Halliday in England derived much of its original inspiration from Firth (above), but Halliday provided a more systematic and comprehensive theory of the structure of language than Firth had, and it has been quite extensively illustrated.

Prague school

Methods of synchronic linguistic analysis

STRUCTURAL LINGUISTICS

This section is concerned mainly with a version of structuralism (which may also be called descriptive linguistics) developed by scholars working in a post-Bloomfieldian tradition.

Phonology. With the great progress made in phonetics in the late 19th century, it had become clear that the question whether two speech sounds were the same or not was more complex than might appear at first sight. Two utterances of what was taken to be the same word might differ quite perceptibly from one occasion of utterance to the next. Some of this variation could be attributed to a difference of dialect or accent and is of no concern here. But even two utterances of the same word by the same speaker might vary from one occasion to the next. Variation of this kind, though it is generally less obvious and would normally pass unnoticed, is often clear enough to the trained phonetician and is measurable instrumentally. It is known that the "same" word is being uttered, even if the physical signal produced is variable, in part, because the different pronunciations of the same word will cluster around some acoustically identifiable norm. But this is not the whole answer, because it is actually impossible to determine norms of pronunciation in purely acoustic terms. Once it has been decided what counts as "sameness" of sound from the linguistic point of view, the permissible range of variation for particular sounds in particular contexts can be measured, and, within certain limits, the acoustic cues for the identification of utterances as "the same" can be determined.

What is at issue is the difference between phonetic and phonological (or phonemic) identity, and for these purposes it will be sufficient to define phonetic identity in terms solely of acoustic "sameness." Absolute phonetic identity is a theoretical ideal never fully realized. From a purely phonetic point of view, sounds are more or less similar, rather than absolutely the same or absolutely different. Speech sounds considered as units of phonetic analysis in this article are called phones, and, following the normal convention, are represented by enclosing the appropriate alphabetic symbol in square brackets. Thus [p] will refer to a *p* sound (*i.e.,* what is described more technically as a voiceless, bilabial stop); and [pit] will refer to a complex of three phones—a *p* sound, followed by an *i* sound, followed by a *t* sound. A phonetic transcription may be relatively broad (omitting much of the acoustic detail) or relatively narrow (putting in rather more of the detail), according to the purpose for which it is intended. A very broad transcription will be used in this article except when finer phonetic differences must be shown.

Phones: speech sounds

Phonological, or phonemic, identity was referred to above as "sameness of sound from the linguistic point of view." Considered as phonological units—*i.e.,* from the point of view of their function in the language—sounds are described as phonemes and are distinguished from phones by enclosing their appropriate symbol (normally, but not necessarily, an alphabetic one) between two slash marks. Thus /p/ refers to a phoneme that may be realized on different occasions of utterance or in different contexts by a variety of more or less different phones. Phonological identity, unlike phonetic similarity, is absolute: two phonemes are either the same or different, they cannot be more or less similar. For example, the English words "bit" and "pit" differ phonemically in that the first has the phoneme /b/ and the second has the phoneme /p/ in initial position. As the words are normally pronounced, the phonetic realization of /b/ will differ from the phonetic realization of /p/ in a number of different ways: it will be at least partially voiced (*i.e.,* there will be some vibration of the vocal cords), it will be without aspiration (*i.e.,* there will be no accompanying slight puff of air, as there will be in the case of the phone realizing /p/), and it will be pronounced with less muscular tension. It is possible to vary any one or all of these contributory differences, making the phones in question more or less similar, and it is possible to reduce the phonetic differences to the point that the hearer cannot be certain which word, "bit" or "pit," has been uttered. But it must be either one or the other; there is no word with an initial sound formed in the same manner as /p/ or /b/ that is halfway between the two. This is what is meant by saying that phonemes are absolutely distinct from one another—they are discrete rather than continuously variable.

How it is known whether two phones realize the same phoneme or not is dealt with differently by different schools of linguists. The "orthodox" post-Bloomfieldian school regards the first criterion to be phonetic similarity. Two phones are not said to realize the same phoneme unless they are sufficiently similar. What is meant by "sufficiently similar" is rather vague, but it must be granted that for every phoneme there is a permissible range of variation in the phones that realize it. As far as occurrence in the same context goes, there are no serious problems.

Classifying phones or phonemes

More critical is the question of whether two phones occurring in different contexts can be said to realize the same phoneme or not. To take a standard example from English: the phone that occurs at the beginning of the word "pit" differs from the phone that occurs after the initial /s/ of "spit." The "*p* sound" occurring after the /s/ is unaspirated (*i.e.,* it is pronounced without any accompanying slight puff of air). The aspirated and unaspirated "*p* sounds" may be symbolized rather more narrowly as [pʰ] and [p] respectively. The question then is whether [pʰ] and [p] realize the same phoneme /p/ or whether each realizes a different phoneme. They satisfy the criterion of phonetic similarity, but this, though a necessary condition of phonemic identity, is not a sufficient one.

The next question is whether there is any pair of words in which the two phones are in minimal contrast (or opposition); that is, whether there is any context in English in which the occurrence of the one rather than the other has the effect of distinguishing two or more words (in the way that [pʰ] versus [b] distinguishes the so-called minimal pairs "pit" and "bit," "pan" and "ban," and so on). If there is, it can be said that, despite their phonetic similarity, the two phones realize (or "belong to") different phonemes—that the difference between them is phonemic. If there is no context in which the two phones are in contrast (or opposition) in this sense, it can be said that they are variants of the same phoneme—that the difference between them is nonphonemic. Thus, the difference between [pʰ] and [p] in English is nonphonemic; the two sounds realize, or belong to, the same phoneme, namely /p/. In several other languages—*e.g.,* Hindi—the contrast between such sounds as [pʰ] and [p] is phonemic, however. The question is rather more complicated than it has been represented here. In particular, it should be noted that [p] is phonetically similar to [b] as well as to [pʰ] and that, although [pʰ] and [b] are in contrast, [p] and [b] are not. It would thus be possible to regard [p] and [b] as variants of the same phoneme. Most linguists, however, have taken the alternative view, assigning [p] to the same phoneme as [pʰ]. Here it will suffice to note that the criteria of phonetic similarity and lack of contrast do not always uniquely determine the assignment of phones to phonemes. Various supplementary criteria may then be invoked.

Free variation and complementary distribution

Phones that can occur and do not contrast in the same context are said to be in free variation in that context, and, as has been shown, there is a permissible range of variation for the phonetic realization of all phonemes. More important than free variation in the same context, however, is systematically determined variation according to the context in which a given phoneme occurs. To return to the example used above: [p] and [pʰ], though they do not contrast, are not in free variation either. Each of them has its own characteristic positions of occurrence, and neither occurs, in normal English pronunciation, in any context characteristic for the other (*e.g.,* only [pʰ] occurs at the beginning of a word, and only [p] occurs after s). This is expressed by saying that they are in complementary distribution. (The distribution of an element is the whole range of contexts in which it can occur.) Granted that [p] and [pʰ] are variants of the same phoneme /p/, it can be said that they are contextually, or positionally, determined variants of it. To use the technical term, they are allophones of /p/. The allophones of a phoneme, then, are its contextually determined variants and they are in complementary distribution.

The post-Bloomfieldians made the assignment of phones to phonemes subject to what is now generally referred to as the principle of bi-uniqueness. The phonemic specification of a word or utterance was held to determine uniquely its phonetic realization (except for free variation), and, conversely, the phonetic description of a word or utterance was held to determine uniquely its phonemic analysis. Thus, if two words or utterances are pronounced alike, then they must receive the same phonemic description; conversely, two words or utterances that have been given the same phonemic analysis must be pronounced alike. The principle of bi-uniqueness was also held to imply that, if a given phone was assigned to a particular phoneme in one position of occurrence, then it must be assigned to the same phoneme in all its other positions of occurrence; it could not be the allophone of one phoneme in one context and of another phoneme in other contexts.

A second important principle of the post-Bloomfieldian approach was its insistence that phonemic analysis should be carried out prior to and independently of grammatical analysis. Neither this principle nor that of bi-uniqueness was at all widely accepted outside the post-Bloomfieldian school, and they have been abandoned by the generative phonologists (see below).

Phonemes of the kind referred to so far are segmental; they are realized by consonantal or vocalic (vowel) segments of words, and they can be said to occur in a certain order relative to one another. For example, in the phonemic representation of the word "bit," the phoneme /b/ precedes /i/, which precedes /t/. But nonsegmental, or suprasegmental, aspects of the phonemic realization of words and utterances may also be functional in a language. In English, for example, the noun "import" differs from the verb "import" in that the former is accented on the first and the latter on the second syllable. This is called a stress accent: the accented syllable is pronounced with greater force or intensity. Many other languages distinguish words suprasegmentally by tone. For example, in Mandarin Chinese the words *haò* "day" and *haǒ* "good" are distinguished from one another in that the first has a falling tone and the second a falling-rising tone; these are realized, respectively, as (1) a fall in the pitch of the syllable from high to low and (2) a change in the pitch of the syllable from medium to low and back to medium. Stress and tone are suprasegmental in the sense that they are "superimposed" upon the sequence of segmental phonemes. The term tone is conventionally restricted by linguists to phonologically relevant variations of pitch at the level of words. Intonation, which is found in all languages, is the variation in the pitch contour or pitch pattern of whole utterances, of the kind that distinguishes (either of itself or in combination with some other difference) statements from questions or indicates the mood or attitude of the speaker (as hesitant, surprised, angry, and so forth). Stress, tone, and intonation do not exhaust the phonologically relevant suprasegmental features found in various languages, but they are among the most important.

Suprasegmental phonemes

A complete phonological description of a language includes all the segmental phonemes and specifies which allophones occur in which contexts. It also indicates which sequences of phonemes are possible in the language and which are not: it will indicate, for example, that the sequences /bl/ and /br/ are possible at the beginning of English words but not /bn/ or /bm/. A phonological description also identifies and states the distribution of the suprasegmental features. Just how this is to be done, however, has been rather more controversial in the post-Bloomfieldian tradition. Differences between the post-Bloomfieldian approach to phonology and approaches characteristic of other schools of structural linguistics will be treated below.

Morphology. The grammatical description of many, if not all, languages is conveniently divided into two complementary sections: morphology and syntax. The relationship between them, as generally stated, is as follows: morphology accounts for the internal structure of words, and syntax describes how words are combined to form phrases, clauses, and sentences.

Distinction between morphology and syntax

There are many words in English that are fairly obviously analyzable into smaller grammatical units. For example, the word "unacceptability" can be divided into *un-, accept, abil-,* and *-ity (abil-* being a variant of *-able).* Of these, at least three are minimal grammatical units, in the sense that they cannot be analyzed into yet smaller grammatical units—*un-, abil-,* and *ity.* The status of *accept,* from this point of view, is somewhat uncertain. Given the existence of such forms as *accede* and *accuse,* on the one hand, and of *except, excede,* and *excuse,* on the other, one might be inclined to analyze *accept* into *ac-* (which might subsequently be recognized as a variant of *ad-*) and *-cept.* The question is left open. Minimal grammatical units like *un-, abil-,* and *-ity* are what Bloomfield called mor-

phemes; he defined them in terms of the "partial phonetic-semantic resemblance" holding within sets of words. For example, "unacceptable," "untrue," and "ungracious" are phonetically (or, phonologically) similar as far as the first syllable is concerned and are similar in meaning in that each of them is negative by contrast with a corresponding positive adjective ("acceptable," "true," "gracious"). This "partial phonetic-semantic resemblance" is accounted for by noting that the words in question contain the same morpheme (namely, *un-*) and that this morpheme has a certain phonological form and a certain meaning.

Bloomfield's definition of the morpheme in terms of "partial phonetic-semantic resemblance" was considerably modified and, eventually, abandoned entirely by some of his followers. Whereas Bloomfield took the morpheme to be an actual segment of a word, others defined it as being a purely abstract unit, and the term morph was introduced to refer to the actual word segments. The distinction between morpheme and morph (which is, in certain respects, parallel to the distinction between phoneme and phone) may be explained by means of an example. If a morpheme in English is posited with the function of accounting for the grammatical difference between singular and plural nouns, it may be symbolized by enclosing the term plural within brace brackets. Now the morpheme [plural] is represented in a number of different ways. Most plural nouns in English differ from the corresponding singular forms in that they have an additional final segment. In the written forms of these words, it is either *-s* or *-es* (*e.g.*, "cat" : "cats"; "dog" : "dogs"; "fish" : "fishes"). The word segments written *-s* or *-es* are morphs. So also is the word segment written *-en* in "oxen." All these morphs represent the same morpheme. But there are other plural nouns in English that differ from the corresponding singular forms in other ways (*e.g.*, "mouse" : "mice"; "criterion" : "criteria"; and so on) or not at all (*e.g.*, "this sheep" : "these sheep"). Within the post-Bloomfieldian framework no very satisfactory account of the formation of these nouns could be given. But it was clear that they contained (in some sense) the same morpheme as the more regular plurals.

Plural morphs in English

Morphs that are in complementary distribution and represent the same morpheme are said to be allomorphs of that morpheme. For example, the regular plurals of English nouns are formed by adding one of three morphs on to the form of the singular: /s/, /z/, or /iz/ (in the corresponding written forms both /s/ and /z/ are written *-s* and /iz/ is written *-es*). Their distribution is determined by the following principle: if the morph to which they are to be added ends in a "sibilant" sound (*e.g., s, z, sh, ch*), then the syllabic allomorph /iz/ is selected (*e.g., fish-es* /fiš-iz/, *match-es* /mač-iz/); otherwise the nonsyllabic allomorphs are selected, the voiceless allomorph /s/ with morphs ending in a voiceless consonant (*e.g., cat-s* /kat-s/) and the voiced allomorph /z/ with morphs ending in a vowel or voiced consonant (*e.g., flea-s* /fli-z/, *dog-s* /dog-z/). These three allomorphs, it will be evident, are in complementary distribution, and the alternation between them is determined by the phonological structure of the preceding morph. Thus the choice is phonologically conditioned.

Past participle morphs in English

Very similar is the alternation between the three principal allomorphs of the past participle ending, /id/, /t/, and /d/, all of which correspond to the *-ed* of the written forms. If the preceding morph ends with /t/ or /d/, then the syllabic allomorph /id/ is selected (*e.g., wait-ed* /weit-id/). Otherwise, if the preceding morph ends with a voiceless consonant, one of the nonsyllabic allomorphs is selected—the voiceless allomorph /t/ when the preceding morph ends with a voiceless consonant (*e.g., pack-ed* /pak-t/) and the voiced allomorph /d/ when the preceding morph ends with a vowel or voiced consonant (*e.g., row-ed* /rou-d/; *tame-d* /teim-d/). This is another instance of phonological conditioning. Phonological conditioning may be contrasted with the principle that determines the selection of yet another allomorph of the past participle morpheme. The final /n/ of *show-n* or *see-n* (which marks them as past participles) is not determined by the phonological structure of the morphs *show* and *see*. For each English word that is similar to "show" and "see" in this respect, it must be stated as a synchronically inexplicable fact that it selects the /n/ allomorph. This is called grammatical conditioning. There are various kinds of grammatical conditioning.

Alternation of the kind illustrated above for the allomorphs of the plural morpheme and the /id/, /d/, and /t/ allomorphs of the past participle is frequently referred to as morphophonemic. Some linguists have suggested that it should be accounted for not by setting up three allomorphs each with a distinct phonemic form but by setting up a single morph in an intermediate morphophonemic representation. Thus, the regular plural morph might be said to be composed of the morphophoneme /Z/ and the most common past-participle morph of the morphophoneme /D/. General rules of morphophonemic interpretation would then convert /Z/ and /D/ to their appropriate phonetic form according to context. This treatment of the question foreshadows, on the one hand, the stratificational treatment and, on the other, the generative approach, though they differ considerably in other respects.

An important concept in grammar and, more particularly, in morphology is that of free and bound forms. A bound form is one that cannot occur alone as a complete utterance (in some normal context of use). For example, *-ing* is bound in this sense, whereas *wait* is not, nor is *waiting*. Any form that is not bound is free. Bloomfield based his definition of the word on this distinction between bound and free forms. Any free form consisting entirely of two or more smaller free forms was said to be a phrase (*e.g.*, "poor John" or "ran away"), and phrases were to be handled within syntax. Any free form that was not a phrase was defined to be a word and to fall within the scope of morphology. One of the consequences of Bloomfield's definition of the word was that morphology became the study of constructions involving bound forms. The so-called isolating languages, which make no use of bound forms (*e.g.*, Vietnamese), would have no morphology.

The principal division within morphology is between inflection and derivation (or word formation). Roughly speaking, inflectional constructions can be defined as yielding sets of forms that are all grammatically distinct forms of single vocabulary items, whereas derivational constructions yield distinct vocabulary items. For example, "sings," "singing," "sang," and "sung" are all inflectional forms of the vocabulary item traditionally referred to as "the verb to sing"; but "singer," which is formed from "sing" by the addition of the morph *-er* (just as "singing" is formed by the addition of *-ing*), is one of the forms of a different vocabulary item. When this rough distinction between derivation and inflection is made more precise, problems occur. The principal consideration, undoubtedly, is that inflection is more closely integrated with and determined by syntax. But the various formal criteria that have been proposed to give effect to this general principle are not uncommonly in conflict in particular instances, and it probably must be admitted that the distinction between derivation and inflection, though clear enough in most cases, is in the last resort somewhat arbitrary.

Inflection and derivation

Bloomfield and most linguists have discussed morphological constructions in terms of processes. Of these, the most widespread throughout the languages of the world is affixation; *i.e.*, the attachment of an affix to a base. For example, the word "singing" can be described as resulting from the affixation of *-ing* to the base *sing*. (If the affix is put in front of the base, it is a prefix; if it is put after the base, it is a suffix; and if it is inserted within the base, splitting it into two discontinuous parts, it is an infix.) Other morphological processes recognized by linguists need not be mentioned here, but reference may be made to the fact that many of Bloomfield's followers from the mid-1940s were dissatisfied with the whole notion of morphological processes. Instead of saying that *-ing* was affixed to *sing* they preferred to say that *sing* and *-ing* co-occurred in a particular pattern or arrangement, thereby avoiding the implication that *sing* is in some sense prior to or more basic than *-ing*. The distinction of morpheme and morph (and the notion of allomorphs) was developed in order to make possible the description of the morphology and

syntax of a language in terms of "arrangements" of items rather than in terms of "processes" operating upon more basic items. Nowadays, the opposition to "processes" is, except among the stratificationalists, almost extinct. It has proved to be cumbersome if not impossible to describe the relationship between certain linguistic forms without deriving one from the other or both from some common underlying form, and most linguists no longer feel that this is in any way reprehensible.

Bloomfield's theory of syntax

Syntax. Syntax, for Bloomfield, was the study of free forms that were composed entirely of free forms. Central to his theory of syntax were the notions of form classes and constituent structure. (These notions were also relevant though less central in the theory of morphology.) Bloomfield defined form classes, rather imprecisely, in terms of some common "recognizable phonetic or grammatical feature" shared by all the members. He gave as examples the form class consisting of "personal substantive expressions" in English (defined as "the forms that, when spoken with exclamatory final pitch, are calls for a person's presence or attention"; *e.g.,* "John," "Boy," "Mr. Smith"); the form class consisting of "infinitive expressions" (defined as "forms which, when spoken with exclamatory final pitch, have the meaning of a command"; *e.g.,* "run," "jump," "come here"); the form class of "nominative substantive expressions" (*e.g.,* "John," "the boys"); and so on. It should be clear from these examples that form classes are similar to though not identical with the traditional parts of speech, and that one and the same form can belong to more than one form class.

What Bloomfield had in mind as the criterion for form class membership (and therefore of syntactic equivalence) may best be expressed in terms of substitutability. Form classes are sets of forms (whether simple or complex, free or bound), any one of which may be substituted for any other in a given construction or set of constructions throughout the sentences of the language.

Constituents and immediate constituents

The smaller forms into which a larger form may be analyzed are its constituents, and the larger form is a construction. For example, the phrase "poor John" is a construction analyzable into, or composed of, the constituents "poor" and "John." Because there is no intermediate unit of which "poor" and "John" are constituents that is itself a constituent of the construction "poor John," the forms "poor" and "John" may be described not only as constituents but also as immediate constituents of "poor John." Similarly, the phrase "lost his watch" is composed of three word forms, "lost," "his," and "watch," all of which may be described as constituents of the construction. Not all of them, however, are its immediate constituents. The forms "his" and "watch" combine to make the intermediate construction "his watch"; it is this intermediate unit that combines with "lost" to form the larger phrase "lost his watch." The immediate constituents of "lost his watch" are "lost" and "his watch"; the immediate constituents of "his watch" are the forms "his" and "watch." By the constituent structure of a phrase or sentence is meant the hierarchical organization of the smallest forms of which it is composed (its ultimate constituents) into layers of successively more inclusive units. Viewed in this way, the sentence "Poor John lost his watch" is more than simply a sequence of five word forms associated with a particular intonation pattern. It is analyzable into the immediate constituents "poor John" and "lost his watch," and each of these phrases is analyzable into its own immediate constituents and so on, until, at the last stage of the analysis, the ultimate constituents of the sentence are reached. The constituent structure of the whole sentence is represented by means of a tree diagram in Figure 1.

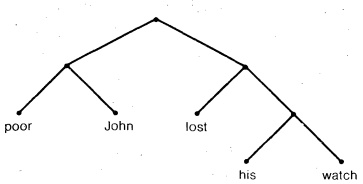


Figure 1: The constituent structure of a simple sentence (see text).

Each form, whether it is simple or composite, belongs to a certain form class. Using arbitrarily selected letters to denote the form classes of English, "poor" may be a member of the form class *A,* "John" of the class *B,* "lost" of the class *C,* "his" of the class *D,* and "watch" of the class *E.* Because "poor John" is syntactically equivalent to (*i.e.,* substitutable for) "John," it is to be classified as a member of *A.* So too, it can be assumed, is "his watch." In the case of "lost his watch" there is a problem. There are very many forms—including "lost," "ate," "stole"—that can occur, as here, in construction with a member of *A* and can also occur alone; for example, "lost" is substitutable for "stole the money," as "stole" is substitutable for either or for "lost his watch." This being so, one might decide to classify constructions like "lost his watch" as members of *C.* On the other hand, there are forms that—though they are substitutable for "lost," "ate," "stole," and so on when these forms occur alone—cannot be used in combination with a following member of *A* (*cf.* "died," "existed"); and there are forms that, though they may be used in combination with a following member of *A,* cannot occur alone (*cf.* "enjoyed"). The question is whether one respects the traditional distinction between transitive and intransitive verb forms. It may be decided, then, that "lost," "stole," "ate" and so forth belong to one class, *C* (the class to which "enjoyed" belongs), when they occur "transitively" (*i.e.,* with a following member of *A* as their object) but to a different class, *F* (the class to which "died" belongs), when they occur "intransitively." Finally, it can be said that the whole sentence "Poor John lost his watch" is a member of the form class *G.* Thus the constituent structure not only of "Poor John lost his watch" but of a whole set of English sentences can be represented by means of the tree diagram given in Figure 2. New sentences of the same type can be constructed by substituting actual forms for the class labels.

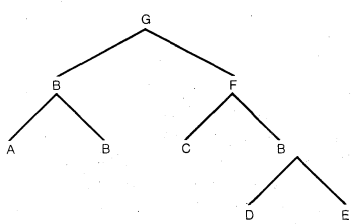


Figure 2: The constituent structure of a class of simple sentences with arbitrary letters used to represent the form class of each constituent (see text).

Endocentric and exocentric constructions

Any construction that belongs to the same form class as at least one of its immediate constituents is described as endocentric; the only endocentric construction in the model sentence above is "poor John." All the other constructions, according to the analysis, are exocentric. This is clear from the fact that in Figure 2 the letters at the nodes above every phrase other than the phrase *A* + *B* (*i.e.,* "poor John," "old Harry," and so on) are different from any of the letters at the ends of the lower branches connected directly to these nodes. For example, the phrase *D* + *E* (*i.e.,* "his watch," "the money," and so forth) has immediately above it a node labelled *B,* rather than either *D* or *E.* Endocentric constructions fall into two types: subordinating and coordinating. If attention is confined, for simplicity, to constructions composed of no more than two immediate constituents, it can be said that subordinating constructions are those in which only one immediate constituent is of the same form class as the whole construction, whereas coordinating constructions are those in which both constituents are of the same form class as the whole construction. In a subordinating construction (*e.g.,* "poor John"), the constituent that is syntactically equivalent to the whole construction is described as the head, and its partner is described as the modifier: thus, in "poor John," the form "John" is the head, and "poor" is its modifier. An example of a coordinating construction is "men and women," in which, it may be assumed, the immediate constituents are the word "men" and the word

"women," each of which is syntactically equivalent to "men and women." (It is here implied that the conjunction "and" is not a constituent, properly so called, but an element that, like the relative order of the constituents, indicates the nature of the construction involved. Not all linguists have held this view.)

Ambiguous constructions

One reason for giving theoretical recognition to the notion of constituent is that it helps to account for the ambiguity of certain constructions. A classic example is the phrase "old men and women," which may be interpreted in two different ways according to whether one associates "old" with "men and women" or just with "men." Under the first of the two interpretations, the immediate constituents are "old" and "men and women"; under the second, they are "old men" and "women." The difference in meaning cannot be attributed to any one of the ultimate constituents but results from a difference in the way in which they are associated with one another. Ambiguity of this kind is referred to as syntactic ambiguity. Not all syntactic ambiguity is satisfactorily accounted for in terms of constituent structure.

Semantics. Bloomfield thought that semantics, or the study of meaning, was the weak point in the scientific investigation of language and would necessarily remain so until the other sciences whose task it was to describe the universe and man's place in it had advanced beyond their present state. In his textbook *Language* (1933), he had himself adopted a behaviouristic theory of meaning, defining the meaning of a linguistic form as "the situation in which the speaker utters it and the response which it calls forth in the hearer." Furthermore, he subscribed, in principle at least, to a physicalist thesis, according to which all science should be modelled upon the so-called exact sciences and all scientific knowledge should be reducible, ultimately, to statements made about the properties of the physical world. The reason for his pessimism concerning the prospects for the study of meaning was his feeling that it would be a long time before a complete scientific description of the situations in which utterances were produced and the responses they called forth in their hearers would be available. At the time that Bloomfield was writing, physicalism was more widely held than it is today, and it was perhaps reasonable for him to believe that linguistics should eschew mentalism and concentrate upon the directly observable. As a result, for some 30 years after the publication of Bloomfield's textbook, the study of meaning was almost wholly neglected by his followers; most American linguists who received their training during this period had no knowledge of, still less any interest in, the work being done elsewhere in semantics.

Result of Bloomfield's physicalist tendencies

Two groups of scholars may be seen to have constituted an exception to this generalization: anthropologically minded linguists and linguists concerned with Bible translation. Much of the description of the indigenous languages of America has been carried out since the days of Boas and his most notable pupil Sapir by scholars who were equally proficient both in anthropology and in descriptive linguistics; such scholars have frequently added to their grammatical analyses of languages some discussion of the meaning of the grammatical categories and of the correlations between the structure of the vocabularies and the cultures in which the languages operated. It has already been pointed out that Boas and Sapir and, following them, Whorf were attracted by Humboldt's view of the interdependence of language and culture and of language and thought. This view was quite widely held by American anthropological linguists (athough many of them would not go as far as Whorf in asserting the dependence of thought and conceptualization upon language).

Also of considerable importance in the description of the indigenous languages of America has been the work of linguists trained by the American Bible Society and the Summer Institute of Linguistics, a group of Protestant missionary linguists. Because their principal aim is to produce translations of the Bible, they have necessarily been concerned with meaning as well as with grammar and phonology. This has tempered the otherwise fairly orthodox Bloomfieldian approach characteristic of the group.

The two most important developments evident in recent work in semantics are, first, the application of the structural approach to the study of meaning and, second, a better appreciation of the relationship between grammar and semantics. The second of these developments will be treated in the following section on *Transformational-generative grammar.* (See also the separate section below, *Semantics.*) The first, structural semantics, goes back to the period preceding World War II and is exemplified in a large number of publications, mainly by German scholars—Jost Trier, Leo Weisgerber, and their collaborators.

The structural approach to semantics is best explained by contrasting it with the more traditional "atomistic" approach, according to which the meaning of each word in the language is described, in principle, independently of the meaning of all other words. The structuralist takes the view that the meaning of a word is a function of the relationships it contracts with other words in a particular lexical field, or subsystem, and that it cannot be adequately described except in terms of these relationships. For example, the colour terms in particular languages constitute a lexical field, and the meaning of each term depends upon the place it occupies in the field. Although the denotation of each of the words "green," "blue," and "yellow" in English is somewhat imprecise at the boundaries, the position that each of them occupies relative to the other terms in the system is fixed: "green" is between "blue" and "yellow," so that the phrases "greenish yellow" or "yellowish green" and "bluish green" or "greenish blue" are used to refer to the boundary areas. Knowing the meaning of the word "green" implies knowing what cannot as well as what can be properly described as green (and knowing of the borderline cases that they are borderline cases). Languages differ considerably as to the number of basic colour terms that they recognize, and they draw boundaries within the psychophysical continuum of colour at different places. Blue, green, yellow, and so on do not exist as distinct colours in nature, waiting to be labelled differently, as it were, by different languages; they come into existence, for the speakers of particular languages, by virtue of the fact that those languages impose structure upon the continuum of colour and assign to three of the areas thus recognized the words "blue," "green," "yellow."

Colour terms in different languages

The language of any society is an integral part of the culture of that society, and the meanings recognized within the vocabulary of the language are learned by the child as part of the process of acquiring the culture of the society in which he is brought up. Many of the structural differences found in the vocabularies of different languages are to be accounted for in terms of cultural differences. This is especially clear in the vocabulary of kinship (to which a considerable amount of attention has been given by anthropologists and linguists), but it holds true of many other semantic fields also. A consequence of the structural differences that exist between the vocabularies of different languages is that, in many instances, it is in principle impossible to translate a sentence "literally" from one language to another.

It is important, nevertheless, not to overemphasize the semantic incommensurability of languages. Presumably, there are many physiological and psychological constraints that, in part at least, determine one's perception and categorization of the world. It may be assumed that, when one is learning the denotation of the more basic words in the vocabulary of one's native language, attention is drawn first to what might be called the naturally salient features of the environment and that one is, to this degree at least, predisposed to identify and group objects in one way rather than another. It may also be that human beings are genetically endowed with rather more specific and linguistically relevant principles of categorization. It is possible that, although languages differ in the number of basic colour categories that they distinguish, there is a limited number of hierarchically ordered basic colour categories from which each language makes its selection and that what counts as a typical instance, or focus, of these universal colour categories is fixed and does not vary from one language to another. If this hypothesis is correct, then it is false to say, as many structural semanticists have said, that languages divide the continuum of colour in a quite

arbitrary manner. But the general thesis of structuralism is unaffected, for it still remains true that each language has its own unique semantic structure even though the total structure is, in each case, built upon a substructure of universal distinctions.

TRANSFORMATIONAL-GENERATIVE GRAMMAR

A generative grammar, in the sense in which Noam Chomsky uses the term, is a rules system formalized with mathematical precision that generates, without need of any information that is not represented explicitly in the system, the grammatical sentences of the language that it describes, or characterizes, and assigns to each sentence a structural description, or grammatical analysis. All the concepts introduced in this definition of "generative" grammar will be explained and exemplified in the course of this section. Generative grammars fall into several types; this exposition is concerned mainly with the type known as transformational (or, more fully, transformational-generative). Transformational grammar was initiated by Zellig S. Harris in the course of work on what he called discourse analysis (the formal analysis of the structure of continuous text). It was further developed and given a somewhat different theoretical basis by Chomsky.

Generation of grammatical sentences

Harris's grammar. Harris distinguished within the total set of grammatical sentences in a particular language (for example, English) two complementary subsets: kernel sentences (the set of kernel sentences being described as the kernel of the grammar) and nonkernel sentences. The difference between these two subsets lies in nonkernel sentences being derived from kernel sentences by means of transformational rules. For example, "The workers rejected the ultimatum" is a kernel sentence that may be transformed into the nonkernel sentences "The ultimatum was rejected by the workers" or "Did the workers reject the ultimatum?" Each of these may be described as a transform of the kernel sentence from which it is derived. The transformational relationship between corresponding active and passive sentences (*e.g.*, "The workers rejected the ultimatum" and "The ultimatum was rejected by the workers") is conventionally symbolized by the rule N_1 V $N_2 \rightarrow N_2$ be V + en by N_1, in which N stands for any noun or noun phrase, V for any transitive verb, en for the past participle morpheme, and the arrow (→) instructs one to rewrite the construction to its left as the construction to the right. (There has been some simplification of the rule as it was formulated by Harris.) This rule may be taken as typical of the whole class of transformational rules in Harris's system: it rearranges constituents (what was the first nominal, or noun, N_1, in the kernel sentence is moved to the end of the transform, and what was the second nominal, N_2, in the kernel sentence is moved to initial position in the transform), and it adds various elements in specified positions (be, en, and by). Other operations carried out by transformational rules include the deletion of constituents; *e.g.*, the entire phrase "by the workers" is removed from the sentence "The ultimatum was rejected by the workers" by a rule symbolized as N_2 be V+en by $N_1 \rightarrow N_2$ be V+en. This transforms the construction on the left side of the arrow (which resulted from the passive transformation) by dropping the by-phrase, thus producing "The ultimatum was rejected."

Functions of Harris's rules

Chomsky's grammar. Chomsky's system of transformational grammar, though it was developed on the basis of his work with Harris, differs from Harris's in a number of respects. It is Chomsky's system that has attracted the most attention and has received the most extensive exemplification and further development. As outlined in *Syntactic Structures* (1957), it comprised three sections, or components: the phrase-structure component, the transformational component, and the morphophonemic component. Each of these components consisted of a set of rules operating upon a certain "input" to yield a certain "output." The notion of phrase structure may be dealt with independently of its incorporation in the larger system. In the following system of rules, S stands for Sentence, NP for Noun Phrase, VP for Verb Phrase, Det for Determiner, Aux for Auxiliary (verb), N for Noun, and V for Verb stem.

(1) S → NP + VP
(2) VP → Verb + NP
(3) NP → Det + N
(4) Verb → Aux + V
(5) Det → *the, a, . . .*
(6) N → *man, ball, . . .*
(7) Aux → *will, can, . . .*
(8) V → *hit, see, . . .*

This is a simple phrase-structure grammar. It generates and thereby defines as grammatical such sentences as "The man will hit the ball," and it assigns to each sentence that it generates a structural description. The kind of structural description assigned by a phrase-structure grammar is, in fact, a constituent structure analysis of the sentence.

In these rules, the arrow can be interpreted as an instruction to rewrite (this is to be taken as a technical term) whatever symbol appears to the left of the arrow as the symbol or string of symbols that appears to the right of the arrow. For example, rule (2) rewrites the symbol VP as the string of symbols Verb + NP, and it thereby defines Verb + NP to be a construction of the type VP. Or, alternatively and equivalently, it says that constructions of the type VP may have as their immediate constituents constructions of the type Verb and NP (combined in that order). Rule (2) can be thought of as creating or being associated with the tree structure in Figure 3.

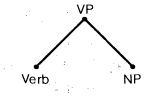


Figure 3: The constituent structure, or phrase structure, assigned by the rule VP → Verb + NP (see text).

Integrated system of rules

Rules (1)–(8) do not operate in isolation but constitute an integrated system. The symbol S (standing mnemonically for "sentence") is designated as the initial symbol. This information is not given in the rules (1)–(8), but it can be assumed either that it is given in a kind of protocol statement preceding the grammatical rules or that there is a universal convention according to which S is always the initial symbol. It is necessary to begin with a rule that has the initial symbol on the left. Thereafter any rule may be applied in any order until no further rule is applicable; in doing so, a derivation can be constructed of one of the sentences generated by the grammar. If the rules are applied in the following order: (1), (2), (3), (3), (4), (5), (5), (6), (6), (7), (8), then assuming that "the" is selected on both applications of (5), "man" on one application of (6), and "ball" on the other, "will" on the application of (7), and "hit" on the application of (8), the following derivation of the sentence "The man will hit the ball" will have been constructed:

(i)	S	
(ii)	NP + VP	by rule (1)
(iii)	NP + Verb + NP	by rule (2)
(iv)	Det + N + Verb + NP	by rule (3)
(v)	Det + N + Verb + Det + N	by rule (3)
(vi)	Det + N + Aux + V + Det + N	by rule (4)
(vii)	*the* + N + Aux + V + Det + N	by rule (5)
(viii)	*the* + N + Aux + V + *the* + N	by rule (5)
(ix)	*the* + *man* + Aux + V + *the* + N	by rule (6)
(x)	*the* + *man* + Aux + V + *the* + *ball*	by rule (6)
(xi)	*the* + *man* + *will* + V + *the* + *ball*	by rule (7)
(xii)	*the* + *man* + *will* + *hit* + *the* + *ball*	by rule (8)

Many other derivations of this sentence are possible, depending on the order in which the rules are applied. The important point is that all these different derivations are equivalent in that they can be reduced to the same tree diagram; namely, the one shown in Figure 4. If this is compared with the system of rules, it will be seen that each application of each rule creates or is associated with a portion (or subtree) of the tree. The tree diagram, or phrase marker, may now be considered as a structural description of the sentence "The man hit the ball." It is a description of the constituent structure, or phrase structure, of the sentence, and it is assigned by the rules that generate the sentence.

It is important to interpret the term generate in a static, rather than a dynamic, sense. The statement that the

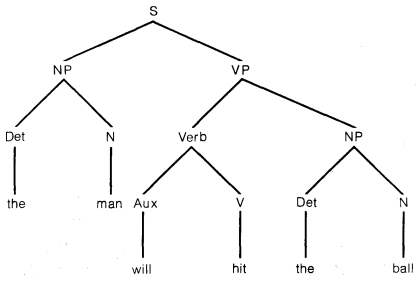


Figure 4: Structural description of the sentence "The man will hit the ball," assigned by the rules of a simple phrase-structure grammar (see text).

grammar generates a particular sentence means that the sentence is one of the totality of sentences that the grammar defines to be grammatical or well formed. All the sentences are generated, as it were, simultaneously. The notion of generation must be interpreted as would be a mathematical formula containing variables. For example, in evaluating the formula $y^2 + y$ for different values of y, one does not say that the formula itself generates these various resultant values (2, when $y = 1$; 5, when $y = 2$; etc.) one after another or at different times; one says that the formula generates them all simultaneously or, better still perhaps, timelessly. The situation is similar for a generative grammar. Although one sentence rather than another can be derived on some particular occasion by making one choice rather than another at particular places in the grammar, the grammar must be thought of as generating all sentences statically or timelessly.

Phrase-structure and transformational rules

It has been noted that, whereas a phrase-structure grammar is one that consists entirely of phrase-structure rules, a transformational grammar (as formalized by Chomsky) includes both phrase-structure and transformational rules (as well as morphophonemic rules). The transformational rules depend upon the prior application of the phrase-structure rules and have the effect of converting, or transforming, one phrase marker into another. What is meant by this statement may be clarified first with reference to a purely abstract and very simple transformational grammar, in which the letters stand for constituents of a sentence (and S stands for "sentence"):

PS rules
(1) $S \rightarrow A + B$
(2) $B \rightarrow C + D$
(3) $A \rightarrow a + b$
(4) $C \rightarrow c + e + f$
(5) $D \rightarrow d + g + h$
T rules
(6) $A + C + D \rightarrow D + A$

The first five rules are phrase-structure rules (PS rules); rule (6) is a transformational rule (T rule). The output of rules (1)–(5) is the terminal string $a + b + c + e + f + d + g + h$, which has associated with it the structural description indicated by the phrase marker shown in Figure 5 (left). Rule (6) applies to this terminal string of the PS rules and the associated phrase marker. It has the effect of deleting C (and the constituents of C) and permuting A and D (together with their constituents). The result is the string of symbols $d + g + h + a + b$, with the associated phrase marker shown in Figure 5 (right).

The phrase marker shown in Figure 5 (left) may be described as underlying, and the phrase marker shown in Figure 5 (right) as derived with respect to rule (6). One of the principal characteristics of a transformational rule is its transformation of an underlying phrase marker into a derived phrase marker in this way. Transformational rules, in contrast with phrase-structure rules, are also formally more heterogeneous and may have more than one symbol on the left-hand side of the arrow. The linguistic importance of these abstract considerations may be explained with reference to the relationship that holds in English between active and passive sentences.

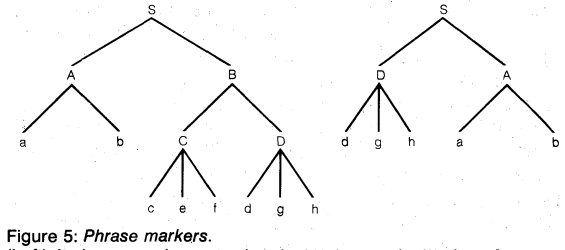


Figure 5: *Phrase markers.*
(Left) A phrase marker associated with the terminal string of a set of phrase-structure rules. (Right) The phrase marker resulting from the application to the phrase marker shown at left of the transformational rule $A + C + D \rightarrow D + A$ (see text).

Chomsky's rule for relating active and passive sentences (as given in *Syntactic Structures*) is very similar, at first sight, to Harris's, discussed above. Chomsky's rule is:

$$NP_1 - Aux - V - NP_2 \rightarrow NP_2 - Aux + be + en - V - by + NP_1$$

Passive transformation

This rule, called the passive transformation, presupposes and depends upon the prior application of a set of phrase-structure rules. For simplicity, the passive transformation may first be considered in relation to the set of terminal strings generated by the phrase-structure rules (1)–(8) given earlier. The string "the + man + will + hit + the + ball" (with its associated phrase marker, as shown in Figure 4) can be treated not as an actual sentence but as the structure underlying both the active sentence "The man will hit the ball" and the corresponding passive "The ball will be hit by the man." The passive transformation is applicable under the condition that the underlying, or "input," string is analyzable in terms of its phrase structure as NP – Aux – V – NP (the use of subscript numerals to distinguish the two NPs in the formulation of the rule is an informal device for indicating the operation of permutation). In the phrase marker in Figure 4 "the" + "man" are constituents of NP, "will" is a constituent of Aux, "hit" is a constituent of V, and "the" + "ball" are constituents of NP. The whole string is therefore analyzable in the appropriate sense, and the passive transformation converts it into the string "the + ball + will + be + en + hit + by + the + man." A subsequent transformational rule will permute "en + hit" to yield "hit + en," and one of the morphophonemic rules will then convert "hit + en" to "hit" (as "ride + en" will be converted to "ridden"; "open + en" to "opened," and so on).

Every transformational rule has the effect of converting an underlying phrase marker into a derived phrase marker. The manner in which the transformational rules assign derived constituent structure to their input strings is one of the major theoretical problems in the formalization of transformational grammar. Here it can be assumed not only that "be + en" is attached to Aux and "by" to NP (as indicated by the plus signs in the rule as it has been formulated above) but also that the rest of the derived structure is as shown in Figure 6. The phrase marker in Figure 6 formalizes the fact, among others, that "the ball" is the subject of the passive sentence "The ball will be hit by the man," whereas "the man" is the subject of the corresponding active "The man will hit the ball" (*cf.* Figure 4).

Although the example above is a very simple one, and only a single transformational rule has been considered independently of other transformational rules in the same system, the passive transformation must operate, not only upon simple noun phrases like "the man" or "the ball," but upon noun phrases that contain adjectives ("the old man"), modifying phrases ("the man in the corner"), relative clauses ("the man who checked in last night"), and so forth. The incorporation, or embedding, of these other structures with the noun phrase will be brought about by the prior application of other transformational rules. It should also be clear that the phrase-structure rules require extension to allow for the various forms of the verb ("is hitting," "hit," "was hitting," "has hit," "has been hitting," etc.) and for the distinction of singular and plural.

Operation of transformations on underlying structures

It is important to note that, unlike Harris's, Chomsky's system of transformational grammar does not convert one sentence into another: the transformational rules operate upon the structures underlying sentences and not upon actual sentences. A further point is that even the simplest sentences (*i.e.*, kernel sentences) require the application of

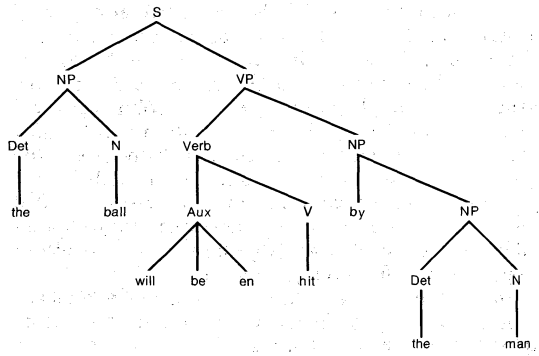


Figure 6: A possible derived phrase marker for a passive sentence (see text).

at least some transformational rules. Corresponding active and passive sentences, affirmative and negative sentences, declarative and interrogative sentences, and so on are formally related by deriving them from the same underlying terminal string of the phrase-structure component. The difference between kernel sentences and nonkernel sentences in *Syntactic Structures* (in Chomsky's later system the category of kernel sentences is not given formal recognition at all) resides in the fact that kernel sentences are generated without the application of any optional transformations. Nonkernel sentences require the application of both optional and obligatory transformations, and they differ one from another in that a different selection of optional transformations is made.

Modifications in Chomsky's grammar. Chomsky's system of transformational grammar was substantially modified in 1965. Perhaps the most important modification was the incorporation, within the system, of a semantic component, in addition to the syntactic component and phonological component. (The phonological component may be thought of as replacing the morphophonemic component of *Syntactic Structures*.) The rules of the syntactic component generate the sentences of the language and assign to each not one but two structural analyses: a deep structure analysis as represented by the underlying phrase marker, and a surface structure analysis, as represented by the final derived phrase marker. The underlying phrase marker is assigned by rules of the base (roughly equivalent to the PS [Phrase-Structure] rules of the earlier system); the derived phrase marker is assigned by the transformational rules. The interrelationship of the four sets of rules is shown diagrammatically in Figure 7. The

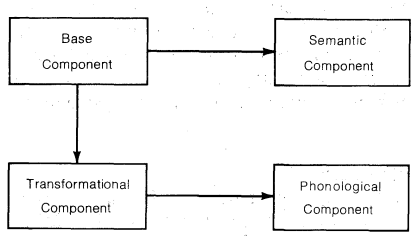


Figure 7: Diagrammatic representation of a transformational grammar (see text).

meaning of the sentence is derived (mainly, if not wholly) from the deep structure by means of the rules of semantic interpretation; the phonetic realization of the sentence is derived from its surface structure by means of the rules of the phonological component. The grammar ("grammar" is now to be understood as covering semantics and phonology, as well as syntax) is thus an integrated system of rules for relating the pronunciation of a sentence to its meaning. The syntax, and more particularly the base, is at the "heart" of the system, as it were: it is the base component (as the arrows in the diagram indicate) that generates the infinite class of structures underlying the well-formed sentences of a language. These structures are then given a semantic and phonetic "interpretation" by the other components.

The base consists of two parts: a set of categorial rules and a lexicon. Taken together, they fulfill a similar function to that fulfilled by the phrase-structure rules of the earlier system. But there are many differences of detail. Among the most important is that the lexicon (which may be thought of as a dictionary of the language cast in a particular form) lists, in principle, all the vocabulary words in the language and associates with each all the syntactic, semantic, and phonological information required for the correct operation of the rules. This information is represented in terms of what are called features. For example, the entry for "boy" might say that it has the syntactic features: [+ Noun], [+ Count], [+ Common], [+ Animate], and [+ Human]. The categorial rules generate a set of phrase markers that have in them, as it were, a number of "slots" to be filled with items from the lexicon. With each such "slot" there is associated a set of features that define the kind of item that can fill the "slot." If a phrase marker is generated with a "slot" for the head of a noun phrase specified as requiring an animate noun (*i.e.*, a noun having the feature [+ Animate]), the item "boy" would be recognized as being compatible with this specification and could be inserted in the "slot" by the rule of lexical substitution. Similarly, it could be inserted in "slots" specified as requiring a common noun, a human noun, or a countable noun, but it would be excluded from positions that require an abstract noun (*e.g.*, "sincerity") or an uncountable noun (*e.g.*, "water"). By drawing upon the syntactic information coded in feature notation in the lexicon, the categorial rules might permit such sentences as "The boy died," while excluding (and thereby defining as ungrammatical) such nonsentences as "The boy elapsed."

Base: categorial rules and a lexicon

One of the most controversial topics in the development of transformational grammar is the relationship between syntax and semantics. Scholars working in the field are now agreed that there is a considerable degree of interdependence between the two, and the problem is how to formalize this interdependence. One school of linguists, called generative semanticists, accept the general principles of transformational grammar but have challenged Chomsky's conception of deep structure as a separate and identifiable level of syntactic representation. In their opinion, the basic component of the grammar should consist of a set of rules for the generation of well-formed semantic representations. These would then be converted by a succession of transformational rules into strings of words with an assigned surface-structure syntactic analysis, there being no place in the passage from semantic representation to surface structure identifiable as Chomsky's deep structure. Chomsky himself has denied that there is any real difference between the two points of view and has maintained that the issue is purely one of notation. That this argument can be put forward by one party to the controversy and rejected by the other is perhaps a sufficient indication of the uncertainty of the evidence. Of greater importance than the overt issues, in so far as they are clear, is the fact that linguists are now studying much more intensively than they have in the past the complexities of the interdependence of syntax, on the one hand, and semantics and logic, on the other. Whether it will prove possible to handle all these complexities within a comprehensive generative grammar remains to be seen.

Generative semanticists

The role of the phonological component of a generative grammar of the type outlined by Chomsky is to assign a phonetic "interpretation" to the strings of words generated by the syntactic component. These strings of words are represented in a phonological notation (taken from the lexicon) and have been provided with a surface-structure analysis by the transformational rules (see Figure 7). The phonological elements out of which the word forms are composed are segments consisting of what are referred to technically as distinctive features (following the usage of the Prague school, see below). For example, the word form "man," represented phonologically, is composed of three segments: the first consists of the features [+ consonantal], [+ bilabial], [+ nasal], etc.; the second of the features [+ vocalic], [+ front], [+ open], etc.; and the third of the features [+ consonantal], [+ alveolar], [+ nasal], etc. (These features should be taken as purely illustrative; there is some doubt about the definitive list of distinctive

features.) Although these segments may be referred to as the "phonemes" /m/, /a/, and /n/, they should not be identified theoretically with units of the kind discussed in the section on *Phonology* under *Structural linguistics.* They are closer to what many American structural linguists called "morphophonemes" or the Prague school linguists labelled "archiphonemes," being unspecified for any feature that is contextually redundant or predictable. For instance, the first segment of the phonological representation of "man" will not include the feature [+ voice]; because nasal consonants are always phonetically voiced in this position in English, the feature [+ voice] can be added to the phonetic specification by a rule of the phonological component.

One further important aspect of generative phonology (*i.e.,* phonology carried out within the framework of an integrated generative grammar) should be mentioned: its dependence upon syntax. Most American structural phonologists made it a point of principle that the phonemic analysis of an utterance should be carried out without regard to its grammatical structure. This principle was controversial among American linguists and was not generally accepted outside America. Not only has the principle been rejected by the generative grammarians, but they have made the phonological description of a language much more dependent upon its syntactic analysis than has any other school of linguists. They have claimed, for example, that the phonological rules that assign different degrees of stress to the vowels in English words and phrases and alter the quality of the relatively unstressed vowel concomitantly must make reference to the derived constituent structure of sentences and not merely to the form class of the individual words or the places in which the word boundaries occur.

Dependence of generative phonology on syntax

TAGMEMICS

The system of tagmemic analysis, as presented by Kenneth L. Pike, was developed for the analysis not only of language but of all of human behaviour that manifests the property of patterning. In the following treatment, only language will be discussed.

Modes of language. Every language is said to be trimodal—*i.e.,* structured in three modes: phonology, grammar, and lexicon. These modes are interrelated but have a considerable degree of independence and must be described in their own terms. Phonology and lexicon should not be seen as mere appendages to grammar, the former simply specifying which phonemes can combine to form morphemes (or morphs), and the latter simply listing the morphemes and other meaningful units with a description of their meaning. There are levels of structure in each of the modes, and the units of one level are not necessarily coterminous with those of another. Phonemes, for example, may combine to form syllables and syllables to form phonological words ("phonological word" is defined as the domain of some phonological process such as accentuation, assimilation, or dissimilation), but the morpheme (or morph) will not necessarily consist of an integral number of syllables, still less of a single syllable. Nor will the word as a grammatical unit necessarily coincide with the phonological word. Similarly, the units of lexical analysis, sometimes referred to as lexemes (in one sense of this term), are not necessarily identifiable as single grammatical units, whether as morphemes, words, or phrases. No priority, then, is ascribed to any one of the three modes.

Tagmeme defined

The originality of tagmemic analysis and the application of the term tagmeme is most clearly manifest in the domain of grammar. By a tagmeme is meant an element of a construction, the element in question being regarded as a composite unit, described in such terms as "slot-filler" or "function-class." For example, one of the tagmemes required for the analysis of English at the syntactic level might be noun-as-subject, in which "noun" refers to a class of substitutable, or paradigmatically related, morphemes or words capable of fulfilling a certain grammatical function, and "subject" refers to the function that may be fulfilled by one or more classes of elements. In the tagmeme noun-as-subject—which, using the customary tagmemic symbolism, may be represented as Subject:noun—the subject slot is filled by a noun. When a particular tagmeme is identified in the analysis of an actual utterance, it is said to be manifested by the particular member of the grammatical class that occurs in the appropriate slot in the utterance. For example, in the utterance "John is asleep," the subject tagmeme is manifested by the noun "John." Tagmemicists insist that tagmemes, despite their bipartite structure, are single units. In grammatical analysis, the distribution of tagmemes, not simply of classes, is stated throughout the sentences of the language. Subject:noun is a different tagmeme from Object:noun, as it is also a different tagmeme from Subject:pronoun.

Hierarchy of levels. Within the grammar of a language there is a hierarchy of levels, units of one level being composed of sequences of units of the level below. In many languages, five such levels are recognized, defined in terms of the following units: morpheme, word, phrase, clause, and sentence. (The term level is being used in a different sense from that in which it was used earlier to refer to phonology and grammar.) The difference between morphology and syntax is simply a difference between two of these five levels, no greater than the difference, for example, between the phrase level and the clause level. Normally, tagmemes at one level are manifested by units belonging to the level below: clause tagmemes by phrases, phrase tagmemes by words, and so on. Intermediate levels may, however, be skipped. For example, the subject tagmeme in a clause may be manifested by a single word in English (*e.g.,* "John," "water") and not necessarily by a phrase ("the young man").

It is also possible for there to be loop-backs in the grammatical hierarchy of a language. This means that a unit of higher level may be embedded within the structure of a unit of lower level; for example, a clause may fill a slot within a phrase (*e.g.,* "who arrived late," in "the man who arrived late").

Tagmemic notation

In regard to the notation of tagmemics, a construction is symbolized as a string of tagmemes (which commonly, though not necessarily, will be sequentially ordered according to the order in which elements manifesting the tagmemes occur in utterances). Each tagmeme is marked as obligatory or optional by having preposed to it a plus sign (+) or a plus-or-minus sign (±), respectively. For example, a formula representing the structure of a clause composed of an obligatory subject tagmeme, an obligatory predicate tagmeme, and an optional object tagmeme might be Cl = + S:n + P:v ± O:n (in which Cl stands for a clause of a certain type and n and v stand for the classes of nouns and verbs, respectively). This formula does not represent in any way the fact (if it is a fact) that the predicate tagmeme and object tagmeme together form a unit that is one of the two immediate constituents of the clause. It is one of the characteristic features of tagmemic grammar that it gives much less emphasis to the notion of constituent structure than other American approaches to grammatical analysis.

STRATIFICATIONAL GRAMMAR

This system of analysis (whose principal advocate is Sydney M. Lamb, a U.S. linguist) is called stratificational because it is based upon the notion that every language comprises a restricted number of structural layers or strata, hierarchically related in such a way that units or combinations of units on one stratum realize units or combinations of units of the next higher stratum. The number of strata may vary from language to language. Four strata have been recognized for English, and it is probable that all languages may have at least these four: the sememic, the lexemic, the morphemic, and the phonemic strata. The sememic stratal system constitutes the semology of the language; the lexemic and morphemic stratal systems constitute the grammar (in the narrower sense of this term); and the phonemic system constitutes the phonology. In some later stratificational work, the term grammar covers the three higher stratal systems—the sememic, the lexemic, and the morphemic—and is opposed to "phonology." The deep structure of sentences is described on the sememic stratum and the surface structure on the morphemic. In the present account, "grammar" is used in

Sememic, lexemic, morphemic, and phonemic strata

the narrower sense and will be opposed to "semology" as well as "phonology."

The originality of stratificational grammar does not reside in the recognition of these three major components of a linguistic description. The stratificational approach to linguistic description is distinguished from others in that it relates grammar to semology and phonology by means of the same notion of realization that it employs to relate the lexemic and the morphemic stratal systems within the grammatical component. Another distinguishing feature of stratificational grammar, in its later development at least, is its description of linguistic structure in terms of a network of relationships, rather than by means of a system of rules; linguistic units are said to be nothing more than points, or positions, in the relational network.

Technical terminology. Lamb has been very careful to make the terminology of stratificational grammar as consistent and perspicuous as possible; but, in fitting some of the more or less established terms into his own theoretical framework, he has reinterpreted them in a potentially confusing manner. Thus, the same terms have been used in different senses in different versions of the system. For example, "morpheme" in stratificational grammar corresponds neither to the unit to which Bloomfield applied the term (*i.e.*, to a word segment consisting of phonemes) nor to the more abstract grammatical unit that a Bloomfieldian morpheme might be described as representing (*e.g.*, the past-tense morpheme that might be variously represented by such allomorphs as /id/, /t/, /d/, etc.). Lamb describes the morpheme as a unit composed of morphons (roughly equivalent to what other linguists have called morphophonemes) that is related to a combination of one or more compositional units of the stratum above, lexons, by means of the relationship of realization. For example, the word form "hated" realizes (on the morphemic stratum) a combination of two lexons, one of which, the stem, realizes the lexeme HATE and the other, the suffix, realizes the PAST TENSE lexeme; each of these two lexons is realized on the stratum below by a morpheme. Another example brings out more clearly the difference between morphemes (the minimal grammatical elements) and lexemes (the minimal meaningful elements). The word form "understood" realizes a combination of three morphemes UNDER, STAND, and PAST. UNDER and STAND jointly realize the single lexeme UNDERSTAND (whose meaning cannot be described as a function of the meanings of UNDER and STAND), whereas the single PAST morpheme directly realizes the single lexeme PAST TENSE.

Morpheme defined in stratificational terms

The stratificational framework, presented in Lamb's work, consistently separates compositional and realizational units, the former being designated by terms ending in the suffix *-on* (semon, lexon, morphon, phonon), the latter by terms ending in the suffix *-eme* (sememe, lexeme, morpheme, phoneme). Ons are components or compounds of emes on the same stratum (semons are components of sememes, lexons are composed of lexemes, etc.) and emes realize ons of the stratum above (phonemes realize morphons, morphemes realize lexons, etc.). Each stratum has its own combinatorial pattern specifying the characteristic combinations of elements on that stratum. Syllable structure is specified on the phonemic stratum, the structure of word forms on the morphemic stratum, the structure of phrases on the lexemic stratum, and the structure of clauses and sentences on the sememic stratum. Phonons are roughly equivalent to phonological distinctive features and include such properties or components of phonemes as labial, nasal, and so on. Sememes are roughly equivalent to what other linguists have called semantic components or features and include such aspects of the meaning of the lexeme "man" as "male," "adult," "human," and so forth. Once again, however, compositional function is distinguished from interstratal realizational function, so that no direct equivalence can be established with nonstratificational terminology. In more recent work in stratificational grammar, the notion that emes are composed of ons has been abandoned, and greater emphasis is laid upon the fact that emes are points, or positions, in a relational network; they are connected to other points in the network but have themselves no internal structure.

Interstratal relationships. One of the principal characteristics of the stratificational approach is that it sets out to describe languages without making use of rules that convert one entity into another. (Reference has been made above to the antipathy many linguists have felt towards describing languages in terms of processes.) The stratificationalist would handle the phenomena in terms of the interstratal relationships of realization. Various kinds of interstratal relationships, other than that of one-to-one correspondence may be recognized: diversification, in which one higher unit has alternative realizations; zero realization, in which a higher unit has no overt realization on the lower stratum; neutralization, in which two or more higher units are mapped into the same lower level unit; and so on. All these interstratal one–many or many–one relations are then analyzed in terms of the logical notions of conjunction and disjunction (AND-relations versus OR-relations), of ordering (x precedes y in an AND-relationship, x is selected in preference to y in an OR-relationship), and the directionality ("upward" towards meaning, or "downward" towards sound). Many of the phenomena that are described by other linguists in terms of processes that derive one unit from another can be described elegantly enough in terms of interstratal relationships of this kind.

Types of interstratal relationships

Critics, however, have objected to the proliferation of strata and theoretical constructs in stratificational grammar, arguing that they result from an a priori commitment to the notion of realization and that the only stratal distinction for which there is any independent evidence is the distinction of phonology and grammar. It has been suggested by Lamb that stratificational grammar provides a model for the way in which linguistic information is stored in the brain and activated during the production and reception of speech. But little is known yet about the neurology of language and speech, and it would be premature to draw any firm conclusions about this aspect of stratificational grammar (see below *Psycholinguistics*).

THE PRAGUE SCHOOL

What is now generally referred to as the Prague school comprises a fairly large group of scholars, mainly European, who, though they may not themselves have been members of the Linguistic Circle of Prague, derived their inspiration from the work of Vilém Mathesius, Nikolay Trubetskoy, Roman Jakobson and other scholars based in Prague in the decade preceding World War II.

Combination of structuralism and functionalism. The most characteristic feature of the Prague school approach is its combination of structuralism with functionalism. The latter term (like "structuralism") has been used in a variety of senses in linguistics. Here it is to be understood as implying an appreciation of the diversity of functions fulfilled by language and a theoretical recognition that the structure of languages is in large part determined by their characteristic functions. Functionalism, taken in this sense, manifests itself in many of the more particular tenets of Prague school doctrine.

One very famous functional analysis of language, which, though it did not originate in Prague, was very influential there, was that of the German psychologist Karl Bühler, who recognized three general kinds of function fulfilled by language: *Darstellungsfunktion, Kundgabefunktion,* and *Appelfunktion.* These terms may be translated, in the present context, as the cognitive, the expressive, and the conative (or instrumental) functions. The cognitive function of language refers to its employment for the transmission of factual information; by expressive function is meant the indication of the mood or attitude of the speaker (or writer); and by the conative function of language is meant its use for influencing the person one is addressing or for bringing about some practical effect. A number of scholars working in the Prague tradition have suggested that these three functions correlate in many languages, at least partly, with the grammatical categories of mood and person. The cognitive function is fulfilled characteristically by 3rd-person nonmodal utterances (*i.e.*, utterances in the indicative mood, making no use of modal verbs); the expressive function by 1st-person utterances in the subjunctive or optative mood; and the conative function by 2nd-person

Various functions of language

utterances in the imperative. The functional distinction of the cognitive and the expressive aspects of language has also been applied by Prague school linguists in their work on stylistics and literary criticism. One of their key principles is that language is being used poetically or aesthetically when the expressive aspect is predominant, and that it is typical of the expressive function of language that this should be manifest in the form of an utterance and not merely in the meanings of the component words.

Phonological contributions. The Prague school is best known for its work on phonology. Unlike the American phonologists, Trubetskoy and his followers did not take the phoneme to be the minimal unit of analysis. Instead, they defined phonemes as sets of distinctive features. For example, in English, /b/ differs from /p/ in the same way that /d/ differs from /t/ and /g/ from /k/. Just how they differ in terms of their articulation is a complex question. For simplicity, it may be said that there is just one feature, the presence of which distinguishes /b/, /d/, and /g/ from /p/, /t/, and /k/, and that this feature is voicing (vibration of the vocal cords). Similarly, the feature of labiality can be extracted from /p/ and /b/ by comparing them with /t/, /d/, /k/, and /g/; the feature of nasality from /n/ and /m/ by comparing them with /t/ and /d/, on the one hand, and with /p/ and /b/, on the other. Each phoneme, then, is composed of a number of articulatory features and is distinguished by the presence or absence of at least one feature from every other phoneme in the language. The distinctive function of phonemes, which depends upon and supports the principle of the duality of structure, can be related to the cognitive function of language. This distinctive feature analysis of Prague school phonology as developed by Jakobson has become part of the generally accepted framework for generative phonology (see above).

Phonemes as distinctive features

Two other kinds of phonologically relevant function are also recognized by linguists of the Prague school: expressive and demarcative. The former term is employed here in the sense in which it was employed above (*i.e.*, in opposition to "cognitive"); it is characteristic of stress, intonation, and other suprasegmental aspects of language that they are frequently expressive of the mood and attitude of the speaker in this sense. The term demarcative is applied to those elements or features that in particular languages serve to indicate the occurrence of the boundaries of words and phrases and, presumably, make it easier to identify such grammatical units in the stream of speech. There are, for example, many languages in which the set of phonemes that can occur at the beginning of a word differs from the set of phonemes that can occur at the end of a word. These and other devices are described by the Prague school phonologists as having demarcative function: they are boundary signals that reinforce the identity and syntagmatic unity of words and phrases.

Theory of markedness. The notion of markedness was first developed in Prague school phonology but was subsequently extended to morphology and syntax. When two phonemes are distinguished by the presence or absence of a single distinctive feature, one of them is said to be marked and the other unmarked for the feature in question. For example, /b/ is marked and /p/ unmarked with respect to voicing. Similarly, in morphology, the regular English verb can be said to be marked for past tense (by the suffixation of *-ed*) but to be unmarked in the present (*cf.* "jumped" versus "jump"). It is often the case that a morphologically unmarked form has a wider range of occurrences and a less definite meaning than a morphologically marked form. It can be argued, for example, that, whereas the past tense form in English (in simple sentences or the main clause of complex sentences) definitely refers to the past, the so-called present tense form is more neutral with respect to temporal reference: it is nonpast in the sense that it fails to mark the time as past, but it does not mark it as present. There is also a more abstract sense of markedness, which is independent of the presence or absence of an overt feature or affix. The words "dog" and "bitch" provide examples of markedness of this kind on the level of vocabulary. Whereas the use of the word "bitch" is restricted to females of the species, "dog" is applicable to both males and females. "Bitch" is the marked and "dog" the unmarked term, and, as is commonly the case, the unmarked term can be neutral or negative according to context (*cf.* "That dog over there is a bitch" versus "It's not a dog, it's a bitch"). The principle of markedness, understood in this more general or more abstract sense, is now quite widely accepted by linguists of many different schools, and it is applied at all levels of linguistic analysis.

Marked and unmarked features

Recent contributions. Current Prague school work is still characteristically functional in the sense in which this term was interpreted in the pre-World War II period. The most valuable contribution made by the postwar Prague school is probably the distinction of theme and rheme and the notion of "functional sentence perspective" or "communicative dynamism." By the theme of a sentence is meant that part that refers to what is already known or given in the context (sometimes called, by other scholars, the topic or psychological subject); by the rheme, the part that conveys new information (the comment or psychological predicate). It has been pointed out that, in languages with a free word order (such as Czech or Latin), the theme tends to precede the rheme, regardless of whether the theme or the rheme is the grammatical subject and that this principle may still operate, in a more limited way, in languages, like English, with a relatively fixed word order (*cf.* "That book I haven't seen before"). But other devices may also be used to distinguish theme and rheme. The rheme may be stressed ("Jóhn saw Mary") or made the complement of the verb "to be" in the main clause of what is now commonly called a cleft sentence ("It's Jóhn who saw Mary").

The general principle that has guided research in "functional sentence perspective" is that the syntactic structure of a sentence is in part determined by the communicative function of its various constituents and the way in which they relate to the context of utterance. A somewhat different but related aspect of functionalism in syntax is seen in current work in what is called case grammar. Case grammar is based upon a small set of syntactic functions (agentive, locative, benefactive, instrumental, and so on) that are variously expressed in different languages but that are held to determine the grammatical structure of sentences. Although case grammar does not derive directly from the work of the Prague school, it is very similar in inspiration.

Case grammar

Historical (diachronic) linguistics

LINGUISTIC CHANGE

All languages change in the course of time. Written records make it clear that 15th-century English is quite noticeably different from 20th-century English, as is 15th-century French or German from modern French or German. It was the principal achievement of the 19th-century linguists not only to realize more clearly than their predecessors the ubiquity of linguistic change but also to put its scientific investigation on a sound footing by means of the comparative method (see the section *History of linguistics: The 19th century*). This will be treated in greater detail in the following section. Here various kinds, or categories, of linguistic change will be listed and exemplified.

Sound change. Since the beginning of the 19th century, when scholars observed that there were a number of systematic correspondences in related words between the sounds of the Germanic languages and the sounds of what were later recognized as other Indo-European languages, particular attention has been paid in diachronic linguistics to changes in the sound systems of languages.

Certain common types of sound change, most notably assimilation and dissimilation, can be explained, at least partially, in terms of syntagmatic, or contextual, conditioning. By assimilation is meant the process by which one sound is made similar in its place or manner of articulation to a neighbouring sound. For example, the word "cupboard" was presumably once pronounced, as the spelling indicates, with the consonant cluster *pb* in the middle. The *p* was assimilated to *b* in manner of articulation (*i.e.*, voicing was maintained throughout the cluster), and subsequently the resultant double consonant *bb* was

simplified. With a single *b* in the middle and an unstressed second syllable, the word "cupboard," as it is pronounced nowadays, is no longer so evidently a compound of "cup" and "board" as its spelling still shows it to have been. The Italian words *notte* "night" and *otto* "eight" manifest assimilation of the first consonant to the second consonant of the cluster in place of articulation (*cf.* Latin *nocte(m), octo*). Assimilation is also responsible for the phenomenon referred to as umlaut in the Germanic languages. The high front vowel *i* of suffixes had the effect of fronting and raising preceding back vowels and, in particular, of converting an *a* sound into an *e* sound. In Modern German this is still a morphologically productive process (*cf. Mann* "man": *Männer* "men"). In English it has left its mark in such irregular forms as "men" (from **manniz*), "feet" (from **fotiz*), and "length" (from **langþa*).

Umlaut

Dissimilation refers to the process by which one sound becomes different from a neighbouring sound. For example, the word "pilgrim" (French *pèlerin*) derives ultimately from the Latin *peregrinus;* the *l* sound results from dissimilation of the first *r* under the influence of the second *r.* A special case of dissimilation is haplology, in which the second of the two identical or similar syllables is dropped. Examples include the standard modern British pronunciations of "Worcester" and "Gloucester" with two syllables rather than three and the common pronunciation of "library" as if it were written "libry." Both assimilation and dissimilation are commonly subsumed under the principle of "ease of articulation." This is clearly applicable in typical instances of assimilation. It is less obvious how or why a succession of unlike sounds in contiguous syllables should be easier to pronounce than a succession of identical or similar sounds. But a better understanding of this phenomenon, as of other "slips of the tongue," may result from current work in the physiological and neurological aspects of speech production.

Not all sound change is to be accounted for in terms of syntagmatic conditioning. The change of *p, t,* and *k* to *f,* θ (the *th* sound in "thin"), and *h* or of *b, d, g* to *p, t,* and *k* in early Germanic cannot be explained in these terms. Nor can the so-called Great Vowel Shift in English, which, in the 15th century, modified the quality of all the long vowels (*cf.* "profane" : "profanity"; "divine" : "divinity"; and others). Attempts have been made to develop a general theory of sound change, notably by the French linguist André Martinet. But no such theory has yet won universal acceptance, and it is likely that the causes of sound change are multiple.

Phonetic vs. phonological sound change

Sound change is not necessarily phonological; it may be merely phonetic (see above *Structural linguistics: Phonology*). The pronunciation of one or more of the phones realizing a particular phoneme may change slightly without affecting any of the previously existing phonological distinctions; this no doubt happens quite frequently as a language is transmitted from one generation to the next. Two diachronically distinct states of the language would differ in this respect in the same way as two coexistent but geographically or socially distinct accents of the same language might differ. It is only when two previously distinct phonemes are merged or a unitary phoneme splits into two (typically when allophonic variation becomes phonemic) that sound change must definitely be considered as phonological. For example, the sound change of *p* to *f, t* to θ(*th*), and *k* to *h,* on the one hand, and of *b* to *p, d* to *t,* and *g* to *k,* on the other, in early Germanic had the effect of changing the phonological system. The voiceless stops did not become fricatives in all positions; they remained as voiceless stops after *s.* Consequently, the *p* sound that was preserved after *s* merged with the *p* that derived by sound change from *b.* (It is here assumed that the aspirated *p* sound and the unaspirated *p* sound are to be regarded as allophones of the same phoneme). Prior to the Germanic sound shift the phoneme to be found at the beginning of the words for "five" or "father" also occurred after *s* in words for "spit" or "spew"; after the change this was no longer the case.

Grammatical change. A language can acquire a grammatical distinction that it did not have previously, as when English developed the progressive ("He is running") in contrast to the simple present ("He runs"). It can also lose a distinction; *e.g.,* modern spoken French has lost the distinction between the simple past (*Il marcha* "he walked") and the perfect (*Il a marché* "he has walked"). What was expressed by means of one grammatical device may come to be expressed by means of another. For example, in the older Indo-European languages the syntactic function of the nouns and noun phrases in a sentence was expressed primarily by means of case endings (the subject of the sentence being in the nominative case, the object in the accusative case, and so on); in most of the modern Indo-European languages these functions are expressed by means of word order and the use of prepositions. It is arguable, although it can hardly be said to have been satisfactorily demonstrated yet, that the grammatical changes that take place in a language in the course of time generally leave its deep structure unaffected and tend to modify the ways in which the deeper syntactic functions and distinctions are expressed (whether morphologically, by word order, by the use of prepositions and auxiliary verbs, or otherwise), without affecting the functions and distinctions themselves. Many grammatical changes are traditionally accounted for in terms of analogy.

Semantic change. Towards the end of the 19th century, a French scholar, Michel Bréal, set out to determine the laws that govern changes in the meaning of words. This was the task that dominated semantic research until the 1930s, when scholars began to turn their attention to the synchronic study of meaning. Many systems for the classification of changes of meaning have been proposed, and a variety of explanatory principles have been suggested. So far no "laws" of semantic change comparable to the phonologist's sound laws have been discovered. It seems that changes of meaning can be brought about by a variety of causes. Most important, perhaps, and the factor that has been emphasized particularly by the so-called words-and-things movement in historical semantics is the change undergone in the course of time by the objects or institutions that words denote. For example, the English word "car" goes back through Latin *carrus* to a Celtic word for a four-wheeled wagon. It now denotes a very different sort of vehicle; confronted with a model of a Celtic wagon in a museum, one would not describe it as a car.

Causes of changes in meaning

Some changes in the meaning of words are caused by their habitual use in particular contexts. The word "starve" once meant "to die" (*cf.* Old English *steorfan,* German *sterben*); in most dialects of English, it now has the more restricted meaning "to die of hunger," though in the north of England "He was starving" can also mean "He was very cold" (*i.e.,* "dying" of cold, rather than hunger). Similarly, the word "deer" has acquired a more specialized meaning than the meaning "wild animal" that it once bore (*cf.* German *Tier*); and "meat," which originally meant food in general (hence, "sweetmeats" and the archaic phrase "meat and drink") now denotes the flesh of an animal treated as food. In all such cases, the narrower meaning has developed from the constant use of the word in a more specialized context, and the contextual presuppositions of the word have in time become part of its meaning.

Borrowing. Languages borrow words freely from one another. Usually this happens when some new object or institution is developed for which the borrowing language has no word of its own. For example, the large number of words denoting financial institutions and operations borrowed from Italian by the other western European languages at the time of the Renaissance testifies to the importance of the Italian bankers in that period. (The word "bank" itself, in this sense, comes through French from the Italian *banca*). Words now pass from one language to another on a scale that is probably unprecedented, partly because of the enormous number of new inventions that have been made in the 20th century and partly because international communications are now so much more rapid and important. The vocabulary of modern science and technology is very largely international.

THE COMPARATIVE METHOD

The comparative method in historical linguistics is concerned with the reconstruction of an earlier language or

Reconstruction of an earlier language

earlier state of a language on the basis of a comparison of related words and expressions in different languages or dialects derived from it. The comparative method was developed in the course of the 19th century for the reconstruction of Proto-Indo-European and was subsequently applied to the study of other language families. It depends upon the principle of regular sound change—a principle that, as explained above, met with violent opposition when it was introduced into linguistics by the Neogrammarians in the 1870s but by the end of the century had become part of what might be fairly described as the orthodox approach to historical linguistics. Changes in the phonological systems of languages through time were accounted for in terms of sound laws.

Grimm's law. The most famous of the sound laws is Grimm's law (though Grimm himself did not use the term law). Some of the correspondences accounted for by Grimm's law are given in Table 1. It will be observed that,

Table 1: Labial and Dental Stops in the Indo-European Languages

Greek	Latin	Gothic	Sanskrit	Slavic
p	p	f	p	p
b	b	p	b	b
ph	f/b	b	bh	b
t	t	θ	t	t
d	d	t	d	d
th	f/d	d	dh	d

when other Indo-European languages, including Latin and Greek, have a voiced unaspirated stop (*b, d*), Gothic has the corresponding voiceless unaspirated stop (*p, t*) and that, when other Indo-European languages have a voiceless unaspirated stop, Gothic has a voiceless fricative (*f*, θ). The simplest explanation would seem to be that, under the operation of what is now called Grimm's law, in some prehistoric period of Germanic (before the development of a number of distinct Germanic languages), the voiced stops inherited from Proto-Indo-European became voiceless and the voiceless stops became fricatives. The situation with respect to the sounds corresponding to the Germanic voiced stops is more complex. Here there is considerable disagreement between the other languages: Greek has voiceless aspirates (*ph, th*), Sanskrit has voiced aspirates (*bh, dh*), Latin has voiceless fricatives in word-initial position (*f*) and voiced stops in medial position (*b, d*), Slavic has voiced stops (*b, d*), and so on. The generally accepted hypothesis is that the Proto-Indo-European sounds from which the Germanic voiced stops developed were voiced aspirates and that they are preserved in Sanskrit but were changed in the other Indo-European languages by the loss of either voice or aspiration. (Latin, having lost the voice in initial position, subsequently changed both of the resultant voiceless aspirates into the fricative *f*, and it lost the aspiration in medial position.) It is easy to see that this hypothesis yields a simpler account of the correspondences than any of the alternatives. It is also in accord with the fact that voiced aspirates are rare in the languages of the world and, unless they are supported by the coexistence in the same language of phonologically distinct voiceless aspirates (as they are in Hindi and other north Indian languages), appear to be inherently unstable.

Proto-Indo-European reconstruction. Reconstruction of the Proto-Indo-European labial stops (made with the lips) and dental stops (made with the tip of the tongue touching the teeth) is fairly straightforward. More controversial is the reconstruction of the Proto-Indo-European sounds underlying the correspondences shown in Table 2. Accord-

Table 2: Velar and Palatal Stops in the Indo-European Languages

Greek	Latin	Gothic	Sanskrit	Slavic
k	k	h	ś	s
g	g	k	j	z
kh	h/g/f	g	h	z
p/t/k	qu	wh	k	k
b/d/g	v/gu	q	g	g
ph/th/kh	f/v/gu	w	gh	g

ing to the most generally accepted hypothesis, there were in Proto-Indo-European at least two distinct series of velar (or "guttural") consonants: simple velars (or palatals), symbolized as **k*, **g*, and **gh*, and labiovelars, symbolized as **k*^{w}, **g*^{w}, and **g*^{w}*h*. The labiovelars may be thought of as velar stops articulated with simultaneous lip-rounding. In one group of languages, the labial component is assumed to have been lost, in another group the velar component; and it is only in the Latin reflex of the voiceless **k*^{w} that both labiality and velarity are retained (*cf.* Latin *quis* from **k*^{w}*i*-). It is notable that the languages that have a velar for the Proto-Indo-European labiovelar stops (*e.g.*, Sanskrit and Slavic) have a sibilant or palatal sound (*s* or *ś*) for the Proto-Indo-European simple velars. Earlier scholars attached great significance to this fact and thought that it represented a fundamental division of the Indo-European family into a western and an eastern group. The western group—comprising Celtic, Germanic, Italic, and Greek—is commonly referred to as the centum group; the eastern group—comprising Sanskrit, Iranian, Slavic, and others—is called the satem (*satəm*) group. (The words centum and satem come from Latin and Iranian, respectively, and mean "hundred." They exemplify, with their initial consonant, the two different treatments of the Proto-Indo-European simple velars.) Nowadays less importance is attached to the centum–satem distinction. But it is still generally held that in an early period of Indo-European, there was a sound law operative in the dialect or dialects from which Sanskrit, Iranian, Slavic and the other so-called satem languages developed that had the effect of palatalizing the original Proto-Indo-European velars and eventually converting them to sibilants.

Controversy concerning Proto-Indo-European velar consonants

Steps in the comparative method. The information given in the previous paragraphs is intended to illustrate what is meant by a sound law and to indicate the kind of considerations that are taken into account in the application of the comparative method. The first step is to find sets of cognate or putatively cognate forms in the languages or dialects being compared: for example, Latin *decem* = Greek *deka* = Sanskrit *daśa* = Gothic *taihun*, all meaning "ten." From sets of cognate forms such as these, sets of phonological correspondences can be extracted; *e.g.*, (1) Latin *d* = Greek *d* = Sanskrit *d* = Gothic *t*; (2) Latin *e* = Greek *e* = Sanskrit *a* = Gothic *ai* (in the Gothic orthography this represents an *e* sound); (3) Latin *c* (*i.e.*, a *k* sound) = Greek *k* = Sanskrit *ś* = Gothic *h*; (4) Latin *em* = Greek *a* = Sanskrit *a* = Gothic *un*. A set of "reconstructed" phonemes can be postulated (marked with an asterisk by the standard convention) to which the phonemes in the attested languages can be systematically related by means of sound laws. The reconstructed Proto-Indo-European word for "ten" is **dekm*. From this form the Latin word can be derived by means of a single sound change, **m* changes to *em* (usually symbolized as **m* > *em*); the Greek by means of the sound change **m* > *a* (*i.e.*, vocalization of the syllabic nasal and loss of nasality); the Sanskrit by means of the palatalizing sound law, **k* > *ś* and the sound change **m* > *a* (whether this is assumed to be independent of the law operative in Greek or not); and the Gothic by means of Grimm's law (**d* > *t*, **k* > *h*) and the sound change **m* > *un*.

Phonological correspondences in cognate forms

Most 19th-century linguists took it for granted that they were reconstructing the actual word forms of some earlier language, that **dekm*, for example, was a pronounceable Proto-Indo-European word. Many of their successors have been more skeptical about the phonetic reality of reconstructed starred forms like **dekm*. They have said that they are no more than formulae summarizing the correspondences observed to hold between attested forms in particular languages and that they are, in principle, unpronounceable. From this point of view, it would be a matter of arbitrary decision which letter is used to refer to the correspondences: Latin *d* = Greek *d* = Sanskrit *d* = Gothic *t*, and so on. Any symbol would do, provided that a distinct symbol is used for each distinct set of correspondences. The difficulty with this view of reconstruction is that it seems to deny the very raison d'être of historical and comparative linguistics. Linguists want to know, if possible, not only that Latin *decem*, Greek *deka*, and so on are related,

but also the nature of their historical relationship—how they have developed from common ancestral form. They also wish to construct, if feasible, some general theory of sound change. This can be done only if some kind of phonetic interpretation can be given to the starred forms. The important point is that the confidence with which a phonetic interpretation is assigned to the phonemes that are reconstructed will vary from one phoneme to another. It should be clear from the discussion above, for example, that the interpretation of **d* as a voiced dental or alveolar stop is more certain than the interpretation of **k* as a voiceless velar stop. The starred forms are not all on an equal footing from a phonetic point of view.

Criticisms of the comparative method. One of the criticisms directed against the comparative method is that it is based upon a misleading genealogical metaphor. In the mid-19th century, the German linguist August Schleicher introduced into comparative linguistics the model of the "family tree." There is obviously no point in time at which it can be said that new languages are "born" of a common parent language. Nor is it normally the case that the parent language "lives on" for a while, relatively unchanged, and then "dies." It is easy enough to recognize the inappropriateness of these biological expressions. No less misleading, however, is the assumption that languages descended from the same parent language will necessarily diverge, never to converge again, through time. This assumption is built into the comparative method as it is traditionally applied. And yet there are many clear cases of convergence in the development of well-documented languages. The dialects of England are fast disappearing and are far more similar in grammar and vocabulary today than they were even a generation ago. They have been strongly influenced by the standard language. The same phenomenon, the replacement of nonstandard or less prestigious forms with forms borrowed from the standard language or dialect, has taken place in many different places at many different times. It would seem, therefore, that one must reckon with both divergence and convergence in the diachronic development of languages: divergence when contact between two speech communities is reduced or broken and convergence when the two speech communities remain in contact and when one is politically or culturally dominant.

Model of the family tree

The comparative method presupposes linguistically uniform speech communities and independent development after sudden, sharp cleavage. Critics of the comparative method have pointed out that this situation does not generally hold. In 1872 a German scholar, Johannes Schmidt, criticized the family-tree theory and proposed instead what is referred to as the wave theory, according to which different linguistic changes will spread, like waves, from a politically, commercially, or culturally important centre along the main lines of communication, but successive innovations will not necessarily cover exactly the same area. Consequently, there will be no sharp distinction between contiguous dialects, but, in general, the further apart two speech communities are, the more linguistic features there will be that distinguish them (see below *Dialectology and linguistic geography*).

Internal reconstruction. The comparative method is used to reconstruct earlier forms of a language by drawing upon the evidence provided by other related languages. It may be supplemented by what is called the method of internal reconstruction. This is based upon the existence of anomalous or irregular patterns of formation and the assumption that they must have developed, usually by sound change, from earlier regular patterns. For example, the existence of such patterns in early Latin as *honos* : *honoris* ("honor" : "of honor") and others in contrast with *orator* : *oratoris* ("orator" : "of the orator") and others might lead to the supposition that *honoris* developed from an earlier **honosis*. In this case, the evidence of other languages shows that **s* became *r* between vowels in an earlier period of Latin. But it would have been possible to reconstruct the earlier intervocalic **s* with a fair degree of confidence on the basis of the internal evidence alone. Clearly, internal reconstruction depends upon the structural approach to linguistics.

Use of irregular forms for reconstruction

The most recent development in the field of historical and comparative linguistics has come from the theory of generative grammar (see above *Transformational-generative grammar*). If the grammar and phonology of a language are described from a synchronic point of view as an integrated system of rules, then the grammatical and phonological similarities and differences between two closely related languages, or dialects, or between two diachronically distinct states of the same language can be described in terms of the similarities and differences in two descriptive rule systems. One system may contain a rule that the other lacks (or may restrict its application more or less narrowly); one system may differ from the other in that the same set of rules will apply in a different order in the one system from the order in which they apply in the other. Language change may thus be accounted for in terms of changes introduced into the underlying system of phonological and grammatical rules (including the addition, loss, or reordering of rules) during the process of language acquisition. So far these principles have been applied principally to sound change. There has also been a little work done on diachronic syntax.

LANGUAGE CLASSIFICATION

There are two kinds of classification of languages practiced in linguistics: genetic (or genealogical) and typological. The purpose of genetic classification is to group languages into families according to their degree of diachronic relatedness. For example, within the Indo-European family, such subfamilies as Germanic or Celtic are recognized; these subfamilies comprise German, English, Dutch, Swedish, Norwegian, Danish, and others, on the one hand, and Irish, Welsh, Breton, and others, on the other. So far, most of the languages of the world have been grouped only tentatively into families, and many of the classificatory schemes that have been proposed will no doubt be radically revised as further progress is made.

Typological classifications

A typological classification groups languages into types according to their structural characteristics. The most famous typological classification is probably that of isolating, agglutinating, and inflecting (or fusional) languages, which was frequently invoked in the 19th century in support of an evolutionary theory of language development. Roughly speaking, an isolating language is one in which all the words are morphologically unanalyzable (*i.e.,* in which each word is composed of a single morph); Chinese and, even more strikingly, Vietnamese are highly isolating. An agglutinating language (*e.g.,* Turkish) is one in which the word forms can be segmented into morphs, each of which represents a single grammatical category. An inflecting language is one in which there is no one-to-one correspondence between particular word segments and particular grammatical categories. The older Indo-European languages tend to be inflecting in this sense. For example, the Latin suffix *-is* represents the combination of categories "singular" and "genitive" in the word form *hominis* "of the man," but one part of the suffix cannot be assigned to "singular" and another to "genitive," and *-is* is only one of many suffixes that in different classes (or declensions) of words represent the combination of "singular" and "genitive."

There is, in principle, no limit to the variety of ways in which languages can be grouped typologically. One can distinguish languages with a relatively rich phonemic inventory from languages with a relatively poor phonemic inventory, languages with a high ratio of consonants to vowels from languages with a low ratio of consonants to vowels, languages with a fixed word order from languages with a free word order, prefixing languages from suffixing languages, and so on. The problem lies in deciding what significance should be attached to particular typological characteristics. Although there is, not surprisingly, a tendency for genetically related languages to be typologically similar in many ways, typological similarity of itself is no proof of genetic relationship. Nor does it appear true that languages of a particular type will be associated with cultures of a particular type or at a certain stage of development. What has emerged from recent work in typology is that certain logically unconnected features tend to occur together, so that the presence of feature A in a given

language will tend to imply the presence of feature B. The discovery of unexpected implications of this kind calls for an explanation and gives a stimulus to research in many branches of linguistics.

Linguistics and other disciplines

PSYCHOLINGUISTICS

The term psycholinguistics was coined in the 1940s and came into more general use after the publication of Charles E. Osgood and Thomas A. Sebeok's *Psycholinguistics: A Survey of Theory and Research Problems* (1954), which reported the proceedings of a seminar sponsored in the United States by the Social Science Research Council's Committee on Linguistics and Psychology.

The boundary between linguistics (in the narrower sense of the term: see the introduction of this article) and psycholinguistics is difficult, perhaps impossible, to draw. So too is the boundary between psycholinguistics and psychology. What characterizes psycholinguistics as it is practiced today as a more or less distinguishable field of research is its concentration upon a certain set of topics connected with language and its bringing to bear upon them the findings and theoretical principles of both linguistics and psychology. The range of topics that would be generally held to fall within the field of psycholinguistics nowadays is rather narrower, however, than that covered in the survey by Osgood and Sebeok.

Language acquisition by children. One of the topics most central to psycholinguistic research is the acquisition of language by children. The term acquisition is preferred to "learning," because "learning" tends to be used by psychologists in a narrowly technical sense, and many psycholinguists believe that no psychological theory of learning, as currently formulated, is capable of accounting for the process whereby children, in a relatively short time, come to achieve a fluent control of their native language. Since the beginning of the 1960s, research on language acquisition has been strongly influenced by Chomsky's theory of generative grammar, and the main problem to which it has addressed itself has been how it is possible for young children to infer the grammatical rules underlying the speech they hear and then to use these rules for the construction of utterances that they have never heard before. It is Chomsky's conviction, shared by a number of psycholinguists, that children are born with a knowledge of the formal principles that determine the grammatical structure of all languages, and that it is this innate knowledge that explains the success and speed of language acquisition. Others have argued that it is not grammatical competence as such that is innate but more general cognitive principles and that the application of these to language utterances in particular situations ultimately yields grammatical competence. Many recent works have stressed that all children go through the same stages of language development regardless of the language they are acquiring. It has also been asserted that the same basic semantic categories and grammatical functions can be found in the earliest speech of children in a number of different languages operating in quite different cultures in various parts of the world.

Theories of language acquisition

Although Chomsky was careful to stress in his earliest writings that generative grammar does not provide a model for the production or reception of language utterances, there has been a good deal of psycholinguistic research directed toward validating the psychological reality of the units and processes postulated by generative grammarians in their descriptions of languages. Experimental work in the early 1960s appeared to show that nonkernel sentences took longer to process than kernel sentences and, even more interestingly, that the processing time increased proportionately with the number of optional transformations involved. More recent work has cast doubt on these findings, and most psycholinguists are now more cautious about using grammars produced by linguists as models of language processing. Nevertheless, generative grammar continues to be a valuable source of psycholinguistic experimentation, and the formal properties of language, discovered or more adequately discussed by generative grammarians than they have been by others, are generally recognized to have important implications for the investigation of short-term and long-term memory and perceptual strategies.

Speech perception. Another important area of psycholinguistic research that has been strongly influenced by recent theoretical advances in linguistics and, more especially, by the development of generative grammar is speech perception. It has long been realized that the identification of speech sounds and of the word forms composed of them depends upon the context in which they occur and upon the hearer's having mastered, usually as a child, the appropriate phonological and grammatical system. Throughout the 1950s, work on speech perception was dominated (as was psycholinguistics in general) by information theory, according to which the occurrence of each sound in a word and each word in an utterance is statistically determined by the preceding sounds and words. Information theory is no longer as generally accepted as it was a few years ago, and more recent research has shown that in speech perception the cues provided by the acoustic input are interpreted, unconsciously and very rapidly, with reference not only to the phonological structure of the language but also to the more abstract levels of grammatical organization.

Information theory and speech perception

Other areas of research. Other areas of psycholinguistics that should be briefly mentioned are the study of aphasia and neurolinguistics. The term aphasia is used to refer to various kinds of language disorders; recent work has sought to relate these, on the one hand, to particular kinds of brain injury and, on the other, to psychological theories of the storage and processing of different kinds of linguistic information. One linguist has put forward the theory that the most basic distinctions in language are those that are acquired first by children and are subsequently most resistant to disruption and loss in aphasia. This, though not disproved, is still regarded as controversial. Two kinds of aphasia are commonly distinguished. In motor aphasia the patient manifests difficulty in the articulation of speech or in writing and may produce utterances with a simplified grammatical structure, but his comprehension is not affected. In sensory aphasia the patient's fluency may be unaffected, but his comprehension will be impaired and his utterances will often be incoherent.

Neurolinguistics should perhaps be regarded as an independent field of research rather than as part of psycholinguistics. In 1864 it was shown that motor aphasia is produced by lesions in the third frontal convolution of the left hemisphere of the brain. Shortly after the connection had been established between motor aphasia and damage to this area (known as Broca's area), the source of sensory aphasia was localized in lesions of the posterior part of the left temporal lobe. More recent work has confirmed these findings. The technique of electrically stimulating the cortex in conscious patients has enabled brain surgeons to induce temporary aphasia and so to identify a "speech area" in the brain. It is no longer generally believed that there are highly specialized "centres" within the speech area, each with its own particular function; but the existence of such a speech area in the dominant hemisphere of the brain (which for most people is the left hemisphere) seems to be well established. The posterior part of this area is involved more in the comprehension of speech and the construction of grammatically and semantically coherent utterances, and the anterior part is concerned with the articulation of speech and with writing. Little is yet known about the operation of the neurological mechanisms underlying the storage and processing of language. (See also the articles entitled PERCEPTION; SPEECH.)

Speech area in the brain

SOCIOLINGUISTICS

Delineation of the field. Just as it is difficult to draw the boundary between linguistics and psycholinguistics and between psychology and psycholinguistics, so it is difficult to distinguish sharply between linguistics and sociolinguistics and between sociolinguistics and sociology. There is the further difficulty that, because the boundary between sociology and anthropology is also unclear, sociolinguistics merges with anthropological linguistics (see below).

It is frequently suggested that there is a conflict between the sociolinguistic and the psycholinguistic approach to the study of language, and it is certainly the case that two distinct points of view are discernible in the literature at the present time. Chomsky has described linguistics as a branch of cognitive psychology, and neither he nor most of his followers have yet shown much interest in the relationship between language and its social and cultural matrix. On the other hand, many modern schools of linguistics that have been very much concerned with the role of language in society would tend to relate linguistics more closely to sociology and anthropology than to any other discipline. It would seem that the opposition between the psycholinguistic and the sociolinguistic viewpoint must ultimately be transcended. The acquisition of language, a topic of central concern to psycholinguists, is in part dependent upon and in part itself determines the process of socialization; and the ability to use one's native language correctly in the numerous socially prescribed situations of daily life is as characteristic a feature of linguistic competence, in the broad sense of this term, as is the ability to produce grammatical utterances. Some of the most recent work in sociolinguistics and psycholinguistics has sought to widen the notion of linguistic competence in this way. So far, however, sociolinguistics and psycholinguistics tend to be regarded as relatively independent areas of research.

Language and socialization

Social dimensions. Language is probably the most important instrument of socialization that exists in all human societies and cultures. It is largely by means of language that one generation passes on to the next its myths, laws, customs, and beliefs, and it is largely by means of language that the child comes to appreciate the structure of the society into which he is born and his own place in that society.

As a social force, language serves both to strengthen the links that bind the members of the same group and to differentiate the members of one group from those of another. In many countries there are social dialects as well as regional dialects, so that it is possible to tell from a person's speech not only where he comes from but what class he belongs to. In some instances social dialects can transcend regional dialects. This is notable in England, where standard English in the so-called Received Pronunciation (RP) can be heard from members of the upper class and upper middle class in all parts of the country. The example of England is but an extreme manifestation of a tendency that is found in all countries: there is less regional variation in the speech of the higher than in that of the lower socioeconomic classes. In Britain and the United States and in most of the other English-speaking countries, people will almost always use the same dialect, regional or social, however formal or informal the situation and regardless of whether their listeners speak the same dialect or not. (Relatively minor adjustments of vocabulary may, however, be made: an Englishman speaking to an American may employ the word "elevator" rather than "lift" and so on.) In many communities throughout the world, it is common for members to speak two or more different dialects and to use one dialect rather than another in particular social situations. This is commonly referred to as code-switching. Code-switching may operate between two distinct languages (*e.g.*, Spanish and English among Puerto Ricans in New York) as well as between two dialects of the same language. The term diglossia (rather than bilingualism) is frequently used by sociolinguists to refer to this by no means uncommon phenomenon.

Variables influencing language usage

In every situation, what one says and how one says it depends upon the nature of that situation, the social role being played at the time, one's status vis-à-vis that of the person addressed, one's attitude towards him, and so on. Language interacts with nonverbal behaviour in social situations and serves to clarify and reinforce the various roles and relationships important in a particular culture. Sociolinguistics is far from having satisfactorily analyzed or even identified all the factors involved in the selection of one language feature rather than another in particular situations. Among those that have been discussed in relation to various languages are: the formality or informality of the situation; power and solidarity relationships between the participants; differences of sex, age, occupation, socioeconomic class, and educational background; and personal or transactional situations. Terms such as style and register (as well as a variety of others) are employed by many linguists to refer to the socially relevant dimensions of phonological, grammatical, and lexical variation within one language. So far there is very little agreement as to the precise application of such terms. (For further treatment of sociolinguistics, see the section *Dialects* in the article LANGUAGE.)

OTHER RELATIONSHIPS

Anthropological linguistics. The fundamental concern of anthropological linguistics is to investigate the relationship between language and culture. To what extent the structure of a particular language is determined by or determines the form and content of the culture with which it is associated remains a controversial question. Vocabulary differences between languages correlate obviously enough with cultural differences, but even here the interdependence of language and culture is not so strong that one can argue from the presence or absence of a corresponding cultural difference. For example, from the fact that English—unlike French, German, Russian, and many other languages—distinguishes lexically between monkeys and apes, one cannot conclude that there is an associated difference in the cultural significance attached to these animals by English-speaking societies. Some of the major grammatical distinctions in certain languages may have originated in culturally important categories (*e.g.*, the distinction between an animate and an inanimate gender). But they seem to endure independently of any continuing cultural significance. The "Whorfian hypothesis" (the thesis that one's thought and even perception are determined by the language one happens to speak), in its strong form at least, is no longer debated as vigorously as it was a few years ago. Anthropologists continue to draw upon linguistics for the assistance it can give them in the analysis of such topics as the structure of kinship. A more recent development, but one that has not so far produced any very substantial results, is the application of notions derived from generative grammar to the analysis of ritual and other kinds of culturally prescribed behaviour.

The "Whorfian hypothesis"

Use of computers in linguistic research

Computational linguistics. By computational linguistics is meant no more than the use of electronic digital computers in linguistic research. At a theoretically trivial level, computers are employed to scan texts and to produce, more rapidly and more reliably than was possible in the past, such valuable aids to linguistic and stylistic research as word lists, frequency counts, and concordances. Theoretically more interesting, though much more difficult, is the automatic grammatical analysis of texts by computer. Considerable progress was made in this area by research groups working on machine translation and information retrieval in the United States, Great Britain, the Soviet Union, France, and a few other countries in the decade between the mid-1950s and the mid-1960s. But much of the original impetus for this work disappeared, for a time at least, in part because of the realization that the theoretical problems involved in machine translation are much more difficult than they were at first thought to be and in part as a consequence of a loss of interest among linguists in the development of discovery procedures. Whether automatic syntactic analysis and fully automatic high-quality machine translation are even feasible in principle remains a controversial question.

Mathematical linguistics. What is commonly referred to as mathematical linguistics comprises two areas of research: the study of the statistical structure of texts and the construction of mathematical models of the phonological and grammatical structure of languages. These two branches of mathematical linguistics, which may be termed statistical and algebraic linguistics, respectively, are typically distinct. Attempts have been made to derive the grammatical rules of languages from the statistical structure of texts written in those languages, but such attempts are generally thought to have been not only unsuccessful so far in practice but also, in principle, doomed to failure. That languages have a statistical structure is a fact well

known to cryptographers. Within linguistics, it is of considerable typological interest to compare languages from a statistical point of view (the ratio of consonants to vowels, of nouns to verbs, and so on). Statistical considerations are also of value in stylistics (see below).

Algebraic linguistics

Algebraic linguistics derives principally from the work of Noam Chomsky in the field of generative grammar (see above *Chomsky's grammar*). In his earliest work Chomsky described three different models of grammar—finite-state grammar, phrase-structure grammar, and transformational grammar—and compared them in terms of their capacity to generate all and only the sentences of natural languages and, in doing so, to reflect in an intuitively satisfying manner the underlying formal principles and processes. Other models have also been investigated, and it has been shown that certain different models are equivalent in generative power to phrase-structure grammars. The problem is to construct a model that has all the formal properties required to handle the processes found to be operative in languages but that prohibits rules that are not required for linguistic description. It is an open question whether such a model, or one that approximates more closely to this ideal than current models do, will be a transformational grammar or a grammar of some radically different character.

Stylistics. The term stylistics is employed in a variety of senses by different linguists. In its widest interpretation it is understood to deal with every kind of synchronic variation in language other than what can be ascribed to differences of regional dialect. At its narrowest interpretation it refers to the linguistic analysis of literary texts. One of the aims of stylistics in this sense is to identify those features of a text that give it its individual stamp and mark it as the work of a particular author. Another is to identify the linguistic features of the text that produce a certain aesthetic response in the reader. The aims of stylistics are the traditional aims of literary criticism. What distinguishes stylistics as a branch of linguistics (for those who regard it as such) is the fact that it draws upon the methodological and theoretical principles of modern linguistics.

Philosophy of language. The analysis of language has always been a subject of particular concern to philosophers, and traditional grammar was strongly influenced by the dominant philosophical attitudes of the day. Modern linguistics and modern philosophical theories have so far had little influence on one another. Some philosophers have shown an interest in Chomsky's controversial suggestion that work in generative grammar lends support to the rationalists in their long-standing dispute about the source of human knowledge. Potentially more fruitful, perhaps, is the interest shown by a number of linguists in philosophical treatments of reference, quantification, and presupposition, in systems of modal logic, and in the work of the so-called philosophers of ordinary language.

Applied linguistics. In the sense in which the term applied linguistics is most commonly used nowadays it

Language teaching

is restricted to the application of linguistics to language teaching. Much of the recent expansion of linguistics as a subject of teaching and research in the universities in many countries has come about because of its value, actual and potential, for writing better language textbooks and devising more efficient methods of teaching languages. Linguistics is also widely held to be relevant to the training of teachers of the deaf and speech therapists. Outside the field of education in the narrower sense, applied linguistics (and, more particularly, applied sociolinguistics) has an important part to play in what is called language planning; *i.e.*, in advising governments, especially in recent created states, as to which language or dialect should be made the official language of the country and how it should be standardized. (J.Lyo./Ed.)

Dialectology and linguistic geography

DIALECT GEOGRAPHY

Dialect study as a discipline—dialectology—dates from the first half of the 19th century, when local dialect dictionaries and dialect grammars first appeared in western Europe. Soon thereafter, dialect maps were developed; most often they depicted the division of a language's territory into regional dialects. The 19th-century rise of nationalism, coupled with the Romantic view of dialects and folklore as manifestations of the ethnic soul, furnished a great impetus for dialectology.

19th-century impetus for dialect study

Early dialect studies. The first dialect dictionaries and grammars were most often written by scholars describing the dialect of their birthplace or by fieldworkers whose main method of investigation was free conversation with speakers of the dialect, usually older persons and, preferably, those who showed the least degree of literacy and who had travelled as little as possible. Many of these grammars and dictionaries recorded dialectal traits that deviated from the standard language. In the second half of the 19th century, when historical and comparative linguistic study was flourishing, it became customary to focus attention on the fate of particular elements of the archaic language in a given dialect; *e.g.*, the changes that Latin vowels and consonants underwent when used in different positions in a particular Romance dialect.

With the accumulation of dialectal data, investigators became increasingly conscious of the inadequacy of viewing dialects as internally consistent units that were sharply differentiated from neighbouring dialects. It became more and more clear that each dialectal element or phenomenon refused to stay neatly within the borders of a single dialect area and that each had its own isogloss; consequently, maps of dialects would have to be replaced by maps showing the distribution of each particular feature. While sound scientifically, the preparation and compilation of such maps, called linguistic atlases, is a difficult, costly, voluminous, and time-consuming job.

Dialect atlases. Dialect atlases are compiled on the basis of investigations of the dialects of a large number of places; a questionnaire provides uniform data. There are two basic methods of data collection: fieldwork and survey by correspondence. Fieldwork, in which a trained investigator transcribes dialectal forms directly (or on tape), affords more precise data and enables the questionnaire to include a greater number of diverse questions; but it implies a necessarily limited number of points to be covered. The advantage of the correspondence method lies in its ability to encompass more points at less cost and with less time expended in gathering the data. On the other hand, rural schoolteachers, normally the persons who complete such questionnaires, can answer only a relatively small number of questions and often imperfectly.

The first large-scale enterprise in linguistic geography was the preparation of the German linguistic atlas. In the 1880s, the initiator of this great undertaking, Georg Wenker, composed 40 test sentences that illustrated most of the important ways in which dialects differed and sent them to schoolmasters in over 40,000 places in the German Empire. The sentences were to be translated into the local dialect. Publication of the results was not begun until 1926; the main cause of the delay was the enormous quantity of material to be arranged and analyzed.

The French atlas

The famous French linguistic atlas of Jules Gilliéron and Edmond Edmont was based on a completely different concept. Using a questionnaire of about 2,000 words and phrases that Gilliéron had composed, Edmont surveyed 639 points in the French-speaking area. The atlas, compiled under the direction of Gilliéron, was published in fascicles from 1902 to 1912 and furnished both a strong stimulus and the basic model for work on linguistic atlases elsewhere in the world. European linguists, especially in Romance- and Germanic-speaking countries, were the first to participate in such atlas projects. One of the most significant contributions is the linguistic atlas of Italy and southern Switzerland by Karl Jaberg and Jakob Jud; it appeared from 1928 to 1940. Particularly noteworthy in its attention to precise definitions of meaning, this atlas often used illustrations and described objects and actions of village life denoted by the questionnaire's words.

At present, dialects of virtually all European languages have been treated in linguistic geography studies. In some countries, data are still being collected and classified and maps are being drawn, but in others a second generation

of atlases is already underway. French dialectologists, for instance, are now working on regional atlases that will complement data contained in the *Atlas linguistique de la France.* In England, work began in 1946, under the direction of Harold Orton and Eugene Dieth; the first volume of the *Survey of English Dialects* was published in 1962. In Slavic-speaking countries, work is now underway both on atlases of separate Slavic languages and on the large general Slavic linguistic atlas that will cover nearly 1,000 locales in all parts of European territory where Slavic languages are spoken. Outside Europe, the greatest amount of work in linguistic geography has been completed in Japan and in the United States.

As early as 1905–06, a committee of Japanese dialectologists published the first linguistic atlas of Japan in two volumes, one devoted to phonology and one to morphology. Subsequent work has been done on a new atlas of Japan as a whole and on several regional atlases. The extensive activity of Chinese specialists has concentrated on descriptions of particular local and regional dialects. The Chinese situation is a peculiar one because of the enormous number of people who speak Chinese, the very significant dialectal differentiation (certain dialects, particularly those in the South of China, would be considered by Western standards as separate languages), and the nature of the Chinese script. Chinese characters do not represent sounds but concepts. Because of this, the written language can be read without difficulty in many different dialect areas, although its spoken form varies greatly from one region to another.

Because of the enormous size of the U.S., atlas surveys were done by region. Between 1931 and 1933, fieldworkers under the direction of the linguist Hans Kurath surveyed 213 New England communities; the results were published in the *Linguistic Atlas of New England* (with 734 maps) in 1939–43. Based on the methodological experience of Jaberg and Jud in their atlas of Italy and southern Switzerland, this work involved systematic investigations not only among the relatively uneducated but also among better educated, more cultured informants and among the very well educated, cultured, and informed members of a community. Thus the dimension of social stratification of language was introduced into linguistic geography, and valuable material about regional linguistic standards became available.

After 1933, fieldwork was extended to the other Atlantic states. Lack of financial support, however, has hindered the publication of these atlases. Nevertheless, several works based on the material gathered have appeared, among them Kurath's *Word Geography of the Eastern United States,* E. Bagby Atwood's *Survey of Verb Forms in the Eastern United States,* and Kurath's and McDavid's *Pronunciation of English in the Atlantic States.* Independent work was carried out in other U.S. regions, mainly with an adapted form of the questionnaire developed for the Atlantic states; only introductions or summaries of material in the files have been published, however, because of lack of funds.

Works on American English

The most effective and thorough—as well as the most expensive—way of presenting data in linguistic atlases is by printing the actual responses to questionnaire items right on the maps. Phenomena of linguistic geography, however, are usually represented by geometric symbols or figures at the proper points on the map, or, even more summarily, by the drawing of isoglosses (linguistic boundaries) or by shading or colouring the areas of particular features.

Only dialect atlases can furnish the complexity of data of the major dialectal phenomena in a multitude of geographic locations in a manner that both assures commensurability of the data and allows a panoramic examination of the whole gamut of data. The inventory of linguistic phenomena is so rich, however, that no one questionnaire can encompass it all. Moreover, the use of a questionnaire unavoidably brings about a schematization of answers that is lacking in spontaneity. For these reasons, other kinds of publications, such as dialect dictionaries or monographs based on extensive free conversation with speakers of local dialects, are indispensable complements to linguistic atlases.

The value and applications of dialectology. The scientific interest of dialectology lies in the fact that dialects are a valuable source of information about popular culture. They reflect not only the history of a language but, to a great extent, the ethnic, cultural, and even political history of a people as well. A knowledge of dialectal facts provides practical guidance to school systems that are trying to teach the standard language to an ever greater number of pupils.

In the 1930s, the value of dialectology to the study of language types became apparent. Because dialects greatly outnumber standard languages, they provide a much greater variety of phenomena than languages and thus have become the main source of information about the types of phenomena possible in linguistic systems. Also, in some languages, but not in others, an extremely wide structural variation among dialects has been found. In Yugoslavia, where two closely related Slavic languages, Serbo-Croatian and Slovene, are spoken, dialects are found with synthetic declension (case endings, as in Latin) and analytic declension (use of prepositions and word order, as in English). In addition, there are among these dialects complex systems of verbal tenses contrasting with simple ones, as well as dialects with or without the dual number or the neuter gender. The dialects of Serbo-Croatian and Slovene also exhibit almost every type of prosodic structure (*e.g.,* tone, stress, length) found in European languages. Some dialects differentiate long and short vowels or rising and falling accents, while others do not; and in some, but not all of them, stress fulfills a grammatical function. Of the several dozen vowel and diphthong sounds that occur in these dialects, only five are common to all of them; all the rest are restricted to relatively small areas. All of this rich variety contrasts sharply with the relative structural uniformity of the English language—not only in the U.S. but wherever it is spoken. (The outstanding exceptions are the creolized dialects, which are distinguished by far-reaching structural peculiarities.)

Variations in Serbo-Croatian and Slovene dialects

SOCIAL DIALECTOLOGY

The methodology of generative grammar was first applied to dialectology in the 1960s, when the use of statistical means to measure the degree of similarity or difference between dialects also became increasingly common. The most important development of that time, however, was the rapid growth of methods for investigating the social variations of dialects; this type of variation, in contrast to geographical variation, is especially great in the U.S., above all in large urban centres. In cities such as New York, a whole scale of speech variations can be found to correlate with the social status and educational levels of the speakers. In addition, age groups exhibit different patterns, but such age-group patterns of variation differ from one social stratum to another. Still another dimension of variation, especially important in the U.S., is connected with the race and the ethnic origin of a speaker as well as his date of immigration. So-called Black English has been influenced by the fact that most of the black population in non-southern U.S. regions originally came from the southeastern states; thus, many Black English peculiarities are in reality transplanted southeastern dialectal traits.

Social correlations with speech variation

Normally, speakers of one of the social dialects of a city possess at least some awareness of the other dialects. In this way, speech characteristics also become subjectively integrated into the system of signs indicating social status. And in seeking to enhance their social status, poorer and less educated speakers may try to acquire the dialect of the socially prestigious. Certain groups—*e.g.,* blacks and the working class—however, will, under certain conditions, show a consciousness of solidarity and a tendency to reject members who imitate either the speech or other types of behaviour of models outside their own social group.

As a consequence of an individual's daily contacts with speakers of the various social dialects of a city, elements of the other dialects are imperceptibly drawn into his dialect. The collective result of such experiences is the spread of linguistic variables; *i.e.,* groups of variants (sounds or grammatical phenomena) primarily determined by social (educational, racial, age, class) influences, an example be-

ing the existence of the two forms "He don't know" and the standard "He doesn't know." Traits representing variables in intergroup relations can become variable features in the speech of individuals as well; *i.e.,* an individual may employ two or more variants for the same feature in his own speech, such as "seeing" and "seein' " or "he don't" and "he doesn't." The frequency of usage for each variable varies with the individual speaker as well as with the social group. There are intermediate stages of frequency between different social groups and entire scales of transitions between different age groups, thus creating even greater variation within the dialect of an individual. The variables also behave differently in the various styles of written or spoken language used by each speaker.

The study of variables is one of the central tasks of any investigation of the dialects of American cities. Applying the statistical methods of modern sociology, linguists have worked out investigative procedures sharply different from those of traditional dialectology. The chief contributor has been William Labov, the pioneer of social dialectology in the U.S. The basic task is to determine the correlation between a group of linguistic variables—such as the different ways of pronouncing a certain vowel—and extralinguistic variables, such as education, social status, age, and race. For a reasonable degree of statistical reliability, one must record a great number of speakers. In general, several examples of the same variable must be elicited from each individual in order to examine the frequency and probability of its usage. Accordingly, the number of linguistic variables that can be examined is quite limited, in comparison with the number of dialectal features normally recorded by traditional fieldworkers in rural communities; in these situations, the investigator is often satisfied with one or two responses for each feature.

Application of modern statistical methods to dialectology

A completely new, flexible, and imaginative method of interviewing is needed for such work in urban centres, as well as new ways of finding and making contact with informants. One example is Labov's method for testing the fate of final and preconsonantal *r* in speakers of different social levels. Choosing three New York City department stores, each oriented to a completely different social stratum, he approached a large number of salesladies, asking each of them about the location of a certain department that he knew to be on the fourth floor. Thus, their answers always contained two words with potential *r*'s—"fourth" and "floor." This shortcut enabled Labov to establish in a relatively short time that the salesladies in the store with richer customers clearly tended to use "*r*-full" forms, whereas those in the stores geared to the poorer social strata more commonly used "*r*-less" forms.

Social dialectology has focused on the subjective evaluation of linguistic features and the degree of an individual's linguistic security, phenomena that have considerable influence on linguistic change. Linguistic scientists, in studying the mechanism of such change, have found that it seems to proceed gradually from one social group to another, always attaining greater frequency among the young. Social dialectology also has great relevance for a society as a whole, in that the data it furnishes will help deal with the extremely complex problems connected with the speech of the socially underprivileged, especially of minority groups. Thus, the recent emphasis on the speech of minority groups, such as the Black English of American cities, is not a chance phenomenon. Specific methods for such investigation are being developed, as well as ways of applying the results of such investigation to educational policies. (P.I./Ed.)

Semantics

Semantics is the philosophical and scientific study of meaning. The term is one of a group of English words formed from the various derivatives of the Greek verb *sēmainō* ("to mean" or "to signify"). The noun semantics and the adjective semantic are derived from *sēmantikos* ("significant"); semiotic (adjective and noun) comes from *sēmeiōtikos* ("pertaining to signs"); semology from *sēma* ("sign") + *logos* ("account"); and semasiology from *sēmasia* ("signification") + *logos* ("account"). It is difficult to formulate a distinct definition for each of these terms because their use largely overlaps in the literature despite individual preferences. Semantics is a relatively new field of study, and its originators, often working independently of one another, felt the need to coin a new name for the new discipline; hence the variety of terms denoting the same subject. The word semantics has ultimately prevailed as a name for the doctrine of meaning, in particular, of linguistic meaning. Semiotic is still used, however, to denote a broader field: the study of sign-using behaviour in general.

MODERN DEVELOPMENT OF SEMANTICS

The concern with meaning, always present for philosophers and linguists, greatly increased in the decades following World War II. The sudden rise of interest in meaning can be attributed to an interaction of several lines of development in various disciplines. From the middle of the 19th century onward, logic, the formal study of reasoning, underwent a period of growth unparalleled since the time of Aristotle. Although the main motivation for a renewed interest in logic was a search for the foundations of mathematics, the chief protagonists of this effort—notably the German mathematician Gottlob Frege and the English philosopher Bertrand Russell—extended their inquiry into the domain of the natural languages, which are the original media of human reasoning. The influence of mathematical thinking, and of mathematical logic in particular, however, left a permanent mark on the subsequent study of semantics.

Semantic theories of the Logical Positivists

Positivist theory. This mark is nowhere more obvious than in the semantic theories offered by the Neopositivists of the Vienna Circle, which flourished in the 1920s and 1930s, and which was composed of philosophers, mathematicians, and scientists who discussed the methodology and epistemology of science. To such "logical" Positivists as the German-born philosopher Rudolf Carnap, for instance, the symbolism of modern logic represented the grammar (syntax) of an "ideal" language. Because the Logical Positivists were, at the same time, radical Empiricists (observationalists) in their philosophy, the semantics of their ideal language has been given in terms of a tie connecting the symbols of this language with observable entities in the world, or the data of one's sense experience, or both. Against such a rigid ideal as logic, natural language appeared to these philosophers as something primitive, vague, inaccurate, and confused. Moreover, since a large part of ordinary and philosophical discourse, particularly that concerning metaphysical and moral issues, could not be captured by the ideal language, the Positivistic approach provided a way to brand all such talk as nonsensical, or at least as "cognitively" meaningless. Accordingly, the Positivists engaged in a prolonged, and largely unsuccessful, effort to formulate a criterion of meaningfulness in terms of empirical verifiability with respect to the sentences formed in natural language.

Whorfian views. Another source of dissatisfaction with the vernacular was made apparent shortly before World War II by the work of the American anthropological linguist Benjamin Lee Whorf. Whorf's famous thesis of linguistic relativity implied that the particular language a person learns and uses determines the framework of his perception and thought. If that language is vague and inaccurate, as the Positivists suggested, or is burdened with the prejudices and superstitions of an ignorant past, as some cultural anthropologists averred, then it is bound to render the user's thinking—and his mental life itself—confused, prejudiced, and superstitious. The Polish-American semanticist Alfred Korzybski, the founder of the movement called General Semantics, believed that the cure for such vague and superstition-laden language lay in a radical revision of linguistic habits in the light of modern science.

School of natural language. Natural language did not remain without champions in the face of this combined onslaught from the logicians and the Whorfians. A reaction started in England, first in Cambridge, then in Oxford. Influenced by the English philosopher George Edward Moore but more so by the "converted" Vienna-born Positivist Ludwig Wittgenstein, the philosophy of "ordinary

language" (also known as the Oxford school) came into its own in the 1940s. According to the philosophers of this group, natural language, far from being the crude instrument the Positivists alleged it to be, provides the basic and unavoidable matrix of all thought, including philosophical reflections. Any "ideal" language, therefore, can make sense only as a parasitical extension of, and never as a substitute for, the natural language. Philosophical problems arise as a result of a failure to see the workings of man's language; they are bound to "dissolve" with improved understanding. These assumptions provided a mighty impetus to reflect upon the vernacular language, including its minute points of grammar and fine nuances of meaning. Indeed, some of the later representatives of this approach, particularly the English philosopher John L. Austin, became renowned as much among linguists as among philosophers.

Modern grammatical influences. In the 1950s the science of linguistics itself rose to the challenges that had been coming chiefly from philosophical quarters. The development of transformational, or generative, grammar, initiated by the work of the U.S. linguists Zellig S. Harris and Noam Chomsky, opened a deeper insight into the syntax of the natural languages. Instead of merely providing a structural description (parsing) of sentences, this approach demonstrates how sentences are built up, step by step, from some basic ingredients. In the hands of the philosopher, this powerful new grammar not only served to counter the positivistic charge of imprecision laid against natural language but aided him in his own work of conceptual clarification. Moreover, the generative approach promised further results: since the late 1960s some steps have been taken to develop a generative semantics for natural languages, in addition to a generative syntax.

Influence of transformational grammar

PHILOSOPHICAL VIEWS ON MEANING

Meaning and reference. On a rather unsophisticated level the problem of meaning can be approached through the following steps. The perception of certain physical entities (objects, marks, sounds, and so on) might lead an intelligent being to the thought of another thing with some regularity. For example, the sight of smoke evokes the idea of fire, footprints on the sand makes one think of the man who must have passed by. The smoke and the footprints are thus signs of something else. They are natural signs, inasmuch as the connection between the sign and the thing signified is a causal link, established by nature and learned from experience. These can be compared with road signs, for example, or such symbols as the outline of a heart pierced by an arrow. The connection between the symbol and the thing signified in these cases is not a natural one; it is established by human tradition or convention and is learned from these sources. These nonnatural signs, or symbols, are widely used in human communication.

In this framework the elements of language appear to be nonnatural signs. The interest in words and phrases reaches beyond their physical appearance: their perception is likely to direct attention or thought to something else. Words, in fact, are the chief media of human communication, and, as the diversity of languages clearly shows, the link involved between words and what they signify cannot be a natural one. Words and sentences are like symbols: they point beyond themselves; they mean something. Smoke means fire, the pierced heart means love. Words mean the thing they make us think of; the meaning of the word is the tie that connects it with that thing.

Words as symbols

There are some words for which this approach seems to work very straightforwardly. The name Paris means (signifies, stands for, refers to, denotes) the city of Paris, the name Aristotle means that philosopher, and so forth. The initial plausibility of such examples created an obsession in the minds of many thinkers, beginning with Plato. Regarding proper names as words par excellence, they tried to extend the referential model of meaning to all of the other classes of words and phrases. Plato's theory of "forms" may be viewed as an attempt to find a referent for such common nouns as "dog" or for abstract nouns like "whiteness" or "justice." As the word Socrates in the sentence "Socrates is wise" refers to Socrates, for example, so the word wise refers to the form of wisdom. Unfortunately, whereas Socrates was a real person in this world, the form of wisdom is not something to be encountered anywhere, at any time, in the world. The difficulty represented by "Platonic" entities of this kind increases as one tries to find appropriate referents for verbs, prepositions, connectives, and so forth. Discussion of abstract entities such as classes (*e.g.,* the class of all running things) and relations (*e.g.,* the relation of being greater than . . .) abound in philosophical literature; Gottlob Frege even postulated "the True" and "the False" as referents for complete propositions.

There are many more serious problems besetting the referential theory of meaning. The first one, eloquently pointed out by Frege, is that two expressions may have the same referent without having the same meaning. For example, "the Morning Star" and "the Evening Star" denote the same planet, yet, clearly, the two phrases do not have the same meaning. If they had, then the identity of the Morning Star and the Evening Star would be as obvious to anybody who understands these phrases as the identity of a vixen with a female fox or a bachelor with an unmarried man is obvious to speakers of English. As it is, the identity of the Morning Star with the Evening Star is a scientific and not a linguistic matter. Thus, even in the case of names, or expressions equivalent to names, one has to distinguish between the denotation (reference, extension) of the name—*i.e.,* the object (or group of objects) it refers to—and its connotation (sense, intension)—*i.e.,* its meaning.

Denotation and connotation

The second problem with the theory of referential meaning arises from phrases that, though meaningful, pretend to refer but, in fact, do not. For example, in the case of such a definite description as "the present king of France," the phrase is meaningful although there is no such person. If the phrase were not meaningful, one would not even know that the phrase has no actual referent. Russell's analysis of these phrases, and the U.S. philosopher Willard V. Quine's similar treatment of such names as Cerberus, effectively detached meaning from reference by claiming that these expressions, when used in sentences, are equivalent to a set of existential propositions; *i.e.,* propositions without definite reference. For example, "The present king of France is bald" comes out as "There is at least, and at most, one person that rules over France, and whoever rules over France is bald." These propositions are meaningful, true or false, without definite reference.

Names, in fact, are very untypical words. The name of the third Secretary General of the United Nations, U Thant, has no meaning in English. Whether it means anything in Burmese does not matter either; the reference is not affected by the meaning or the lack of meaning of the name. Names, as such, do not belong to the vocabulary of a language; most dictionaries do not list them. Thus, in spite of the initial plausibility, the idea of reference does not help in understanding the nature of linguistic meaning.

Meaning and truth. Despite the failure of referential meaning, many philosophers were quite unwilling to give up the idea that the meaning of linguistic expressions has something to do with objects, events, and states of affairs in the world. They reasoned that if language is used to talk about the physical environment, then there must be some connection between man's words and the things around him. If reference fails to provide the link, something else must.

In the face of referential failure Russell fell back on truth. The Positivists suggested verifiability as the criterion of empirical meaning. Indeed, at least in many cases, it stands to reason that to understand a sentence is to know what state of affairs would make it true or false. Such considerations motivate the aletheic (Greek *alētheia,* "truth") semantic theories, which claim that the notion of meaning is best explained in terms of truth rather than reference.

The most influential discussion of the notion of truth was offered by the Polish-born mathematician and logician Alfred Tarski in the 1930s. His semantic definition of truth is contained in the following formula (which he called [T]):

Tarski's semantic definition of truth

(T) X is true if, and only if, p

in which "p" is a variable representing any sentence and "X" is a variable representing the name, or unique de-

scription, of that sentence. The easiest way to obtain such a unique description is to put the sentence in quotation marks. Thus, we get such instances of (T) as

"Snow is white" is true if, and only if, snow is white.

The above formula implies a distinction between the object language and the metalanguage. "*X*" represents the name of a sentence in the object language—*i.e.,* roughly, the language used to talk about things in the world. The instances of (T) themselves, however, are in the metalanguage, the language in which one can talk about both things in the world and sentences of the object language. Tarski claimed that no language can contain its own truth predicate, for if it did then it would permit the formation of such sentences as:

(S) This very sentence is false.

Is the sentence (S) true or false? Clearly, it is true if, and only if, it is false, which is an intolerable paradox. Consequently, for any language the predicate ". . . is true" and other semantical predicates must belong to a language of a higher order (a metalanguage).

For this reason Tarski restricted his theory to clearly formalized, artificial languages, a decision that was very much in line with the positivistic tendencies of the 1930s. Nevertheless, the Tarski formula remained attractive even to some semanticists concerned with meaning in natural languages. For one thing, it seemed to succeed in pairing linguistic entities (named by the values of "*X*"; *e.g.,* the sentence "Snow is white") and nonlinguistic entities (named by the values of "*p*"; *e.g.,* the fact, or possible state of affairs, that snow is white). This correlation, however, is not very helpful because each one of the nearly infinite number of sentences one may form would have its "fact" as a counterpart, identifiable only by means of that very same sentence. Consequently, if linguistic meaning consisted in these correlations, no one could learn the meaning of any sentence at all and certainly not the meaning of all the sentences the speakers of a language are able to understand.

What was needed was a theory explaining the contribution of individual words—a clearly finite set—to the truth of sentences. Tarski himself, as well as other writers, suggested a repeatable procedure based on the notion of satisfaction. Snow, for example, satisfies the sentential function "*x* is white" because "Snow is white" is true. In much the same way, 3 satisfies the function "$2 \cdot x = 6$" because "$2 \cdot 3 = 6$" is true. Simply stated, the meaning of the predicate ". . . is white" is determined—and is learned—in terms of the set of objects of which it is true.

Limitations to the "meaning as truth" theory

As this approach is extended to cover the wide variety of words that exist in a natural language, however, its initial simplicity—and thereby its attractiveness—becomes progressively lost. This can be illustrated by "egocentric" words like "I," "you," "here," and "now"; by connectives like "since," "however," and "nevertheless"; and, if these appear trivial, by such crucial words as "believe," "know," and "intend," on the one hand, and "good" and "beautiful" on the other. Whereas it is plausible to say that, for instance, Joe and Mary satisfy the function "*X* loves *Y*," provided Joe loves Mary, it is more complicated to determine what would satisfy such functions as "*X* believes *Y*," "*X* knows *Y*," or "*X* intends *Y*." "*X*" is satisfiable by people, but the satisfaction of "*Y*" poses a problem. If one suggests such things as propositions, facts, and possibilities, one is confronted with abstract entities of a kind similar to those encountered in the Platonic theory of referential meaning. Again, can it be said that John and a unicorn will satisfy the function "*X* looks for *Y*"—if it is true that John looks for a unicorn?

Another difficulty arises concerning "good," "beautiful," and other words of moral or aesthetic judgment. If, for example, beauty is indeed "in the eye of the beholder," then what one person calls beautiful might not similarly impress another, yet two people might keep arguing as to whether the thing is beautiful or is not. Thus, people may seem to agree on the meaning of the word, yet remain at odds about its proper application. The meaning of such "emotive" words cannot be decided in terms of truth alone.

A more serious objection to the aletheic (truth) theory arises from the fact that many significant utterances of natural language are not true or false at all. Whereas statements, testimonies, and reports are true or false, orders, promises, laws, regulations, proposals, prayers, curses, and so forth are not assessed in terms of truth or falsity. It is not obvious that the employment of words in these speech acts is less relevant to their meaning than their use in speech acts of the truth-bearing kind.

Meaning and use. The difficulties just mentioned lead to another view concerning the notion of meaning, a theory that may be called the use theory. This view admits that not all words refer to something, and not all utterances are true or false. What is common to all words and all sentences, without exception, is that people use them in speech. Consequently, their meaning may be nothing more than the restrictions, rules, and regularities that govern their employment.

The use theory has several sources. First, in trying to understand the nature of moral and aesthetic discourse certain authors suggested that such words as "good" and "beautiful" have an emotive meaning instead of (or in addition to) the descriptive meaning other words have; in using them one expresses approval or commendation. If one says, for instance, that helping the poor is good, one does not describe that action, but says, in effect, something like "I approve of helping the poor, do so as well." Such is the role of these words, according to these thinkers, and to understand this role is to know their meaning.

Wittgenstein's language games

The second, and more important, stimulus for the use theory was provided by the work of Ludwig Wittgenstein. This philosopher not only pointed out the wide variety of linguistic moves mentioned above but in order to show that none of these moves enjoys a privileged status proposed the idea of certain language games in which one or another of these moves plays a dominant or even an exclusive role. One can imagine, for instance, a tribe whose language consists of requests only. Members of the tribe make requests and the other members comply or refuse. There is no truth in this language, yet the words used to make requests would have meaning. Human language as it exists in reality is more complex; it is a combination of a great many language games. Yet the principle of meaning, according to this theory, is the same: the meaning of a word is the function of its employment in these games. To Wittgenstein the question "What is a word really?" is analogous to "What is a piece in chess?"

Finally, John L. Austin offered a systematic classification of the variety of speech acts. According to him, to say something is to do something, and what one does in saying something is typically indicated by a particular performative verb prefixing the "normal form" of the utterance. These verbs, such as "state," "declare," "judge," "order," "request," "promise," "warn," "apologize," "call," and so on, mark the illocutionary force of the utterance in question. If one says, for instance, "I shall be there," then, depending on the circumstances, this utterance may amount to a prediction, a promise, or a warning. Similarly, the words of the commanding officer, "You will retreat" may have the force of a simple forecast, or of an order. If the circumstances are not clear, the speaker always can be more explicit and use the normal form; *e.g.,* "I promise that I shall be there" or "I order you to retreat."

To rephrase the conclusion already stated: the dimension of truth and falsity is not invoked by all the utterances of the language; therefore, it cannot provide an exclusive source of meaning. There are other dimensions, such as feasibility (in case of orders and promises), utility (in case of regulations and prescriptions), and moral worth (in case of advices and laws). These dimensions may be as much involved in the understanding of what one said and, consequently, in the meaning of the words the speaker used, as the dimension of truth.

As previously mentioned, philosophers professing the aletheic theory claimed that the meaning of a word should be explained in terms of its contribution to the truth or falsity of the sentences in which it can occur. The latest form of the use theory is an appropriate extension of the same idea. According to some exponents, the meaning of

a word is nothing but its illocutionary act potential—*i.e.*, its contribution to the nature of the speech acts that can be performed by using that word. One difficulty with this view is that the definition is too broad, to the extent of being unilluminating or useless. Given this definition, nobody would know what any word means without knowing the entire language completely because the possibilities of employing a given word are not only without limit but extend to every conceivable context and circumstance. As Wittgenstein stated so forcefully,

> The sign (the sentence) gets its significance from the system of signs, from the language to which it belongs. Roughly: understanding a sentence means understanding a language.

Difficulties with the use theory

If this be the case, how can one account for the obviously gradual and prolonged process of learning a language? Indeed, the definition of the meaning of a word as illocutionary act potential seems to overstate the case. The obvious truth that the meaning of performative verbs, and other words closely tied to one illocutionary aspect or other, cannot be divorced from the nature of that type of speech act, does not entail that the meaning of an ordinary word, like "cat" or "running" is affected by any illocutionary force. Such words can occur in utterances bearing all kinds of illocutionary forces, so the contribution of these forces, as it were, cancel out. Nevertheless, what remains is the fact that all words are used to say something, in one way or another. The use theory would put a strong emphasis on the word "used" in the previous sentence. The next, and final, approach to meaning would stress the word "say."

Meaning and thought. In Wittgenstein's chess example, moves are made by moving the pieces. In a language, moves (saying something) are made by using words. And, according to the use theory, as a piece is defined by its move potential, so (the meaning of) a word is defined by its "saying" potential.

This analogy works only up to a certain limit. Whereas chess is only a game, the use of language is much more. One plays chess—or any other game—for its own sake; one speaks, however, with other ends in mind. Games, as it were, do not point beyond themselves; speech does. In order to see this, compare an ordinary conversation with such word games as children play—*e.g.*, exchanging words that rhyme, or words that begin with the same letter. These word games are language games and nothing more because the children use words according to certain rules. In doing so, however, they do not say anything except in the trivial sense of uttering words; nor are they called to understand what the other children say beyond the minimal feat of recognizing the words.

In real speech the situation is radically different. The point of using words in a real speech act is to be understood. If someone says, "It will rain tomorrow," his aim is to make the hearer believe that it will rain tomorrow. It is possible, of course, that the hearer or listener will not believe the speaker. Nevertheless, if the hearer understands what the speaker says, he will at least know that this is what the other person wants him to believe by saying what he says. Similarly, if one says, "Go home," and the listener understands what is said, then, whether or not the listener will actually go, he will at least know that this is what the speaker wants to bring about by using these words. Thus, the notion of saying something is inseparably tied to such concepts as belief, intention, knowledge, and understanding.

Speech as an expression of thought

The view just outlined is a reformulation of a very traditional idea, namely, that speech is essentially the expression of thought. Words are used not to play a game with fixed rules but to express beliefs and judgments, intentions and desires; that is, to make others know, by the use of words according to fixed rules, that one has certain beliefs, desires, and so forth, and that one invites others to share them.

"Expression of thought" sounds rather vague. For one thing, what is a thought? Suppose John believes that Joe has stolen his watch. John can express this belief by saying, "Joe has stolen my watch" or "My watch has been stolen by Joe" or "It is Joe who has stolen my watch" and so on. Moreover, if John is a multilingual person, he can express the same belief in German, French, and so forth. These variants, called paraphrases and translations respectively, will express the same belief, the same thought. But whereas it makes sense to ask for the exact words of John's statement or to ask about the language in which it was made, it would be foolish to ask for the exact words of John's belief or to ask about the language in which it is framed. The alternative in "Do you believe that Joe has stolen the watch, or that the watch was stolen by Joe?" does not make sense. Consequently, the same thought—the same proposition, as some philosophers prefer to call it—can be expressed by using various linguistic media. In other words, the same thought can be encoded in various codes (languages) and in various ways in the same code (paraphrases) in much the same way as the same idea can be expressed in speech or in writing and the same numbers can be written by using Roman or Arabic numerals.

From this point of view, it appears that saying something involves encoding a thought and that understanding what one said involves decoding and recovering the same thought. The meaning of a sentence will consist in its relation to the thought it is used to encode. This may be viewed as the fundamental thesis of the psychological theory of meaning.

As previously explained, no theory of meaning can be adequate as long as it treats sentences as indivisible units. For, in the first place, the potentially infinite number of sentences would defy any attempt to learn their meaning one by one, and, second, such a theory could not account for the obvious ability of fluent speakers to understand entirely novel sentences. There must be, therefore, a correlation between certain recurring elements of sentences (roughly, words) and certain recurring elements of thoughts (roughly, concepts or ideas). Accordingly, the learning of the semantic component of a language will consist in the learning of these connections.

Synonyms and analytic propositions

In this learning process two notions play a prominent role: synonymy and analyticity. As the sentences that express the same thought stand in the relation of paraphrase (or translation), so words or phrases that code the same idea are related as synonyms—*e.g.*, "vixen" and "female fox." Again, because one concept may include another, the sentence expressing this relation will record a conceptual truth or analytic proposition; *e.g.*, "A dog is an animal." A definition, finally, will exhibit all parts of a concept by a combination of such propositions.

What concepts are and how they are related to words are topics that have been discussed throughout the history of philosophy. The following problems related to concepts pertain to the core of philosophical psychology: whether all concepts are derived from experience, as Aristotle and the Empiricists believed, or whether some of them at least are innate, as Plato and the Rationalists maintained; whether concepts exist prior to and independent from their verbal encodings, as the Realists and Conceptualists claimed, or whether they are nothing but a certain "field of force" accompanying the words, as the Nominalists thought.

It should be noted that these disputes, in a modern garb, still continue with undiminished force. Contemporary Empiricists still try to reduce most concepts to a configuration of sense data, or a pattern of nerve stimulation, while the Behaviourists attempt to explain understanding in terms of overt behaviour. Modern-day Rationalists reply by insisting on the unique spontaneity of human speech and by reviving the theory of innate ideas.

MEANING IN LINGUISTICS

Semantics in the theory of language. The science of linguistics is concerned with the theory of language expressed in terms of linguistic universals—*i.e.*, features that are common to all natural languages. According to the widely adopted schema of the U.S. scholar Charles W. Morris, this theory must embrace three domains: pragmatics, the study of the language user as such; semantics, the study of the elements of a language from the point of view of meaning; and syntax, the study of the formal interrelations that exist between the elements of a language (*i.e.*, sounds, words) themselves. Subsequently, certain authors spoke of three levels: the phonetic, the syntactic (the pho-

Pragmatics, semantics, and syntax

netic and syntactic together are often called grammatical), and the semantic level. On each of these levels a language may be studied in isolation or in comparison with other languages. In another dimension, the investigation might be restricted to the state of a language (or languages) at a given time (synchronic study), or it might be concerned with the development of a language (or languages) through a period of time (diachronic study).

Semantics, then, is one of the main fields of linguistic science. Yet, except for borderline investigations, the linguist's interest in semantic matters is quite distinct from the philosopher's concern. Whereas the philosopher asks the question "What is meaning?", the typical questions the linguist is likely to ask include: "How is the meaning of words encoded in a language?" "How is this meaning to be determined?" "What are the laws governing change of meaning?" and "How can the meaning of a word be given, expressed, or defined?"

A few examples will suffice to illustrate some of these problems, and to show how the linguist's approach differs from that of the philosopher. In the matter of encoding, words are arbitrary signs; to some authors, particularly to the Swiss linguist Ferdinand de Saussure, this feature of arbitrariness represents an essential characteristic of all real languages. Nevertheless, in all languages there are clear cases of onomatopoeia—*i.e.,* the occurrence of imitative words, such as "whisper," "snore," "slap," and, more remotely, "cuckoo."

There are several other issues that pertain to the question of encoding. Certain languages show a marked preference for very specific words, at least in certain domains, while lacking the corresponding general terms, which are the only ones occurring in other languages. The Eskimos, for instance, have a number of words denoting various kinds of snow, but no single word for snow. Similarly, in English, although there are distinct names for hundreds of animal species, there is no name for the very familiar animal species of which the female member is called cow and the male member bull.

There are also languages, such as English or Chinese, that for the most part prefer single words to the compounded phrases that other languages (*e.g.,* German) seem to favour. Accordingly, whereas the English vocabulary is larger, the German words are more pliable, capable of entering into compounds often of great length and complexity. Such differences support Saussure's distinction between lexicological and grammatical languages.

Context-bound and context-free words

Another distinction can be drawn concerning the relative frequency and importance of context-bound and context-free words. The meaning of such English words as "take," "put," and "get" depends almost entirely on the context—*e.g.,* putting up with somebody has very little to do with putting off something or other. These can be compared with verbs like "canter" or "promulgate," which, by their very specific meaning, almost determine the context, rather than having their meaning determined by the context. Clearly, the context-bound type of word, such as "take" or "put," lends itself to idiomatic use, rather than the context-free word.

There are some obvious regularities in the change of meaning that are of interest to the linguist. One such regularity is the extension or transference of meaning based upon some similarities—*i.e.,* the phenomenon of metaphor. For example, one can speak of the leg of the table, the mouth of the river, the eye of the needle, and the crown of the tree. These are anthropomorphic metaphors: the transfer goes from something belonging to an individual or close to him (his body, garments) to something more remote. The same principle operates in the extension of meaning from a domain close in interest rather than in physical proximity. Baseball-minded people are apt to speak of "not getting to first base," "striking out," or "scoring a hit" in contexts often remote from baseball. For a similar reason, many abstract concepts are denoted by words transplanted from the concrete domain. Such phrases as "grasping ideas," "seeing the point of a joke," "body of knowledge," "in the back of my mind," and many others, are the result of this very important move from the abstract to the concrete.

Meaning changes of another type are the result of emotive factors. The word democracy, for instance, has all but lost its original meaning and has become a word applicable to any system the speaker wants to praise. The contrary development is exhibited in the recent history of such words as "Fascist" and "aggression." In order to avoid derogatory connotations one is often forced, by social pressure, to use euphemisms, often to the detriment of clarity. Examples of this include the switch from "underdeveloped nations" to "developing nations," from "retarded children" to "exceptional children," and from "old people" to "senior citizens."

The preceding are but a few examples concerning coding and meaning change. The questions about the ways of finding out what a word means and about the manner of giving an adequate definition of a word deserve a more detailed account.

Meaning, structure, and context. Foreigners in a strange country and linguists are often confronted with the task of learning a new language. It is important to realize that in doing so they do not set out with a completely blank mind: they expect to learn a language (*i.e.,* a system of communication describable in terms of a large set of linguistic universals). They expect—consciously in the case of the linguist and unconsciously in the case of the layman—to find words and sentences, grammatical structures, and illocutionary forces in that language. And, on the semantic level, they expect words that will fit into the familiar semantic classes. They are confident, in other words, that the language they intend to master will be intertranslatable with their own.

Structure and language learning

Therefore, although at the very beginning their learning remains on the ostensive level (trying to find out the name of this or that kind of object), very soon they proceed to the level of first guessing, then establishing, the meaning of words from the contexts in which they occur. This has to be the case, for words that in any way can be viewed as "names" of objects (*i.e.,* that could be learned ostensively) form but a fraction of the vocabulary of any language. Anyone who doubts this should but try to list the words from this paragraph that could be learned ostensively. Moreover, linguists find no great difficulties in learning dead languages—*e.g.,* that of the ancient Egyptians—without any contact with any speaker, provided that a sufficiently large corpus of texts is available and that some clues are provided to the meaning of at least some words.

If any more evidence concerning this point is needed, one should remember that "pictorial" dictionaries are bound to remain on the kindergarten level, and that the mark of a good dictionary is the abundance of appropriate contexts. Thus, the contexts show the concept.

These intuitions are behind the U.S. philosopher Paul Ziff's semantic theory. According to Ziff, the meaning of a word is a function, first, of its complementary set, which consists of all the acceptable sentences in which the word can occur, and, second, of its contrastive set, which consists of all of the words that can replace that word in all of these sentences without rendering the sentences deviant. Clearly, the elaboration of the contrastive set will produce words more and more similar in meaning to the word in question, the limiting case being synonyms that can occur wherever the word in question can occur.

This theory is in need of further refinement. In the sentence "The cat sleeps," the fact that "cat" can co-occur with "sleep" undoubtedly casts some light on the meaning of these words (a cat is a kind of thing that can sleep). But there are a great number of sequences that could complete the frames "The cat sleeps and . . . ," ". . . said that the cat sleeps," and so forth. Clearly, the near infinity of the resulting sentences will not contribute anything to the meaning of "cat" beyond what the segment "the cat sleeps" already contributes.

Surface forms and underlying structures

Transformational grammar can be of assistance at this point. According to this approach, the sentences just considered are simply surface forms, each corresponding to an underlying structure, in which "cat" and "sleep" appear as forming an elementary, or kernel sentence (roughly: "a cat sleeps"). The essence of Ziff's insight can be rein-

terpreted in terms of the notions developed by Zellig S. Harris: co-occurrence (instead of complementary set) and co-occurrence difference (instead of contrastive set), both restricted to kernels. Because the vocabulary of a language is limited and the number of kernel structures is very small, the meaning of a word can be determined on the basis of a finite set of elementary sentences.

The contribution of grammar to semantic theory is by no means exhausted by this step. For the grammatical restrictions on a word represent, as it were, the "skeleton" of its meaning before the "flesh" is put on by the co-occurrences. The very first step in giving the meaning of a word is to specify its grammatical category—noun, verb, adjective, adverb, connective, and so forth—and not to speak of grammatical constants (such as the first, but not the second, "to" in "I want to go to Paris"), the meaning of which, if any, is entirely determined by their grammatical role. A refined grammar yields much more: the fact that the adjective "good," for example, unlike adjectives like "yellow" or "fat," can occur in the frames "(He is) good at (playing chess)"; "(The root is) good to (eat)"; "It is good that (it is raining)"; "It was good of (him) to (come)" says a great deal about the meaning of that word. The co-occurrences then complete the picture.

Lexical entries. Good dictionaries offer a variety of contexts for the items listed, but, obviously, this is not enough. For one thing, no dictionary can list all the co-occurrences. There must be devices to sum up, as it were, the information revealed by the contexts. This is the role of dictionary definitions. The branch of scientific semantics that is concerned with the form and adequacy conditions of dictionary entries is called lexicography.

A systematic study of dictionary entries was presented in the 1960s by the U.S. philosophers Jerrold J. Katz and Jerry A. Fodor. According to them, the standard form of a dictionary entry comprises three kinds of ingredients: grammatical markers, semantic markers, and distinguishers. The grammatical markers describe the syntactic behaviour of the item in question in terms of a refined system of grammatical categories. The traditional division of words into nouns, adjectives, verbs, adverbs, and so on is but the first step in this direction. The class of nouns, for example, has to be subdivided into count nouns (like "cat"), mass nouns (like "water"), abstract nouns (like "love"), and so forth. The class of adjectives must be classified into subclasses that are fine enough to capture the grammatical peculiarities of such adjectives as "unlikely" or "good." The traditional subdivision of verbs into transitives and intransitives has to be completed to account for such verbs as "compare" or "order," which obviously involve three noun phrases ("someone compares something to something"), often of a particular kind ("human" nouns or noun clauses).

The idea of a semantic marker is merely a further elaboration of the traditional notions of genus and species. The result is a system of semantic markers that comprise such items as "physical object," "animate," "human," "male," "young" (in the case of the entry for "boy"), and others. Katz claims that the problems of synonymy, analyticity, and contradiction can be handled, at least in part, in terms of lexical items sharing some or all of their semantic markers.

Finally, the distinguisher completes the dictionary entry by giving, as it were, the leftover, if any, of the semantic information. There is no general form for the distinguisher; it may give the atomic weight (for elements), purpose (for tools), concise description (for animals), and so forth.

Generative semantics. According to the original formulation of generative or transformational grammar, the semantic and the syntactic components were regarded as distinct elements in the deep structure of a sentence. The syntactic component consisted of a relation of phrase markers giving the transformational structure of the sentence, usually represented in terms of "tree" diagrams with such nodes as "S" for sentence, "NP" for noun phrase. The semantic content entered through the process of lexical insertion—*i.e.*, the replacement of some of the nodes by words, the carriers of meaning. Lexical insertion was supposed to take place at the very beginning of the series of transformations leading up to the surface form of the sentence. The original input of meaning, as it were, was carried through the transformations yielding the semantical reading, or sense, of the whole sentence.

Lexical insertion in a generative grammar

Several modern studies have attempted to demonstrate that this separation between syntax and semantics cannot be maintained. It appears that certain words in themselves indicate a structure analogous to syntactic structures. For example, consider "harden" and "break." To harden something is to cause that thing to become hard (or harder); to break something is to cause something to become broken. Because "harden" consists of two elements, "hard-" and "-en," thus it could be argued that the word itself is structured; "break" does not indicate any structure, yet its meaning clearly involves one. "Broken," therefore, carries a more basic semantic unit than "break." Again, in the case of such verbs as "remind," "allege," "blame," or "forgive," one feels that their meaning is highly structured, and, with some thinking, one can articulate the presuppositions of these words, which involve a great number of semantic units in a very complex relationship.

These and similar reasons support the theory of generative semantics, which denies a clear distinction between the semantic and the syntactic components. The transformations, in this theory, connect the surface structure of the sentence with its semantic representation (or according to some linguists, its logical form). Words, then, can encode either a semantic primitive (such as "blue") or a whole structure (such as "forgive") within the semantic representation. Thus, there is no definite point in the transformational history of the sentence at which lexical insertion must occur. The ultimate conclusion of this view is that, instead of the threefold division of semantics, syntax, and phonetics, all that is needed is the simple distinction between semantics and phonetics, corresponding to the distinction between meaning (as structured) and its verbal encoding. How much, finally, of the semantic structure can be attributed to a particular language, and how much can be ascribed to common (and possibly innate) elements of the human mind, remains a fascinating problem for continued study. (Z.V.)

The study of writing

This section deals with the various fields of study relating to written sources and writing systems. Of these, the most important are philology, which deals mainly with written and oral source materials; and linguistics and grammatology, which deal mainly with systems of signs, which in turn are based on primary source materials. The use of the word "mainly" implies that, while certain fields deal predominantly with certain aspects of study, there are no sharp divisions between them, and some areas of study can be treated by both disciplines.

THE SCOPE OF THE STUDY OF WRITING

Sources of information. In a study of any system of writing, a basic distinction must be made between the raw materials (subject matter) that are studied and the systems deduced from them. Spoken (or written) utterances represent the raw materials of a language; a grammar or a lexicon presents a linguistic system abstracted and reconstructed from these materials. Similarly, written texts represent the raw materials of writing from which an alphabet can be reconstructed.

Manifold sources of information bear on systems of writing: (1) Written sources proper include texts, inscriptions, books, and manuscripts. The ability to understand these sources can be handed down traditionally from generation to generation, as in the case of the Hebrew or Chinese writings, or it must be recovered by a process of decipherment, as in the case of the Egyptian hieroglyphic. (2) Lists of signs, alphabets, syllabaries, and so on can be devised by the "inventors" of a writing at the time when the writing is first introduced; or they can be reconstructed in later times from the written sources by teachers or scholars for didactic or scholarly purposes. (3) Studies are based on

Written sources

written sources or lists of signs, alphabets, syllabaries, and so on. Such studies can be primary, when they are based directly on sources, or secondary, when they must rely on primary studies for information.

A partial study, which is delimited by the subject matter, times, and area, must be distinguished from a general, comprehensive study of the subject. Thus an analysis of the writing of Middle English, Egyptian hieroglyphic writing, or the Greek alphabet may be contrasted with an analysis of the structure of alphabetic writings or a study of cursive or monumental writing.

Definitions of terms. *Semiotics.* As has been noted above, men communicate with each other by means of various systems of signs, of which the most universal are language, a system of auditory communication; and gesture language and writing, two systems of visual communication. For the general science of signs, several terms have been proposed, of which the term semiotic, or semiotics, as here preferred, may perhaps be the most appropriate. Semiotics covers a much broader area than the term semantics, which deals with the meaning of linguistic elements.

Philology, linguistics, and grammatology. Philology, involved mainly in the study of the linguistic sources of a people or a group of peoples, forms the basic means for the comprehension of their respective cultures. It deals less with oral sources than with written sources, mainly literature (whatever its exact meaning). Philology deals with the formal aspects of writing under the topic of epigraphy and paleography. Linguistics is concerned with the study of linguistic systems as reconstructed mainly from oral sources. Pursued less than the study of "oral language," the study of the "written language"—that is, of the language as it is used in written sources—is also a matter of linguistics. Linguistics deals with the structural aspects of writing under the heading of graphemics.

The field of study that deals with writing in the broadest sense is called grammatology. Equally appropriate terms for this subject are grammatonomy and graphonomy.

GRAMMATOLOGY

Three main approaches. Three main approaches to grammatology can be distinguished: descriptive-historical, typological-structural, and formal.

Descriptive-historical. The traditional descriptive-historical approach to the study of writing is by far the most common. This is a simple narrative approach to the description of writing in its historical evolution. The apparent shortcoming of descriptive-historical texts is the general lack of systematic typology—that is, systematic classification by type. Good studies on individual writings, such as hieroglyphic Egyptian and the Greek alphabet, are not wanting. What is entirely missing is theoretical and comparative evaluation of the various types of writing, such as discussions of various types of syllabaries, alphabets, word signs, and logo-syllabic writings.

The historical approach is further vitiated through confusion with considerations of a geographic nature, as evidenced by such chapter titles as "Asiatic Writings" and "American Writings," or "Writings of Asia" and "Writings of America," which are frequently found in the standard manuals on writing.

Typological-structural. The typological-structural approach is based on the realization of the importance of structure and typology in the study of writing. In contrast to the traditional approach, according to which the writings of the world are described in their evolutionary progress, the new approach requires first a thorough analysis of the structure of the individual writing systems and then their classification by type within the framework of writing in general.

The meaning of structure

Under "structure" is meant here primarily what is sometimes called the "inner structure," which is concerned with the function of a writing system, in contrast to the "outer structure," which involves its formal characteristics. Considerations of function are based on such points as word writing (logography), syllable writing (syllabography), and letter writing (alphabetography), and the typology of logographic, syllabic, and alphabetic systems of signs. Considerations of form involve such points as the contrast between the pictorial and linear systems or between the monumental and cursive writings. While, theoretically, both approaches are acceptable, the emphasis placed here on function, rather than form, may be illustrated by considering the following case. From the point of view of function, the Morse alphabet represents the same type of alphabet as the Latin writing and its descendants; from the point of view of form, the Morse alphabet is formally independent of the Latin writing. Based on considerations of function, the Morse alphabet is considered as being of the Latin type, despite its different formal, outer structure.

Formal. There are two kinds of formal approach to the field of grammatology: the traditional approach, as practiced mainly in the philological disciplines under the topic of epigraphy and paleography (see below), and the formal approach to sign analysis recently initiated in the United States. The aim of the latter is to provide a scientific account of cursive writing as practiced in that country. Its procedure consists first of breaking up the letters of the cursive writing into its component segments, which appear in the form of bars, hooks, arches, and loops, and then of providing formation rules governing the linking together of these segments into letters and into larger strings corresponding to words of the language. It is said that 18 such segments are sufficient to describe all English lower- and upper-case letters.

Subdivisions of grammatology. Three main subdivisions of grammatology can be distinguished: subgraphemics, graphemics, and metagraphemics. They are treated mainly by scholars versed in philological and linguistic disciplines.

Subgraphemics. The field of subgraphemics deals mainly with primitive forerunners of writing that utilize visual marks having no set correspondences in language. In its sublinguistic aspects, subgraphemics can be compared to kinesics, the study of the various communicational aspects of learned, patterned body motion behaviour; and cherology, the study of gesture language.

Graphemics. The field of graphemics deals with full writing or phonography, as represented in systems of writing in which written signs generally have set correspondences in elements of language. The field of graphemics thus deals with writing after it became a secondary transfer of the language, a vehicle by which elements of the spoken language were expressed in a more or less exact form by means of visual signs used conventionally. This took place for the first time about 5,000 years ago in the Sumerian and Egyptian writings.

Instead of "graphemics," other scholars use the terms "graphics" or "graphic linguistics." All three terms are frequently misused by scholars who limit the terms to the study of alphabetic writings, overlooking or paying scant attention to all other types of writing, such as the logosyllabic and syllabic systems.

Little work has been done in the field of relations of writing to language. Philologists have been concerned mainly with the historical evolution of writing and have paid little attention to the interrelations between writing and language. Linguists have been more concerned with the spoken language than with the written language. When interested in written languages, they have often limited their study to living written languages, neglecting the rich sources of information that can be culled from ancient written languages and from pre-alphabetic systems. The question of the relationship of writing to language has been pursued in recent years mainly by scholars with a background in linguistics. Because of their interest in modern languages and writings, this implies generally relations between the alphabet and language. A general treatment of the subject can be found in the respective chapters of the introductory manuals to linguistics. Linguists generally have stressed the independent character of writing and have studied it as an independent system rather than as a system ultimately based on and related to the underlying language.

While the connections between language and writing are close, there has never been a one-to-one correspondence between the elements of language and the signs of writing. The "fit" (*i.e.*, the correspondence) between language and

writing is generally stronger in the earlier stages of a certain system of writing and weaker in its later stages. This is because a writing system when first introduced generally reproduces rather faithfully the underlying phonemic structure (structure of sounds). In the course of time, writing, more conservative than language, generally fails to keep up with the continuous changes of language and, as time progresses, diverges more and more from its linguistic counterpart. A good example is the old Latin writing, with its relatively good "fit" between graphemes (the written letters or group of letters that represent one phoneme or individual sound) and phonemes as compared with the present-day French or English writing, with their tremendous divergences between graphemes and phonemes. In some cases, recent spelling reforms have helped to remedy the existing discrepancies between writing and language. The best "fit" between phonemes and graphemes has been achieved in the Korean writing in the 16th century and in the Finnish and Czech writings of modern times.

The growing divergence of language from writing

Families of writings are not related to families of languages. Note, for example, that English and Finnish are written in the Latin writing but belong to two different families of languages, and that the cuneiform writing was used in antiquity by peoples speaking many different languages.

The temporal primacy of language over writing has been taken for granted by most scholars, especially the American linguists. It has been contested by some European scholars, who claim that writing is as old as oral language and gesture language. The fact is that full writing, expressing linguistic elements, originated only about 5,000 years ago in Mesopotamia and Egypt and that full writing is therefore much younger than language. Only if the semasiographic stage is included under writing can the assumption of equal temporal hierarchy of writing and language be admitted. (Semasiography is the use of marks to convey meaning without the presence of linguistic elements.) As noted elsewhere, however, the semasiographic stage should not be treated as full writing, but as a forerunner of writing.

Metagraphemics. A study of the various metagraphic devices (*e.g.,* punctuation marks, capital or italic letter forms) that are used besides or in addition to writing proper may be called metagraphemics or paragraphemics. This is still an obscure field, and its relationship to both subgraphemics and graphemics needs a thorough investigation.

EPIGRAPHY AND PALEOGRAPHY

The investigation of writing from the formal point of view has been traditionally the prime domain of the epigrapher and paleographer. Epigraphy is concerned mainly with inscriptions written in characters that are incised or scratched with a sharp tool on hard material, such as stone or metal; paleography deals mainly with manuscripts written in characters that are drawn or painted with pen, pencil, or brush on soft material, such as leather, papyrus, or paper. Since epigraphy means "writing upon something" and paleography means "old writing," it is clear that the distinction made above between epigraphy and paleography cannot be justified on etymological grounds. The distinction has grown artificially over the years, as one scholar or another began to apply one or the other term to his own branch of study of written sources. Because of the close interrelations between epigraphy and paleography, some scholars refuse to admit any distinction between the two and prefer to use only the term paleography.

Definitions of epigraphy and paleography

The main characteristics of epigraphy and paleography as listed above may be applied, with some leeway, to the ancient Near East (Mesopotamia, Egypt, Anatolia), the classical world, China, India, the Islāmic world, and, in general, to the Western writings from the Middle Ages down to the introduction of the printing press. But there are some difficulties: Mesopotamian and Aegean clay tablets are soft, and the writing is cursively executed, both points characteristic of paleography, but the tablets are also durable and bulky and have incised, concave characters executed with a stylus, all points characteristic of epigraphy. Similarly, the wax tablets of the classical world are soft, perishable, and cursively executed, characteristics of paleography, but are written with concave characters, incised with a stylus, characteristic of epigraphy. There are likewise some difficulties in the classification of tablets of wood; they are soft and perishable, and the writing on them is generally cursive, characteristic of paleography, but they have characters incised or scratched with a sharp implement, characteristic of epigraphy.

Formal and interpretive studies

Paleography and epigraphy are involved in the study of written sources from two points of view: the purely formal aspect and the hermeneutical (interpretive) aspect.

The study of the purely formal aspect, possible without any understanding of the contents or without an extended study of the contents, is concerned, for example, with the kind, form, and size of the materials; the technique of writing; and the form, order, and direction of writing. Hermeneutics, possible only with study of the contents, is concerned, for example, with the dating and localizing of written sources, their authorship, linguistic interpretation, and content evaluation.

A general scientific discipline of epigraphy and paleography does not exist. There are no studies that treat of the subject from a general, theoretical point of view, encompassing all the written sources, wherever they may be found. There are, for example, no treatises listing and discussing the various materials, or shapes and sizes of materials, used for writing throughout the world, just as there are no structural-typological studies that treat of the formal evolution of signs from pictorial to linear or from round to angular. Among other potential topics that await investigation are: trends in ductus (hand; the general shape and style of letters), such as individual, national, and regional; the direction of writing, indication of prosodic features (quantity, stress, and tone), and names of signs (letters). The narrow fields that are represented are, for example, West Semitic epigraphy, Arabic paleography, Greek and Latin epigraphy and paleography, or Chinese epigraphy and paleography. In all cases, these narrow fields of study form subdivisions of wider but still linguistically or geographically defined fields of study, such as Semitic or Arabic philology, classical philology, Assyriology, and Sinology.

HISTORY OF THE STUDY OF WRITING

The first students of writing were doubtless the very originators ("inventors") of a new writing system. By "writing system" is meant here a full writing in which the individual signs of the writing stand for the corresponding elements of the language, which is to be contrasted with the forerunners of writing, in which the individual signs have but loose connection with language. As a result of a discrete analysis of the language for which a writing system was devised, lists of the elements of a language and their proposed written counterparts must have been first compiled and experimented with in actual practice. The establishment of a full system of writing also required conventionalization of forms and principles. Forms of signs had to be standardized so that the users would draw the signs in approximately the same way. Regulation of the system had to take place in the matter of the orientation of signs and the direction, form, and order of the lines, columns, and the sides of a text. Correspondences established between signs and words were paralleled by those between signs and definite syllabic values. After the initial period of trial and error, the established correspondences were conventionalized by being taught in schools.

Studies prior to the 18th century. The activities involved in setting up a full system of writing are indirectly attested in the Sumerian school texts, known almost from the beginnings of the Sumerian writing, which appear in the form of lists of signs and words, and scribal and literary exercises. The scribal activities of the Sumerians and, in the later periods, Akkadians are matched, albeit to a smaller degree, as far as actual attestation is concerned, by those of other peoples of the ancient Near East, such as the Egyptians and Hittites.

The Sumerian school texts

The study of the language and the corresponding writing was highly developed among the Chinese and Indic peoples, as best exemplified by the great Indic grammarian Pāṇini (about the 4th century BC) and his school, as

well as among the Greeks and to a lesser degree among the Romans. Beginning in the Middle Ages, the Arabs and Jews showed great interest in matters pertaining to their languages and writings. Important contributions were made by the Arab scholars Sībawayh (8th century) and az-Zamakhsharī (1075–1143) and by the Jews Rashi (1040–1105) and David Kimhi (*c.* 1160–*c.* 1235). Interesting, but largely fantastic, is a collection of several dozen alphabets and sign lists put together by the Arab Aḥmad ibn Abū Bakr ibn Waḥshīyah (*c.* 800). Among the early Europeans were the Spanish bishop Diego de Landa (1524–79), with his analysis of the Maya writing, and the German scholar Athanasius Kircher (1602–80), with his frequently mystic ideas about writing, especially the Egyptian hieroglyphic.

Modern Western studies. Modern general studies of writing in the West began in the second half of the 18th century. This group of early studies was based almost exclusively on Greco-Latin writing, with its further developments in the Middle Ages and modern times, and supplemented by a scattering of Semitic alphabets, such as the Hebrew, Arabic, and Syriac. The books of this first period are of no more than historical interest today.

In contrast to these early studies, the second group consists of the larger and much more serious undertakings of François Lenormant (1872), Heinrich Wuttke (1872), Carl Faulmann (1880), Isaac Taylor (1883), and Philippe Berger (1891). These books became standard manuals in the field of writing and served that function until the first half of the 20th century.

Next in time is a group of studies generally smaller in size and more limited in coverage than the group discussed just above, as represented by Walter James Hoffmann (1895), R. Stübe (1907), Theodore Wilhelm Danzel (1912), Karl Weule (1915), and William A. Mason (1920). On the basis of such monumental works on the American Indian writings as those of Henry R. Schoolcraft (1851), Garrick Mallery (1886 and 1893), and other similar works, the authors of this group of studies have emphasized much more than did the earlier scholars the importance of the forerunners of writing in the whole field of the study of writing.

The main characteristic of the next period, the first half of the 20th century, during which great manuals on writing were produced, was the descriptive-historical approach, as it was of all the preceding group of studies on writing.

The typological-structural approach to the study of writing appeared in the latter half of the 20th century, along with interest in the relationship of writing to society. Representatives of this and of the previous group are briefly characterized in the bibliography. (I.J.G./Ed.)

BIBLIOGRAPHY

General linguistics: ROBERT H. ROBINS, *A Short History of Linguistics,* 2nd ed. (1979), *General Linguistics: An Introductory Survey,* 3rd ed. (1980), is a comprehensive and balanced treatment of the whole field. LEONARD BLOOMFIELD, *Language* (1933), a classic introduction to the subject, is still not completely superseded and is essential reading for an understanding of subsequent American work. CHARLES F. HOCKETT, *A Course in Modern Linguistics* (1958), a comprehensive, stimulating, though somewhat personal textbook, represents the post-Bloomfieldian period in the United States. JOHN LYONS has produced a number of notable surveys: *Introduction to Theoretical Linguistics* attempts to synthesize more traditional and more modern ideas on language, paying particular attention to generative grammar and semantics; *New Horizons in Linguistics* (ed., 1970), contains previously unpublished chapters on developments in most areas of linguistics; *Language and Linguistics: An Introduction* (1981) is a textbook covering theoretical developments. MARTIN JOOS (ed.), *Readings in Linguistics* (1957), is an excellent selection of key articles on structuralism in the post-Bloomfieldian period. Z.S. HARRIS, *Methods in Structural Linguistics* (1951), offers the most extreme and most consistent expression of the distributional approach to linguistic analysis—important for the development of generative grammar. NOAM CHOMSKY, *Syntactic Structures* (1957), is the first generally accessible and relatively non-technical treatment of generative grammar, widely recognized as one of the most revolutionary books on language to appear in the 20th century; J.P.B. ALLEN and PAUL VAN BUREN (eds.), *Chomsky: Selected Readings* (1971), contains an annotated selection of key passages from Chomsky's main works. S. PIT CORDER (ed.), *The Edinburgh Course in Applied Linguistics,* 4 vol. (1973–77), is a collection of readings covering a wide range of views. RICHARD C. OLDFIELD and J.C. MARSHALL (eds.), *Language* (1968), J.A. FODOR, T.G. BEVER, and M.F. GARRETT, *The Psychology of Language* (1974); and JOSEPH F. KESS, *Psycholinguistics* (1976), are important works in psycholinguistics. DELL HYMES (ed.), *Language in Culture and Society* (1964), is an excellent selection of articles in sociolinguistics and anthropological linguistics. Valuable information is found in ROY HARRIS, *The Language Myth* (1981); GEORGE A. MILLER, *Language and Speech* (1981); SUSAN BASSNETT-MCGUIRE, *Translation Studies* (1981); ERIC WANNER and LILA R. GLEITMAN (eds.), *Language Acquisition: The State of the Art* (1982); and HANS AARSLEFF, *From Locke to Saussure: Essays on the Study of Language and Intellectual History* (1982).

Semantics: W.P. ALSTON, *Philosophy of Language* (1964), is the best current introduction to philosophical semantics. STEPHEN ULLMANN, *Semantics: An Introduction to the Science of Meaning* (1962), has become a classic work. M. BLACK, *Language and Philosophy* (1949), discusses some earlier views. L. BLOOMFIELD, *Language* (1933), contains a classic discussion of scientific semantics. B.L. WHORF, *Language, Thought and Reality,* ed. by J.B. CARROLL (1956), raises the issue of linguistic relativism. J.J. KATZ, *The Philosophy of Language* (1966), offers a semantic theory tied to generative grammar, the best expression of which is found in N. CHOMSKY, *Aspects of the Theory of Syntax* (1965). W.V. QUINE, *Word and Object* (1960); and P. ZIFF, *Semantic Analysis* (1960), represent two different but influential semantic theories. Philosophy of language is explored also in the following monographs: IRMENGARD RAUCH and GERALD F. CARR (eds.), *The Signifying Animal: The Grammar of Language and Experience* (1980); CHARLES ALTIERI, *Act and Quality: A Theory of Literary Meaning and Humanistic Understanding* (1981); GRAHAM D. MARTIN, *The Architecture of Experience: A Discussion of the Role of Language and Literature in the Construction of the World* (1981); NATHAN U. SALMON, *Reference and Essence* (1981); DERECK BICKERTON, *Roots of Language* (1981); SÁNDOR HERVEY, *Semiotic Perspectives* (1982); JEREMY CAMPBELL, *Grammatical Man: Information, Enthropy, Language, and Life* (1982); ANNETTE LAVERS, *Roland Barthes, Structuralism and After* (1982); and DAVID LIGHTFOOT, *The Language Lottery: Toward a Biology of Grammars* (1982).

Periodicals: Language, Word, International Journal of American Linguistics (United States); *Philological Society Transactions, Journal of Linguistics* (Great Britain); *Lingua, Studies in Language* (Holland); *Bulletin de la Société de Linguistique de Paris* (France).

(Ed.)

Lisbon

Lisbon (Portuguese: Lisboa), the capital of Portugal and of the district that bears its name, is the nation's chief port, largest city, and commercial, political, and tourist centre. It stands on the westernmost point of land of continental Europe. The city's name is a modification of the ancient Olisipo (variant Ulyssipo), and its founding has been variously attributed to Ulysses (Greek: Odysseus), the hero of Homer's *Odyssey,* to Elisha, grandson of the Hebrew patriarch Abraham, and, more credibly, to Phoenician colonists.

Lisbon owes its historical prominence to its superb natural harbour, one of the most beautiful in the world. The city lies on the north bank of the Tagus River (Rio Tejo), about eight miles (13 kilometres) from the river's entrance into the Atlantic Ocean. From the ocean upstream to the city, the river is almost straight and about two miles wide. It is spanned, on the west side of the city, by the 25th of April Bridge (formerly called the Salazar Bridge), the longest suspension bridge in western Europe. Just east of the bridge, the Tagus suddenly broadens into a bay seven miles wide called the Sea of Straw (Mar de Palha)—a reference to the sheen of the water. Scenically spectacular though it may be, this hill-cradled bay of burnished water lies on a strategic sea route and serves as a busy port, handling much of the exports and imports of Portugal and Spain.

This article is divided into the following sections:

Physical and human geography 81
- Character of the city 81
- The landscape 81
  - The city site
  - Climate
  - The city layout
  - Housing
- The people 82
- The economy 82
  - Industry
  - Commerce and finance
  - Transportation
- Administration and social conditions 83
  - Government
  - Health
  - Education
- Cultural life 83

History 83
- The early period 83
  - Prehistoric to Moorish times
  - The Portuguese conquest
  - The Age of Discovery
- Evolution of the modern city 84
  - Disaster and reconstruction
  - 19th-century expansion
  - The 20th century

Bibliography 85

Physical and human geography

CHARACTER OF THE CITY

Once a remote outpost on what was thought to be the farthest edge of the known world, by the 15th century Lisbon was established as the centre of operations for Portuguese exploration. Although this seagirt city of white houses and elegant parks and gardens is no longer the capital of a vast overseas empire, it remains a busy commercial and tourist centre. Lisbon has exchanged the sounds of the past—the cries of Galician water-carriers and of bakers bearing huge baskets of bread, the whistles of knife-grinders, the bagpipes of peasants from the north—for the honking of congested motor traffic and the clang of trams.

Some traditions remain, however. One can still see in the streets *varinas* (fish sellers), dressed in long, black skirts, carrying their wares in baskets on their heads; and one can still hear sung in the little cafés of the medieval Alfama quarter the sad, romantic music called fado. The port maintains an intimacy with its city that was common in the days before steam. Amid the freighters, warships, liners, and ferryboats, a picturesque note is struck by the *fragatas,* of Phoenician origin; these crescent-shaped boats with their striking black hulls and pink sails still perform most of the harbour's lighterage. Vessels tie up at quays open to the everyday life of the town, where the clang of the trolley cars blends with the sound of ships' bells. At dawn, fishing smacks deposit their catch at the town's front doorstep for noisy auction to Lisbon dealers, while the *varinas* wait to fill the baskets they peddle through the streets. Farther within, the fish market gives way to the equally colourful and clamorous fruit and vegetable market. Despite modernization, Lisbon, in many ways, retains the air of a 19th-century city.

THE LANDSCAPE

The city site. Lisbon is built in a succession of terraces up the slopes of a range of low, rolling hills, which rise from the banks of the Tagus River and the Sea of Straw northwest toward the Sintra Mountains, whose covering of lush Mediterranean and north European flora provides an attractive retreat for the city's population. Sections of the city vary considerably in height, especially in the older areas along the water's edge, offering splendid views of the river and the low cliffs that line the river's southern shore.

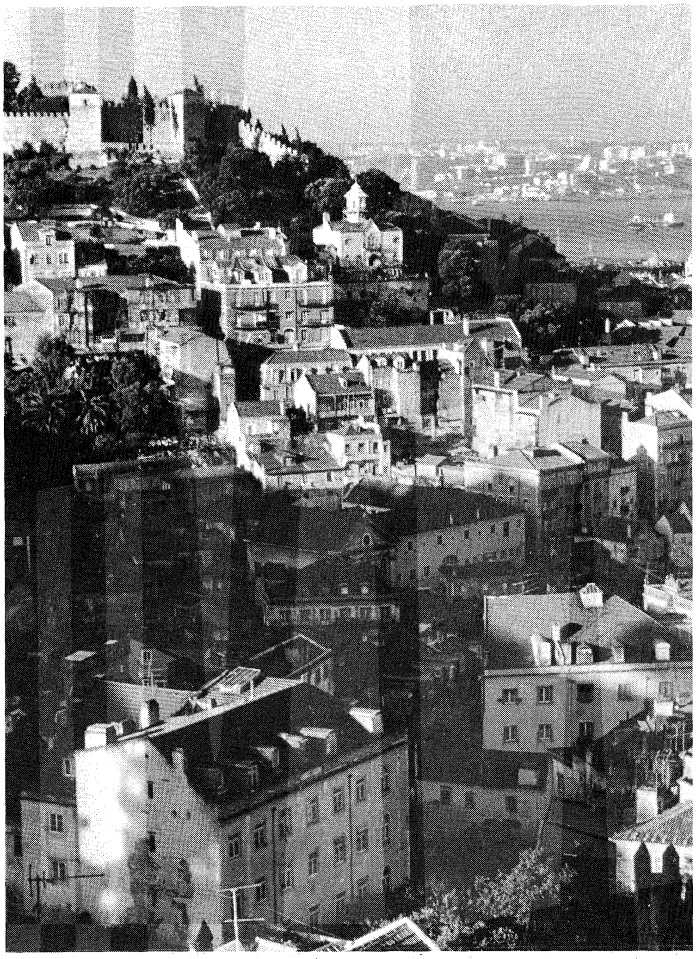
Several geologic faults cross Lisbon and the surrounding region. A major earthquake devastated the city in 1755, but seismic activity in the 20th century has been limited to slight tremors.

Climate. Lisbon has a mild and equable climate, with a mean annual temperature of 63° F (17° C). The proximity of the Atlantic and the frequency of sea fogs keep the atmosphere humid, and summers can be somewhat oppressive, although the city has been esteemed as a winter health resort since the 18th century. Average annual rainfall is 26.6 inches (666 millimetres).

The role of the river

The city layout. It is traditional for Lisbon's poets to refer to the entwining Tagus as the city's lover. The river is indeed an ever-present part of the city's decor, and the official entrance to Lisbon is a broad marble stair mounting from the water to the vast, arcaded Commerce Square (Praça do Comércio). The three landward sides of the square are surrounded by uniform 18th-century buildings, sea-green and white. This formal, Baroque-flavoured composition is pierced by a monumental archway, built a century later, marking the entry into the central city. In the middle of the square, surrounded each day by a regiment of parked automobiles, stands a bronze statue of King Joseph (José) I, on horseback, from which derives the nickname given the area by English sailors, Black Horse Square.

The square lies at the south end of Lisbon's central district, the Cidade Baixa ("Lower City"). The Baixa area was completely rebuilt after the earthquake in 1755 under the supervision of Joseph I's prime minister, Sebastião de Carvalho, later the marquês de Pombal. The streets are laid out in a grid pattern broken by spacious squares. A series of parallel streets, each named for its original intended occupants (*e.g.,* Rua Áurea ["Golden Street"] for the goldsmiths), run north from Commerce Square to Dom Pedro IV Square, commonly known as Rossio Square. The Rossio is a traditional centre of activity and the starting point of the city's main promenade, the wide, gently sloping Avenida da Liberdade. This tree-lined boulevard leads north from the city centre to more modern sections of the town.



Central Lisbon and the Tagus River, with the Castle of St. George on the hilltop.
Robert Frerck/Odyssey Productions

To the east of the Baixa lies the Alfama, the oldest part of the city, where narrow, winding streets crowd down to the river between a jumble of houses. In this area, from the hill where Lisbon was first founded, the Castle of St. George (São Jorge) watches over the city. The castle, like most of the buildings in the Alfama, is Moorish in origin; it was named for England's patron saint, in honour of an alliance made in 1386. Just below it, the austere, white Church of St. Vincent-Outside-the-Walls guards the remains of the saint, which were—according to legend—miraculously brought to the city in a ship guided by two ravens. To commemorate the event, the birds are depicted on the Lisbon coat of arms.

A number of neighbourhoods extend west of Cidade Baixa toward the suburb of Belém. Each possesses its own distinctive character, reflecting the epoch in which it was built. The Bairro Alto ("Upper District"), for example, dates primarily from the 17th century. It is characterized by its straight, narrow, steeply inclined streets. Some of its streets, especially those leading down to the Baixa, are so steep that they give way to stairs, cable cars, and even an elevator (an iron structure designed by the French engineer Gustav Eiffel).

Despite new construction, the general outlines of the city remain as they were. It is still a city of balconies and vistas. On 17 of its prominences (many Lisboans profess to see only seven traditional hills, as in Rome) the *miradouros,* which are garden balconies maintained by the municipality, are still frequented by citizens of all ages.

Housing. New housing projects, hotels, and offices have begun to change the city. The pastel-tinted and somewhat sleepy Lisbon that offered a neutral, 19th-century haven to 200,000 war refugees in the 1940s has disappeared in the din and dust of new construction. Lisbon has emerged as a bustling modern metropolis.

Pombal's Baixa remains rigorously protected from change, but the four-story buildings of the Avenida da Liberdade and its ancillary streets have been almost totally replaced by 10-story buildings in a bland modern style. The new construction has also gained the hills, even the Alfama, whose established residents continue to hang laundry across the narrow streets and to grill sardines on doorstep braziers.

The municipality has built new neighbourhoods in the northern and northwestern sectors of town. Other developments have pushed westward toward Belém as a replacement for decayed neighbourhoods. These modern structures, some of them 14 stories high, are designed by Lisbon architects, who produce handsome, colourful, contemporary buildings. Most of the new lodgings are reserved for the poorest families (a certain number pay nothing at all) and for people of moderate means. The Chelas District project, implanted on heathland previously considered too difficult to build upon, houses about 10 percent of the Lisboan population. Despite these projects, adequate housing remains a problem, and a number of shantytowns have developed on the periphery of the city. Many affluent families have moved out of town, and more and more villas have appeared in the countryside of the "Portuguese Riviera," which lies between Lisbon and the town of Estoril, 16 miles to the west.

New housing

THE PEOPLE

Although the district of Lisbon occupies only about 3 percent of Portugal's total area, it contains more than 20 percent of the country's population. Similarly, the city proper occupies approximately 3 percent of Lisbon district, but roughly 40 percent of the district's residents live in the city. The area has long been a magnet for rural immigration. The population is predominantly Portuguese, but there are some foreign residents, mostly diplomats and merchants in the import-export business. Lisbon, judged by net median family income, is the capital of western Europe's poorest people.

For centuries the Lisboans have discussed the symptoms of an affliction said to be endemic in this strip of the Iberian Peninsula: *saudade,* generally translated as "melancholy," a variety of a state of anxiety tempered by fatalism. *Saudade* is said to be reflected in the fado, which is sung in its original form only in Lisbon, and in Lisbon only in the two hillside precincts of Alfama and Bairro Alto. The word fado means "fate," usually an unkind fate described in the songs that are throbbingly sung to a penetrating but melodic music.

Saudade

Portugal is an essentially Roman Catholic country, and Lisbon is distinguished as one of the three places in the world whose chief Roman Catholic clergyman bears the title of patriarch. The Lisboans are typically less devout than northern Portuguese, however, using the church mainly for family occasions: christenings, weddings, and funerals. Religious processions are generally subdued affairs, without the colour and the drama found in Spain. The June feasts of the popular saints (St. Anthony, St. John, and St. Peter) are exceptions. The Lisboans celebrate by donning imaginative costumes, jumping over bonfires, and dancing in the streets until dawn. Indeed, these lively events, held in the small quarters near the Castle of St. George, retain all the pagan elements of a midsummer festival.

THE ECONOMY

Industry. To the long-established local industries of soapmaking, munitions, and steel manufacture have been added glassmaking, electronics, margarine manufacture, and diamond cutting. The petroleum refinery, largely state supported, has also been expanded.

The greatest development in recent years, however, has come just outside the city limits on the south bank of the Tagus; the district has become Portugal's most important manufacturing centre. The industrial belt has continued to grow from the river down to Setúbal, which is 25 miles south. One of the world's largest cement plants is found on the far bank, along with grain elevators, a steelmaking complex, a cork factory, and a plastics plant.

The port of Lisbon, with 19 miles of docks, now has

special facilities for the handling of container-ship cargoes and for car ferries, and it continues to undergo expansion. Cotton, grain, and coal are important imports.

The role of tourism

Commerce and finance. Tourism and commerce have played a major part in Lisbon's modernization, and revenues from tourism have helped offset usually negative national trade balances. With the increase in population and the spread of hotels, offices, and apartment blocks throughout the city, banks have proliferated. Shops are concentrated around the Rua Garrett and the Baixa, though good shopping centres have grown up in the new residential districts toward the airport. The city's major market is located in the riverside square near the Cais do Sodré railway station.

The commercial docks, situated in Alcântara, west of Cidade Baixa, have refrigeration plants located close by to handle the catches of sardines and tunny on which the trawler fleet's livelihood has long been based.

Transportation. Lisbon is connected by rail and road to the interior of Portugal and to the rest of Europe. The airport, at Portela de Sacavém, some six miles beyond the city, has flights to Europe, the Americas, Africa, and the Middle East. The 25th of April Bridge has been the main roadway into the city since it was built in the mid-1960s. The first bridge in Lisbon's long history to span the Tagus, it is more than 7,470 feet (2,277 metres) from anchorage to anchorage, with a central span of about 3,323 feet suspended 230 feet above mean water level. There is space under the roadway to carry two railroad tracks.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. As the capital of Portugal, Lisbon and its surrounding suburbs house all of the principal institutions of the republic. Many government offices occupy the 18th-century buildings that surround Commerce Square. The nation's parliamentary body meets at the Palace of the National Assembly (Palácio da Assembléia Nacional), which also houses the National Archives. Located on the west side of the Bairro Alto, the 17th-century buildings were originally occupied by the convent of São Bento da Saúde. Farther west, toward Belém, the Necessidades Palace (Palácio das Necessidades) houses the Ministry of Foreign Affairs. Built in the mid-18th century on the site of a chapel (for which it is named), it was, until 1910, a royal palace. Another royal palace, built in Belém in 1700, is now the official residence of the president of Portugal.

Like Portugal's 21 other administrative districts, the district of Lisbon is headed by a civil governor who is appointed by and responsible to the central government through the Minister of the Interior. The district is divided into municipalities (*concelhos*), which are further subdivided into wards (*bairros*) and parishes (*freguesias*). Representatives to parish assemblies are chosen through local elections. They then elect an executive body, the parish committee. Lisbon's Municipal Assembly consists of representatives chosen by their parish committees and members directly elected by the local citizens. It serves as the legislative branch of local government and elects the executive branch, the Municipal Chamber, which is headed by a president appointed by the district governor.

Municipal services

Services under the jurisdiction of the municipality include the city waterworks, road maintenance, and sanitation. The money for these services is provided by grants from the central government and through local taxes.

Health. As a major urban centre, Lisbon has a higher percentage of doctors and other health professionals than the rest of Portugal. The city's hospitals include state, private, and military establishments. There are no hospitals run by the municipality. State hospitals are handicapped by the scarcity of public funding. As a result they are insufficiently staffed and, in some cases, inadequately equipped to cope with the demands put on their services. English, French, and Hebrew hospitals cater largely to their small, respective communities.

Education. Officially, education in Portugal is free and compulsory for children ages six through 14, but the number of public schools in Lisbon is insufficient for the city's population. Private schools fill the gap to a certain extent. The University of Lisbon was founded in 1290 and remained Portugal's only university until the 16th century. It was moved back and forth between Lisbon and Coimbra until 1537, when it was renamed and permanently located in Coimbra. Thus, Lisbon was left without a university until 1911, when a second University of Lisbon was established. The Technical University of Lisbon was founded in 1931, and three more universities were opened in the city during the 1970s. Despite this effort to expand educational opportunities, the number of university applicants far exceeds the number of available places.

CULTURAL LIFE

Lisbon's Calouste Gulbenkian Foundation and Museum, a cultural centre named for its benefactor, an Armenian oil-lease negotiator, presents music and ballet as well as exhibits of the other fine arts. It also houses the broad-ranging personal collection that Calouste Gulbenkian, who lived in Lisbon from 1942 until his death in 1955, willed to the Portuguese nation.

The city has more than a dozen other museums, including those of modern, antique, sacred, decorative, and folk arts. In the latter are found many beautiful native artifacts. Two specialized, rather unusual museums are the Azulejo Museum and the Coach Museum. The former, located in the convent of Madre de Deus, boasts a large and varied collection of the painted tiles (*azulejos*) for which the Iberian Peninsula is famous. The Coach Museum occupies a wing of the Portuguese president's official residence and contains an impressive display of carved and gilded coaches.

Lisbon's municipal orchestra was founded in 1971. The National Conservatory offers advanced instruction in both music and drama. The city has two principal theatres, the St. Charles (São Carlos) and the National Theatre of Dona Maria II. The St. Charles, which was constructed in the late 18th century, has a beautiful elliptical interior, and the National, which was built about 1845, displays a facade of six giant columns saved from the convent church of St. Francisco, which was destroyed by earthquake. The interior, gutted by fire in 1966, has been restored.

Theatres and theatrical churches

Neither of these edifices is as theatrical as the interiors of some of the churches built or restored after the 1755 earthquake. In gold, marble, carved wood, and rare tiles, these interiors are decorated in Baroque, Rococo, or rocaille. One outstanding example is the 16th-century church of São Roque, whose unpretentious exterior belies its opulent collection of *azulejos,* paintings, and mosaics inlaid with semiprecious stones.

The old, red brick bullring with its Moorish arches and cupolas still finds an audience for its spectacles. In Portuguese bullfights the bull's first opponent is either a *cavaleiro* on horseback or a *toureiro* on foot. The performance of either man is judged by his dexterity, courage, and proximity to the bull. Next, a squad of acrobatic men, the *forcados,* wrestle the bull, eventually immobilizing it with their bare hands. The bull is not slain.

Lisbon has several other sports and recreational areas. Residents can travel several miles north or west to one of the three major football (soccer) stadiums. Many of the housing developments are planted with trees and grass, their small parks adding to Lisbon's collection of more than 40 public gardens. The largest public park, Monsanto, covers about eight square miles and has numerous recreational facilities. Its rolling hills were planted in the 1920s to provide a windbreak for the town and are now thickly forested. There are also two botanical gardens and a zoological garden within the city.

History

THE EARLY PERIOD

Prehistoric to Moorish times. The valley in which the heart of Lisbon now lies was, in prehistoric times, the bed of a forked branch of the Tagus. (The subway now forks at the same spot.) No evidence has been uncovered to show who were the first residents on the hills surrounding the valley. Although it seems likely that the city was founded *c.* 1200 BC as a trading station by the far-ranging Phoenicians, there is no unassailable proof of the story.

Origins of the city

The city's ancient name, Olisipo, may be derived from the Phoenician *alis ubbo* ("delightful little port") or from the legend that the city's founder was Ulysses.

Whatever the city's origins, it is known that the area was under Roman domination from 205 BC to *c.* AD 409 and that Julius Caesar raised the settlement to the dignity of a *municipium* and named it Felicitas Julia. A few inscribed stones remain as evidence of the Roman presence. The Romans lost the city to the migratory peoples known as the Alani, who were driven out by the Suebi, who in turn were conquered by the Visigoths. The base plan of the original fortifications is thought to be Visigothic and, if so, is the sole vestige of their reign.

The Muslims of North Africa took Lisbon when they overran the Iberian Peninsula in the 8th century; they stayed for 433 years, despite incursions by the Normans in 844 and by Alfonso VI of Castile and León in 1093. Under the Moors the city was known under variations of "Lisbon": Luzbona, Lixbuna, Ulixbone, and Olissibona. Some authorities contend that the Muslims took this name from the conquered Roman castle, but Lisbon historians suggest that it derives from *água boa* ("good water").

The Portuguese conquest. Behind their walls, the Moors were able to hold out for months when the city was assailed by crusader forces—English, Flemish, Norman, and Portuguese under Afonso Henriques, the Portuguese king. The city finally fell in 1147 and successfully resisted Moorish attempts to win it back. The Moorish alcazar was transformed into a Portuguese royal palace, and, according to legend, the Lisbon Cathedral (Sé Patriarcal) was converted from a mosque (with subsequent restorations in the styles of many periods after fires and earthquakes). There is no evidence, however, of a building on the site of the cathedral before the time of Afonso Henriques.

Although 1,400 years of occupation and invasion have left almost no trace among the stones of the capital, the presence of the outlander is still visible in the faces of the inhabitants, which range in cast from the Scandinavian to the Mauritanian.

After winning Lisbon, King Afonso established his court 105 miles to the north-northeast, atop a cliff at Coimbra. Lisbon did not become the national capital until more than a century later, in 1256. Within its Moorish walls, of which some traces still remain, medieval Lisbon measured 1,443 feet at its widest and 1,984 feet at its longest, descending the hill below the castle. Even before the Portuguese conquest, some houses had already been built outside the walls toward the river. The site of this first Lisbon is occupied by the lively Alfama quarter, which has kept the labyrinthine medieval street plan.

The medieval core

King Dinis I (1279–1325) decreed that Portuguese, the dialect of the Porto region, was to be the national language. He founded the university in Lisbon in 1290, and during his reign, other hilltops around the central valley were crowned with convents and churches.

In 1372–73 Lisbon was besieged and burned by the Castilians, who forced King Ferdinand I, an unsuccessful contender for the Castilian throne, to repudiate his alliance with England; thereafter the King swiftly erected new defenses. His wall—more than three miles long, with 77 towers and 38 gates and enclosing more than 247 acres—withstood the renewed Castilian attack of 1384, which followed Ferdinand's death.

The Age of Discovery. When the Portuguese Age of Discovery (1415–1578) began, a census of Lisbon showed 65,000 inhabitants occupying 23 parishes. A considerable number of these residents became rich, and the city was endowed with larger and more luxurious buildings. African slaves became a familiar Lisbon sight, the trade in slaves being one in which Portugal played a major role. After the great explorer Vasco da Gama led a Portuguese fleet to India in 1498, the Venetian monopoly on Oriental trade was broken; and colonies of German, Flemish, Dutch, English, and French traders established themselves in Lisbon. Greeks, Lombards, and Genoese who had lost their trading enclaves in Constantinople when that city fell to the Turks in 1453 also came to Lisbon.

King Manuel I (1495–1521) dominated this epoch, and under his rule Portugal developed its sole contribution to European architecture, an extreme style of late Gothic decoration that celebrated the voyages of discovery, Manuel, and God. The prime examples of Manueline style at Lisbon, the Tower of Belém and the Jerónimos Monastery, about four miles downstream from the city centre, are far less exuberant than those at the rival Portuguese cities of Batalha and Tomar. The tower and the monastery are nevertheless the most important architectural monuments in the Lisbon area. The five-story Tower of Belém, located on the riverbank, was built in 1515 as a fort in the middle of the Tagus, which subsequently altered course. Girt by a cable carved in the stone, it has a stern Gothic interior but exhibits its North African touches on its turrets and crenellations and presents rounded Renaissance arches for the windows. The monastery with its church and cloisters was begun in 1502 by Boytac (Boitaca), an architect of French origin, and was not finished until the end of the century. Four other architects worked on the project, their styles passing from the Gothic through the Renaissance to the Baroque. Smoothed by time, the ensemble is harmonious and proudly Portuguese.

Manuel I promoted the urbanization of the central valley between Lisbon's hills, creating a city square, the Rossio, which at once became a popular meeting place. By the Tagus he constructed a new palace, the Paços da Ribeira, with a large square laid out along its eastern flank. The area between the Rossio and the Palace Terrace (Terreiro do Paço) was soon crisscrossed with streets, along which rose the new shops, churches, and hospitals of what had become a phenomenally prosperous city.

The prosperity was chimerical, however. John (João) the Pious, who had succeeded Manuel, permanently transferred (1537) the university to the royal palace at Coimbra, far from the capital's excesses. He also invited the Jesuits and the Inquisition to come to Portugal to counter the ungodly materialism of Lisbon. The Inquisition office, located in the Rossio, was particularly ferocious in its persecution of the Jews, who were the bankers, financiers, and moneylenders of the time. Many wealthy Jews had their property and goods confiscated; some emigrated to Holland and other countries, taking their money and financial expertise with them. As a result, Lisbon's connections with foreign markets were disrupted and the country's economy suffered severe financial constraints.

Decline in the 16th century

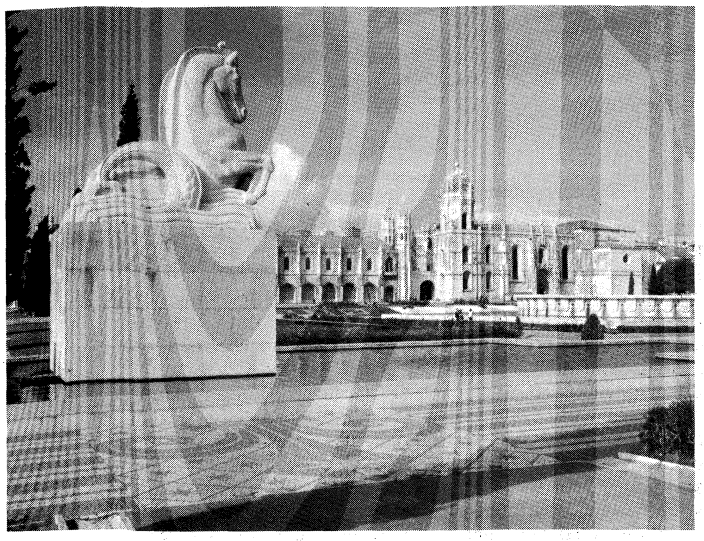
Lisbon was visited with plagues and earthquakes during this time, but they proved easier to meet than the cost of 50 years of glory. Literally half the nation's population had vanished in pursuit of wealth in the new colonies. With farms deserted, food was imported from other European countries at crippling prices, and with so many skilled men absent, wages rose sharply, as did the cost of building and manufacturing materials. The colonial treasures, which had made Lisbon such a sybaritic queen of the seas, in the end cost more than they could fetch.

In 1578 King Sebastian of Portugal was killed in a disastrous invasion of Morocco: two years later, the Spanish pushed into Portugal, and Philip II of Spain became king of both countries. In 1588 it was from Lisbon that the Invincible Armada sailed against England, Portugal's oldest ally. In the half century that followed, Lisbon lived relatively well as a port for the riches of the Spanish Main. In 1640 a conspiracy of Lisbon nobles struck for freedom and drove out the Spaniards, restoring Portugal's independence. The square just north of the Rossio, Restoration Square (Praça dos Restauradores), is named for them.

With the Cromwellian treaty of 1654, following British military assistance to the Portuguese in the war with Spain, the British merchants trading and living in Lisbon set up a corporation, which became known as the British Factory. The Factory negotiated with the Portuguese government for trade concessions and other privileges, appealing to the British government to put pressure on the Portuguese authorities when necessary. Britain's economic and political influence on Portugal was strong, and the Factory remained in existence until 1810.

EVOLUTION OF THE MODERN CITY

Disaster and reconstruction. In the first half of the 18th century the profits from the plantations and the gold



The Jerónimos Monastery in Lisbon.
Colour Library International

and diamond deposits of Brazil brought a new flurry of optimism and excitement to Lisbon. This ended on the morning of Nov. 1, 1755. The churches were crowded to honour the dead on All Saints' Day when the city was devastated by one of the greatest earthquakes ever recorded. There were two shocks, 40 minutes apart; and the waters of the Tagus, lifted from their bed, roared through the city, followed by fire. It is believed that 30,000 lives were lost, and more than 9,000 buildings were destroyed.

The great earthquake of 1755

Physically, Lisbon recovered with a celerity astonishing for the time, but the shock left its mark upon the thinking of generations to come. The reconstruction—a good deal of foreign aid was forthcoming—was achieved by Joseph I's prime minister, Sebastião José de Carvalho, the virtual ruler of the nation. He put Manuel da Maia, engineer in chief of the realm, in charge of five architects and soon had a plan for remaking the totally devastated centre of the town, the Baixa. The riverside palace had been destroyed, and its terrace was expanded to create the new Commerce Square. Northward from there, a grid of 48 streets led inland to the Rossio and a neighbouring new square, Figueira. The two-story, uniform buildings were topped by two tiers of dormers projecting from tiled roofs. The corners of the eaves, in the Lisbon tradition, turned up, in faint echo of a pagoda. The building style, evolved for fast, cheap construction, was Baroque but virtually stripped of decoration. After the minister was rewarded with the title of marquês de Pombal, the style became known as the *estilo pombalino.*

The Sé and most of the churches were repaired or rebuilt, but the 14th-century Carmel (Carmo) Church was left as it was. Looming from its hilltops over the Baixa, the roofless Gothic shell now serves as an archaeological museum, while its cloister serves as the barracks for the National Republican Guard, a paramilitary security force. The Palace of the Inquisition, utterly flattened, was not rebuilt when Pombal enlarged and realigned the Rossio, and on its site, 90 years later, the National Theatre of Dona Maria II was erected. Pombal banished the Jesuit order and transformed their establishment into St. Joseph's Hospital to replace the destroyed All Saints Hospital. The medical school scrambled for room at St. Joseph's until it acquired a new building of its own late in the 19th century. The Jesuit novice house was converted to serve as the Nobles' School. Later governments expelled more religious orders, whose buildings became barracks, hospitals, royal academies, and government offices.

19th-century expansion. During the Peninsular War of the early 1800s, Lisbon alternated between French and British control, and after Napoleon's defeat it was embroiled in civil war until 1834. That conflict was followed by 10 years of revolutionary outbursts. Nineteenth-century Lisbon nevertheless continued to expand and, by 1885, embraced 20,378 acres, while the population had doubled in 100 years to reach 300,000. Public buildings, such as the new city hall and the Ajuda Royal Palace, had been built, and the harbour had been modernized and quays constructed on land reclaimed from the river. The railway had appeared, and a system of horse cars served the lower town.

The greatest change in the city, and the one most important for its future growth, was the opening of a new main street in 1880—the Avenida da Liberdade. The municipality bordered the central six-lane carriageway with wide blue mosaic sidewalks graced with palms and shade trees, fountains, and ornamental waters stocked with goldfish and swans. So the street remains today, with the addition of outdoor cafés beneath the trees.

The opening of Avenida da Liberdade

In conjunction with the new thoroughfare, a series of new streets, the "Avenidas Novas," expanded the city northward, and new neighbourhoods sprang up as well on the borders of the Avenida da Liberdade. In 1901 the electric streetcar made its appearance, enabling more people to live farther away from their employment in the Baixa. Three cable cars shuttled up and down the adjacent hills, and Eiffel designed the giant elevator that hisses grandly between the town's upper and lower levels.

New water supplies, augmenting those of the 1748 aqueduct of Águas Livres, were introduced from Alviela. Consequently, water was piped directly into houses, eliminating the ancient calling of *galego,* or water porter.

The 20th century. In 1908 Portugal's king and crown prince were assassinated on the northwest corner of Commerce Square. Two years later the new king, Manuel II, abdicated. A republic was declared, and a period of national instability ensued. When António Salazar took control of the near-bankrupt nation in 1932, he erected a corporate state of which he alone determined the policies until his retirement in 1968. There was considerable development in Lisbon throughout this time, and during World Wars I and II the city was able to offer refuge to some 200,000 foreigners.

In the 1960s national policy began to change, allowing economic expansion. The 30-year-old austerity program of stability and self-sufficiency (at an admittedly low level of investment and consumption) was somewhat softened, and international tourists and international corporations began to be accommodated. In 1966, well ahead of schedule (and with $77,000,000 in aid from the United States), the Salazar (now the 25th of April) Bridge was completed.

An era of expansion

On April 25, 1974, the government of Marcelo Caetano, Salazar's successor, was overthrown by a military coup. By the early 1980s, however, political instability and economic difficulties remained serious problems and hindered the nation's, and the city's, efforts to bring about social and economic reforms. The restriction of government funds to the municipality of Lisbon led to a bitter struggle within the city council, where resistance mounted to speculative building projects that would affect the environment in historical parts of the town.

BIBLIOGRAPHY

General: Fodor's Lisbon, 1986 (1986), provides general descriptive information. DAVID WRIGHT and PATRICK SWIFT, *Lisbon: A Portrait and a Guide* (1971), gives thorough coverage of all quarters of the city, its history, monuments, cultural institutions, and contemporary life, and also includes excursions outside Lisbon. Also see CAROL WRIGHT, *Lisbon* (1971), on the city and the life of its inhabitants at various times of day, as well as information for the visitor to Lisbon and its environs; and VIVIAN ROWE, *The Road to Lisbon* (1962), mainly concerned with the journey from France to Lisbon but including some material on the attractions of the city itself.

History: JÚLIO DE CASTILHO, *Lisboa Antiga: O Bairro Alto de Lisboa,* 2nd ed., 5 vol. (1902–04), and *Lisboa Antiga: Bairros Orientais,* 2nd ed., 12 vol. (1934–38); THOMAS D. KENDRICK, *The Lisbon Earthquake* (1956), a study of the 1755 earthquake and its impact on European philosophers and theologians; JANET SCHAW, *Journal of a Lady of Quality: Being the Narrative of a Journey from Scotland to the West Indies, North Carolina, and Portugal, in the Years 1774–1776,* 3rd ed. edited by EVANGELINE WALKER ANDREWS and CHARLES MCLEAN ANDREWS (1939, reissued 1971); and ROSE MACAULAY, *They Went to Portugal* (1946, reissued 1985), on the British in Portugal.

(B.E./L. de S.R.)

The Art of Literature

The current meaning generally attached to the term literature—a body of writing by a people or by peoples using the same language—is a relatively modern one. The term itself, derived from the Latin word *littera* ("letter of the alphabet"; *litterae,* "letters"), is ancient enough; but in ancient times literature tended to be considered separately in terms of kinds of writing, or genres as they came to be called in the 18th century when the term literature took on its modern meaning. Thus Aristotle's *Poetics,* though it is concerned with and gives examples from Greek epic and dithyrambic poetry and comedy, has as its central concern tragedy.

The sections that make up this article are arranged roughly in a chronological order according to the mediums of verse and prose with attention, again mainly chronological, to various genres. No such classification is totally satisfactory. Furthermore, although the derivation of the word literature implies writing, there is much oral literature, a general treatment of which may be found in the article FOLK ARTS: *Literature.*

There is a further complication: literature as a whole and in its parts means varying things to various writers, critics, and historians. At one extreme, it may be held that anything written is literature. Though this position is seldom held, that at the other extreme—literature is only the *Iliad,* the *Odyssey,* and *Hamlet*—is slightly more popularly held. Between these extremes, attitudes vary widely. For some critics, a hierarchy exists: tragedy is superior to comedy; the short story is inferior to the novel. For other critics, qualitative criteria apply: poetry is verse that succeeds; the limerick and nonsense verse are failed poetry. Critics also differ on the purpose or ends of literature. Many ancient critics—and some modern ones—hold that the true ends of literature are to instruct and delight. Others—a majority of the modern ones, probably—hold that pleasure is the sole end. All of these divergences and other similar ones appear in the treatments that follow.

For historical treatment of various literatures see the articles LITERATURE, THE HISTORY OF WESTERN; and sections in the articles AFRICAN ARTS, CENTRAL ASIAN ARTS, OCEANIC ARTS, and SOUTHEAST ASIAN ARTS. Some literatures are treated separately by language, by nation, or by special subject (*e.g.,* CELTIC LITERATURE; LATIN LITERATURE; FRENCH LITERATURE; JAPANESE LITERATURE; BIBLICAL LITERATURE).

The article is divided into the following sections:

The scope of literature 87
- Literary composition
- Content of literature
- Literature and its audience
- Literature and its environment
- Literature as a collection of genres
- Writings on literature

Poetry 94
- The nature of poetry 94
  - Attempts to define poetry
  - Poetry and prose
  - Form in poetry
  - Poetry as a mode of thought: the Protean encounter
- Prosody 98
  - Elements of prosody
  - Prosodic style
  - Theories of prosody

Narrative fiction 105
- Epic 105
  - General characteristics
  - Bases
  - Early patterns of development
  - Later variations
- Fable, parable, and allegory 110
  - Nature and objectives
  - Historical development in Western culture
  - Allegorical literature in the East
- Ballad 116
  - Elements
  - Composition
  - Types of balladry
  - Subject matter
  - Chronology
- Romance 119
  - The component elements
  - Medieval verse romances
  - Medieval prose romances
  - Later developments
- Saga 123
  - Nonfictional saga literature
  - Legendary and historical fiction
- Novel 125
  - Elements
  - Uses
  - Style
  - Types of novel
  - The novel in English
  - Europe
  - Asia, Africa, Latin America
  - Social and economic aspects
  - Evaluation and study
  - The future of the novel
- Short story 147
  - Analysis of the genre
  - History
  - The 20th century

Drama 152
- Dramatic literature 152
  - General characteristics
  - Drama as an expression of a culture
  - Influences on the dramatist
  - The range of dramatic forms and styles
- Comedy 160
  - Origins and definitions
  - Theories
  - Kinds of comedy in diverse historical periods
  - The comic in other arts
- Tragedy 169
  - Development
  - Tragedy and modern drama
  - Theory of tragedy

Other genres 182
- Satire 182
  - The nature of satire
  - Satirical media
  - The satirist, the law, and society
- Nonfictional prose 185
  - Nature
  - Elements
  - Approaches
  - The essay
  - History
  - Doctrinal, philosophical, and religious prose
  - Political, polemical, and scientific prose
  - Other forms
- Biographical literature 195
  - Aspects
  - Kinds
  - Historical development
  - Biographical literature today
- Literary criticism 203
  - Functions
  - Historical development
  - The 20th century

Children's literature 207
- Definition of terms
- The case for a children's literature
- Some general features and forces
- The development of children's literature
- Historical sketches of the major literatures

THE SCOPE OF LITERATURE

Literature is a form of human expression. But not everything expressed in words—even when organized and written down—is counted as literature. Those writings that are primarily informative—technical, scholarly, journalistic—would be excluded from the rank of literature by most, though not all, critics. Certain forms of writing, however, are universally regarded as belonging to literature as an art. Individual attempts within these forms are said to succeed if they possess something called artistic merit and to fail if they do not. The nature of artistic merit is less easy to define than to recognize. The writer need not even pursue it to attain it. On the contrary, a scientific exposition might be of great literary value and a pedestrian poem of none at all.

Distinction between poetry and verse

The purest (or, at least, the most intense) literary form is the lyric poem, and after it comes elegiac, epic, dramatic, narrative, and expository verse. Most theories of literary criticism base themselves on an analysis of poetry, because the aesthetic problems of literature are there presented in their simplest and purest form. Poetry that fails as literature is not called poetry at all but verse. Many novels—certainly all the world's great novels—are literature, but there are thousands that are not so considered. Most great dramas are considered literature (although the Chinese, possessors of one of the world's greatest dramatic traditions, consider their plays, with few exceptions, to possess no literary merit whatsoever).

The Greeks thought of history as one of the seven arts, inspired by a goddess, the muse Clio. All of the world's classic surveys of history can stand as noble examples of the art of literature, but most historical works and studies today are not written primarily with literary excellence in mind, though they may possess it, as it were, by accident.

The essay was once written deliberately as a piece of literature: its subject matter was of comparatively minor importance. Today most essays are written as expository, informative journalism, although there are still essayists in the great tradition who think of themselves as artists. Now, as in the past, some of the greatest essayists are critics of literature, drama, and the arts.

Some personal documents (autobiographies, diaries, memoirs, and letters) rank among the world's greatest literature. Some examples of this biographical literature were written with posterity in mind, others with no thought of their being read by anyone but the writer. Some are in a highly polished literary style; others, couched in a privately evolved language, win their standing as literature because of their cogency, insight, depth, and scope.

Many works of philosophy are classed as literature. The *Dialogues* of Plato (4th century BC) are written with great narrative skill and in the finest prose; the *Meditations* of the 2nd-century Roman emperor Marcus Aurelius are a collection of apparently random thoughts, and the Greek in which they are written is eccentric. Yet both are classed as literature, while the speculations of other philosophers, ancient and modern, are not. Certain scientific works endure as literature long after their scientific content has become outdated. This is particularly true of books of natural history, where the element of personal observation is of special importance. An excellent example is Gilbert White's *Natural History and Antiquities of Selbourne* (1789).

Role of oratory in literature

Oratory, the art of persuasion, was long considered a great literary art. The oratory of the American Indian, for instance, is famous, while in classical Greece, Polymnia was the muse sacred to poetry and oratory. Rome's great orator Cicero was to have a decisive influence on the development of English prose style. Abraham Lincoln's Gettysburg Address is known to every American schoolchild. Today, however, oratory is more usually thought of as a craft than as an art. Most critics would not admit advertising copywriting, purely commercial fiction, or cinema and television scripts as accepted forms of literary expression, although others would hotly dispute their exclusion. The test in individual cases would seem to be one of enduring satisfaction and, of course, truth. Indeed, it becomes more and more difficult to categorize literature, for in modern civilization words are everywhere. Man is subject to a continuous flood of communication. Most of it is fugitive, but here and there—in high-level journalism, in television, in the cinema, in commercial fiction, in westerns and detective stories, and in plain, expository prose—some writing, almost by accident, achieves an aesthetic satisfaction, a depth and relevance that entitle it to stand with other examples of the art of literature.

LITERARY COMPOSITION

Critical theories. *Western.* If the early Egyptians or Sumerians had critical theories about the writing of literature, these have not survived. From the time of classical Greece until the present day, however, Western criticism has been dominated by two opposing theories of the literary art, which might conveniently be called the expressive and constructive theories of composition.

The Greek philosopher and scholar Aristotle (384–322 BC) is the first great representative of the constructive school of thought. His *Poetics* (the surviving fragment of which is limited to an analysis of tragedy and epic poetry) has sometimes been dismissed as a recipe book for the writing of potboilers. Certainly, Aristotle is primarily interested in the theoretical construction of tragedy, much as an architect might analyze the construction of a temple, but he is not exclusively objective and matter of fact. He does, however, regard the expressive elements in literature as of secondary importance, and the terms he uses to describe them have been open to interpretation and a matter of controversy ever since.

The 1st-century Greek treatise *On the Sublime* (conventionally attributed to the 3rd-century Longinus) deals with the question left unanswered by Aristotle—what makes great literature "great"? Its standards are almost entirely expressive. Where Aristotle is analytical and states general principles, the pseudo-Longinus is more specific and gives many quotations: even so, his critical theories are confined largely to impressionistic generalities.

Constructive and expressive theories of criticism

Thus, at the beginning of Western literary criticism, the controversy already exists. Is the artist or writer a technician, like a cook or an engineer, who designs and constructs a sort of machine that will elicit an aesthetic response from his audience? Or is he a virtuoso who above all else expresses himself and, because he gives voice to the deepest realities of his own personality, generates a response from his readers because they admit some profound identification with him? This antithesis endures throughout western European history—Scholasticism versus Humanism, Classicism versus Romanticism, Cubism versus Expressionism—and survives to this day in the common judgment of our contemporary artists and writers. It is surprising how few critics have declared that the antithesis is unreal, that a work of literary or plastic art is at once constructive and expressive, and that it must in fact be both.

Eastern. Critical theories of literature in the Orient, however, have been more varied. There is an immense amount of highly technical, critical literature in India. Some works are recipe books, vast collections of tropes and stylistic devices; others are philosophical and general. In the best period of Indian literature, the cultural climax of Sanskrit (*c.* 320–490), it is assumed by writers that expressive and constructive factors are twin aspects of one reality. The same could be said of the Chinese, whose literary manuals and books on prosody and rhetoric are, as with the West, relegated to the class of technical handbooks, while their literary criticism is concerned rather with subjective, expressive factors—and so aligns itself with the pseudo-Longinus' "sublime." In Japan, technical, stylistic elements are certainly important (Japanese discrimination in these matters is perhaps the most refined in the world), but both writer and reader above all seek qualities of subtlety and poignancy and look for intimations

of profundity often so evanescent as to escape entirely the uninitiated reader.

Broad and narrow conceptions of poetry. Far Eastern literary tradition has raised the question of the broad and narrow definitions of poetry (a question familiar in the West from Edgar Allan Poe's advocacy of the short poem in his "Poetic Principle" [1850]). There are no long epic poems in Chinese, no verse novels of the sort written in England by Robert Browning or Alfred Lord Tennyson in the 19th century. In Chinese drama, apart from a very few of the songs, the verse as such is considered doggerel. The versified treatises on astronomy, agriculture, or fishing, of the sort written in Greek and Roman times and during the 18th century in the West, are almost unknown in the Far East. Chinese poetry is almost exclusively lyric, meditative, and elegiac, and rarely does any poem exceed 100 lines—most are little longer than Western sonnets; many are only quatrains. In Japan this tendency to limit length was carried even further. The ballad survives in folk poetry, as it did in China, but the "long poem" of very moderate length disappeared early from literature. For the Japanese, the tanka is a "long poem": in its common form it has 31 syllables; the *sedōka* has 38; the *dodoitsu,* imitating folk song, has 26. From the 17th century and onward, the most popular poetic form was the haiku, which has only 17 syllables.

This development is relevant to the West because it spotlights the ever-increasing emphasis which has been laid on intensity of communication, a characteristic of Western poetry (and of literature generally) as it has evolved since the late 19th century. In the Far East all cultivated people were supposed to be able to write suitable occasional poetry, and so those qualities that distinguished a poem from the mass consequently came to be valued above all others. Similarly, as modern readers in the West struggle with a "communication avalanche" of words, they seek in literature those forms, ideas, values, vicarious experiences, and styles that transcend the verbiage to be had on every hand.

Literary language. In some literatures (notably classical Chinese, Old Norse, Old Irish), the language employed is quite different from that spoken or used in ordinary writing. This marks off the reading of literature as a special experience. In the Western tradition, it is only in comparatively modern times that literature has been written in the common speech of cultivated men. The Elizabethans did not talk like Shakespeare nor 18th-century people in the stately prose of Samuel Johnson or Edward Gibbon (the so-called Augustan plain style in literature became popular in the late 17th century and flourished throughout the 18th, but it was really a special form of rhetoric with antecedent models in Greek and Latin). The first person to write major works of literature in the ordinary English language of the educated man was Daniel Defoe (1660?–1731), and it is remarkable how little the language has changed since. *Robinson Crusoe* (1719) is much more contemporary in tone than the elaborate prose of 19th-century writers like Thomas De Quincey or Walter Pater. (Defoe's language is not, in fact, so very simple: simplicity is itself one form of artifice.)

Ambiguity. Other writers have sought to use language for its most subtle and complex effects and have deliberately cultivated the ambiguity inherent in the multiple or shaded meanings of words. Between the two world wars, "ambiguity" became very fashionable in English and American poetry and the ferreting out of ambiguities—from even the simplest poem—was a favourite critical sport. T.S. Eliot in his literary essays is usually considered the founder of this movement. Actually, the platform of his critical attitudes is largely moral, but his two disciples, I.A. Richards in *Principles of Literary Criticism* (1924) and William Empson in *Seven Types of Ambiguity* (1930), carried his method to extreme lengths. The basic document of the movement is C.K. Ogden and I.A. Richards' *The Meaning of Meaning* (1923), a work of enormous importance in its time. Only a generation later, however, their ideas were somewhat at a discount.

Translation. Certainly, William Blake or Thomas Campion, when they were writing their simple lyrics, were unaware of the ambiguities and multiple meanings that future critics would find in them. Nevertheless, language is complex. Words do have overtones; they do stir up complicated reverberations in the mind that are ignored in their dictionary definitions. Great stylists, and most especially great poets, work with at least a half-conscious, or subliminal, awareness of the infinite potentialities of language. This is one reason why the essence of most poetry and great prose is so resistant to translation (quite apart from the radically different sound patterns that are created in other-language versions). The translator must project himself into the mind of the original author; he must transport himself into an entirely different world of relationships between sounds and meanings, and at the same time he must establish an equivalence between one infinitely complex system and another. Since no two languages are truly equivalent in anything except the simplest terms, this is a most difficult accomplishment. Certain writers are exceptionally difficult to translate. There are no satisfactory English versions, for example, of the Latin of Catullus, the French of Baudelaire, the Russian of Pushkin, or of the majority of Persian and Arabic poetry. The splendour of Sophocles' Greek, of Plato at his best, is barely suggested even in the finest English versions. On the other hand, the Germans insist that Shakespeare is better in German than he is in English, a humorous exaggeration perhaps. But again, Shakespeare is resistant to translation into French. His English seems to lack equivalents in that language.

The very greatest translations may become classics in their own right, of enduring literary excellence (the King James Version of the Bible, appearing in 1611, is an outstanding example), but on the whole the approximate equivalence of most translations to their originals seems to have a very short life. The original work remains the same, of lasting value to its own people, but the translation becomes out of date with each succeeding generation as the language and criteria of literary taste change. Nothing demonstrates the complexity of literary language more vividly. An analogous process takes place when a reader experiences a literary work in his own language; each generation gets a "new version" from its own classics.

Yet the values of great literature are more fundamental than complexity and subtleties of meaning arising from language alone. Works far removed from contemporary man in time and in cultural background, composed in a variety of languages utterly different from one another in structure, have nevertheless been translated successfully enough to be deeply moving. The 20th century has seen an immense mass of the oral literature of preliterate peoples and of the writings of all the great civilizations translated into modern languages. Understanding the growth of literature and its forms in other civilizations has greatly enriched the understanding of our own.

Craftsmanship. *Prosody.* Literature, like music, is an art of time, or "tempo": it takes time to read or listen to, and it usually presents events or the development of ideas or the succession of images or all these together in time. The craft of literature, indeed, can be said to be in part the manipulation of a structure in time, and so the simplest element of marking time, rhythm, is therefore of basic importance in both poetry and prose. Prosody, which is the science of versification, has for its subject the materials of poetry and is concerned almost entirely with the laws of metre, or rhythm in the narrowest sense. It deals with the patterning of sound in time; the number, length, accent, and pitch of syllables; and the modifications of rhythm by vowels and consonants. In most poetry, certain basic rhythms are repeated with modifications (that is to say, the poem rhymes or scans or both) but not in all. It most obviously does neither in the case of the "free forms" of modern poetry; but neither does it in the entire poetry of whole cultures. Since lyric poetry is either the actual text of song or else is immediately derived from song, it is regular in structure nearly everywhere in the world, although the elements of patterning that go into producing its rhythm may vary. The most important of these elements in English poetry, for example, have been accent, grouping of syllables (called feet), number of syllables in the line,

Elements of prosody

and rhyme at the end of a line (and sometimes within it). Other elements such as pitch, resonance, repetition of vowels (assonance), repetition of consonants (alliteration), and breath pauses (cadence) have also been of great importance in distinguishing successful poetry from doggerel verse, but on the whole they are not as important as the former, and poets have not always been fully conscious of their use of them. Greek and Latin poetry was consciously patterned on the length of syllables (long or short) rather than on their accent; but all the considerations of "sound" (such as assonance and alliteration) entered into the aesthetically satisfactory structure of a poem. Similarly, both the French and Japanese were content simply to count the syllables in a line—but again, they also looked to all the "sound" elements.

The rhythms of prose are more complicated, though not necessarily more complex, than those of poetry. The rules of prose patterning are less fixed; patterns evolve and shift indefinitely and are seldom repeated except for special emphasis. So the analysis of prose rhythm is more difficult to make than, at least, the superficial analysis of poetry.

Structure. The craft of writing involves more than mere rules of prosody. The work's structure must be manipulated to attract the reader. First, the literary situation has to be established. The reader must be directly related to the work, placed in it—given enough information on who, what, when, or why—so that his attention is caught and held (or, on the other hand, he must be deliberately mystified, to the same end).

Aristotle gave a formula for dramatic structure that can be generalized to apply to most literature: presentation, development, complication, crisis, and resolution. Even lyric poems can possess plot in this sense, but by no means are all literary works so structured, nor does such structure ensure their merit—it can be safely said that westerns, detective stories, and cheap melodramas are more likely to follow strictly the rules of Aristotle's *Poetics* than are great novels. Nevertheless, the scheme does provide a norm from which there is infinite variation. Neoclassical dramatists and critics, especially in 17th-century France, derived from Aristotle what they called the unities of time, action, and place. This meant that the action of a play should not spread beyond the events of one day and, best of all, should be confined within the actual time of performance. Nor should the action move about too much from place to place—best only to go from indoors to outdoors and back. There should be only one plot line, which might be relieved by a subplot, usually comic. These three unities—of time, place, and action—do not occur in Aristotle and are certainly not observed in Classical Greek tragedy. They are an invention of Renaissance critics, some of whom went even further, insisting also on what might be called a unity of mood. To this day there are those who, working on this principle, object to Shakespeare's use of comic relief within the tragic action of his plays—to the porter in *Macbeth,* for instance, or the gravediggers in *Hamlet.*

Unities of time, place, and action

Assiduous critics have found elaborate architectural structures in quite diffuse works—including Miguel de Cervantes' *Don Quixote* (1605–15), Sterne's *Tristram Shandy* (1759–67), Casanova's *Icosameron* (1788; 1928). But their "discoveries" are too often put there after the event. Great early novels such as the Chinese *Dream of the Red Chamber* (1754; first published in English 1929) and the Japanese *Tale of Genji* (early 11th century) usually develop organically rather than according to geometrical formulas, one incident or image spinning off another. Probably the most tightly structured work, in the Neoclassicists' sense, is the Icelandic *Njál's saga.*

The 19th century was the golden age of the novel, and most of the more famous examples of the form were systematically plotted, even where the plot structure simply traced the growth in personality of an individual hero or heroine. This kind of novel, of which in their very diverse ways Stendhal's *The Red and the Black* (1830) and Dickens' *David Copperfield* (1850) are great examples, is known as *Bildungsroman.* Gustave Flaubert's *Madame Bovary* (1857) is as rigorously classicist in form as the 17th-century plays of Racine and Corneille, which were the high point of the French classical theatre, although Flaubert obeys laws more complex than those of the Aristotelians. Novels such as Tolstoy's *War and Peace* (1865–69), Dostoyevsky's *Brothers Karamazov* (1880), and the works of Balzac owe much of their power to their ability to overwhelm the reader with a massive sense of reality. The latter 19th and early 20th centuries witnessed an attack on old forms, but what the new writers evolved was simply a new architecture. A novel like James Joyce's *Ulysses* (1922), which takes place in a day and an evening, is one of the most highly structured ever written. Novelists such as Joseph Conrad, Ford Madox Ford, Virginia Woolf, and, to a lesser extent, Henry James developed a multiple-aspect narrative, sometimes by using time shifts and flashbacks and by writing from different points of view, sometimes by using the device (dating back to Classical Greek romances) of having one or more narrators as characters within the story. (This technique, which was first perfected in the verse novels of Robert Browning, in fact reached its most extreme development in the English language in poetry: in Ezra Pound's *Cantos,* T.S. Eliot's *The Waste Land,* William Carlos Williams' *Paterson,* and the many long poems influenced by them.)

Narrative devices in the novel

CONTENT OF LITERATURE

The word as symbol. The content of literature is as limitless as the desire of human beings to communicate with one another. The thousands of years, perhaps hundreds of thousands, since the human species first developed speech have seen built up the almost infinite systems of relationships called languages. A language is not just a collection of words in an unabridged dictionary but the individual and social possession of living human beings, an inexhaustible system of equivalents, of sounds to objects and to one another. Its most primitive elements are those words that express direct experiences of objective reality, and its most sophisticated are concepts on a high level of abstraction. Words are not only equivalent to things, they have varying degrees of equivalence to one another. A symbol, says the dictionary, is something that stands for something else or a sign used to represent something, "as the lion is the symbol of courage, the cross the symbol of Christianity." In this sense all words can be called symbols, but the examples given—the lion and the cross—are really metaphors: that is, symbols that represent a complex of other symbols, and which are generally negotiable in a given society (just as money is a symbol for goods or labour). Eventually a language comes to be, among other things, a huge sea of implicit metaphors, an endless web of interrelated symbols. As literature, especially poetry, grows more and more sophisticated, it begins to manipulate this field of suspended metaphors as a material in itself, often as an end in itself. Thus, there emerge forms of poetry (and prose, too) with endless ramifications of reference, as in Japanese waka and haiku, some ancient Irish and Norse verse, and much of the poetry written in western Europe since the time of Baudelaire that is called modernist. It might be supposed that, at its most extreme, this development would be objective, constructive—aligning it with the critical theories stemming from Aristotle's *Poetics.* On the contrary, it is romantic, subjective art, primarily because the writer handles such material instinctively and subjectively, approaches it as the "collective unconscious," to use the term of the psychologist Carl Jung, rather than with deliberate rationality.

Themes and their sources. By the time literature appears in the development of a culture, the society has already come to share a whole system of stereotypes and archetypes: major symbols standing for the fundamental realities of the human condition, including the kind of symbolic realities that are enshrined in religion and myth. Literature may use such symbols directly, but all great works of literary art are, as it were, original and unique myths. The world's great classics evoke and organize the archetypes of universal human experience. This does not mean, however, that all literature is an endless repetition of a few myths and motives, endlessly retelling the first stories of civilized man, repeating the Sumerian *Epic of Gilgamesh* or Sophocles' *Oedipus the King.* The subject matter of literature is as wide as human experience it-

Use of symbols and myths in literature

self. Myths, legends, and folktales lie at the beginning of literature, and their plots, situations, and allegorical (metaphorical narrative) judgments of life represent a constant source of literary inspiration that never fails. This is so because mankind is constant—men share a common physiology. Even social structures, after the development of cities, remain much alike. Whole civilizations have a life pattern that repeats itself through history. Jung's term "collective unconscious" really means that mankind is one species, with a common fund of general experience. Egyptian scribes, Soviet bureaucrats, and junior executives in New York City live and respond to life in the same ways; the lives of farmers or miners or hunters vary only within narrow limits. Love is love and death is death, for a South African Bushman and a French Surrealist alike. So the themes of literature have at once an infinite variety and an abiding constancy. They can be taken from myth, from history, or from contemporary occurrence, or they can be pure invention (but even if they are invented, they are nonetheless constructed from the constant materials of real experience, no matter how fantastic the invention).

The writer's personal involvement. As time goes on, literature tends to concern itself more and more with the interior meanings of its narrative, with problems of human personality and human relationships. Many novels are fictional, psychological biographies which tell of the slowly achieved integration of the hero's personality or of his disintegration, of the conflict between self-realization and the flow of events and the demands of other people. This can be presented explicitly, where the characters talk about what is going on in their heads; either ambiguously and with reserve, as in the novels of Henry James, or overtly, as in those of Dostoyevsky. Alternatively, it can be presented by a careful arrangement of objective facts, where psychological development is described purely in terms of behaviour, and where the reader's subjective response is elicited by the minute descriptions of physical reality, as in the novels of Stendhal and the greatest Chinese novels like the *Dream of the Red Chamber,* which convince the reader that through the novel he is seeing reality itself, rather than an artfully contrived semblance of reality.

Presentation of characters

Literature, however, is not solely concerned with the concrete, with objective reality, with individual psychology, or with subjective emotion. Some deal with abstract ideas or philosophical conceptions. Much purely abstract writing is considered literature only in the widest sense of the term, and the philosophical works that are ranked as great literature are usually presented with more or less of a sensuous garment. Thus, Plato's *Dialogues* rank as great literature because the philosophical material is presented in dramatic form, as the dialectical outcome of the interchange of ideas between clearly drawn, vital personalities, and because the descriptive passages are of great lyric beauty. Karl Marx's *Das Kapital* (1867–95) approaches great literature in certain passages in which he expresses the social passion he shares with the Hebrew prophets of the Old Testament. Euclid's *Elements* and St. Thomas Aquinas' *Summa theologica* give literary, aesthetic satisfaction to some people because of their purity of style and beauty of architectonic construction. In short, most philosophical works that rank as great literature do so because they are intensely human. The reader responds to Pascal's *Pensées,* to Montaigne's *Essays,* and to Marcus Aurelius' *Meditations* as he would to living men. Sometimes the pretense of purely abstract intellectual rigour is in fact a literary device. The writings of the 20th-century philosopher Ludwig Wittgenstein, for example, owed much of their impact to this approach, while the poetry of Paul Valéry borrows the language of philosophy and science for its rhetorical and evocative power.

Relation of form to content. Throughout literary history, many great critics have pointed out that it is artificial to make a distinction between form and content, except for purposes of analytical discussion. Form determines content. Content determines form. The issue is, indeed, usually only raised at all by those critics who are more interested in politics, religion, or ideology than in literature; thus, they object to writers who they feel sacrifice ideological orthodoxy for formal perfection, message for style.

Style. But style cannot really be said to exist on paper at all; it is the way the mind of the author expresses itself in words. Since words represent ideas, there cannot be abstract literature unless a collection of nonsense syllables can be admitted as literature. Even the most avant-garde writers associated with the Cubist or Non-Objective painters used language, and language is meaning, though the meaning may be incomprehensible. Oscar Wilde and Walter Pater, the great 19th-century exponents of "art for art's sake," were in fact tireless propagandists for their views, which dominate their most flowery prose. It is true that great style depends on the perfect matching of content and form, so that the literary expression perfectly reflects the writer's intention; "poor style" reveals the inability of a writer to match the two—in other words, reveals his inability to express himself. This is why we say that "style expresses the man." The veiled style of Henry James, with its subtleties, equivocations, and qualifications, perfectly reflects his complicated and subtle mind and his abiding awareness of ambiguity in human motives. At the other extreme, the style of the early-20th-century U.S. novelist Theodore Dreiser—bumbling, clumsy, dogged, troubled—perfectly embodies his own attitudes toward life and is, in fact, his constant judgment of his subject matter. Sometimes an author, under the impression that he is simply polishing his style, may completely alter his content. As Flaubert worked over the drafts of *Madame Bovary,* seeking always the apposite word that would precisely convey his meaning, he lifted his novel from a level of sentimental romance to make it one of the great ironic tragedies of literature. Yet, to judge from his correspondence, he seems never to have been completely aware of what he had done, of the severity of his own irony.

Coordination of content and form in good style

Literature may be an art, but writing is a craft, and a craft must be learned. Talent, special ability in the arts, may appear at an early age; the special personality called genius may indeed be born, not made. But skill in matching intention and expression comes with practice. Naïve writers, "naturals" like the 17th-century English diarist Samuel Pepys, the late 18th-century French naïf Restif de la Bretonne, the 20th-century American novelist Henry Miller, are all deservedly called stylists, although their styles are far removed from the deliberate, painstaking practice of a Flaubert or a Turgenev. They wrote spontaneously whatever came into their heads; but they wrote constantly, voluminously, and were, by their own standards, skilled practitioners.

Objective–subjective expression. There are certain forms of literature that do not permit such highly personal behaviour—for instance, formal lyric poetry and classic drama. In these cases the word "form" is used to mean a predetermined structure within whose mold the content must be fitted. These structures are, however, quite simple and so cannot be said to determine the content. Racine and Corneille were contemporaries; both were Neoclassic French dramatists; both abided by all the artificial rules—usually observing the "unities" and following the same strict rules of prosody. Yet their plays, and the poetry in which they are written, differ completely. Corneille is intellectually and emotionally a Neoclassicist—clear and hard, a true objectivist, sure of both his verse and the motivations of his characters. Racine was a great romantic long before the age of Romanticism. His characters are confused and tortured; his verse throbs like the heartbeats of his desperate heroines. He is a great sentimentalist in the best and deepest meaning of that word. His later influence on poets like Baudelaire and Paul Valéry is due to his mastery of sentimental expression, not, as they supposed, to his mastery of Neoclassic form.

Predetermined structure of classic drama

Verse on any subject matter can of course be written purely according to formula. The 18th century in England saw all sorts of prose treatises cast in rhyme and metre, but this was simply applied patterning. (Works such as *The Botanic Garden* [2 vol., 1794–95] by Erasmus Darwin should be sharply distinguished from James Thomson's *The Seasons* [1726–30], which is true poetry, not versified natural history—just as Virgil's *Georgics* is not an agricultural handbook.) Neoclassicism, especially in its 18th-century developments, confused—for ordinary minds, at

any rate—formula with form and so led to the revolt called Romanticism. The leading theorists of that revolt, the poets William Wordsworth and Samuel Taylor Coleridge, in the "Preface" (1800) to *Lyrical Ballads* urged the observance of a few simple rules basic to all great poetry and demanded a return to the integrity of expressive form. A similar revolution in taste was taking place all over Europe and also in China (where the narrow pursuit of formula had almost destroyed poetry). The Romantic taste could enjoy the "formlessness" of William Blake's prophetic books, or Walt Whitman's *Leaves of Grass,* or the loose imagination of Shelley—but careful study reveals that these writers were not formless at all. Each had his own personal form.

Time passes and the pendulum of taste swings. In the mid-20th century, Paul Valéry, T.S. Eliot, and Yvor Winters would attack what the latter called "the fallacy of expressive form," but this is itself a fallacy. All form in literature is expressive. All expression has its own form, even when the form is a deliberate quest of formlessness. (The automatic writing cultivated by the surrealists, for instance, suffers from the excessive formalism of the unconscious mind and is far more stereotyped than the poetry of the Neoclassicist Alexander Pope.) Form simply refers to organization, and critics who attack form do not seem always to remember that a writer organizes more than words. He organizes experience. Thus, his organization stretches far back in his mental process. Form is the other face of content, the outward, visible sign of inner spiritual reality.

Form and formlessness in modern literature

LITERATURE AND ITS AUDIENCE

Folk and elite literatures. In preliterate societies oral literature was widely shared; it saturated the society and was as much a part of living as food, clothing, shelter, or religion. In barbaric societies, the minstrel might be a courtier of the king or chieftain, and the poet who composed liturgies might be a priest. But the oral performance itself was accessible to the whole community. As society evolved its various social layers, or classes, an "elite" literature began to be distinguishable from the "folk" literature of the people. With the invention of writing this separation was accelerated until finally literature was being experienced individually by the elite (reading a book), while folklore and folk song were experienced orally and more or less collectively by the illiterate common people.

Elite literature continuously refreshes itself with materials drawn from the popular. Almost all poetic revivals, for instance, include in their programs a new appreciation of folk song, together with a demand for greater objectivity. On the other hand folk literature borrows themes and, very rarely, patterns from elite literature. Many of the English and Scottish ballads that date from the end of the Middle Ages and have been preserved by oral tradition share plots and even turns of phrase with written literature. A very large percentage of these ballads contain elements that are common to folk ballads from all over western Europe; central themes of folklore, indeed, are found all over the world. Whether these common elements are the result of diffusion is a matter for dispute. They do, however, represent great psychological constants, archetypes of experience common to the human species, and so these constants are used again and again by elite literature as it discovers them in folklore.

Modern popular literature. There is a marked difference between true popular literature, that of folklore and folk song, and the popular literature of modern times. Popular literature today is produced either to be read by a literate audience or to be enacted on television or in the cinema; it is produced by writers who are members, however lowly, of an elite corps of professional literates. Thus, popular literature no longer springs from the people; it is handed to them. Their role is passive. At the best they are permitted a limited selectivity as consumers.

Certain theorists once believed that folk songs and even long, narrative ballads were produced collectively, as has been said in mockery "by the tribe sitting around the fire and grunting in unison." This idea is very much out of date. Folk songs and folk tales began somewhere in one human mind. They were developed and shaped into the forms in which they are now found by hundreds of other minds as they were passed down through the centuries. Only in this sense were they "collectively" produced. During the 20th century, folklore and folk speech have had a great influence on elite literature—on writers as different as Franz Kafka and Carl Sandburg, Selma Lagerlöf and Kawabata Yasunari, Martin Buber and Isaac Bashevis Singer. Folk song has always been popular with bohemian intellectuals, especially political radicals (who certainly are an elite). Since World War II the influence of folk song upon popular song has not just been great; it has been determinative. Almost all "hit" songs since the mid-century have been imitation folk songs; and some authentic folk singers attract immense audiences.

Influence of folklore on modern literature

Popular fiction and drama, westerns and detective stories, films and television serials, all deal with the same great archetypal themes as folktales and ballads, though this is seldom due to direct influence; these are simply the limits within which the human mind works. The number of people who have elevated the formulas of popular fiction to a higher literary level is surprisingly small. Examples are H.G. Wells's early science fiction, the western stories of Gordon Young and Ernest Haycox, the detective stories of Sir Arthur Conan Doyle, Georges Simenon, and Raymond Chandler.

The latter half of the 20th century has seen an even greater change in popular literature. Writing is a static medium: that is to say, a book is read by one person at a time; it permits recollection and anticipation; the reader can go back to check a point or move ahead to find out how the story ends. In radio, television, and the cinema the medium is fluent; the audience is a collectivity and is at the mercy of time. It cannot pause to reflect or to understand more fully without missing another part of the action, nor can it go back or forward. Marshall McLuhan in his book *Understanding Media* (1964) became famous for erecting a whole structure of aesthetic, sociological, and philosophical theory upon this fact. But it remains to be seen whether the new, fluent materials of communication are going to make so very many changes in civilization, let alone in the human mind—mankind has, after all, been influenced for thousands of years by the popular, fluent arts of music and drama. Even the most transitory television serial was written down before it was performed, and the script can be consulted in the files. Before the invention of writing, all literature was fluent because it was contained in people's memory. In a sense it was more fluent than music, because it was harder to remember. Man in mass society becomes increasingly a creature of the moment, but the reasons for this are undoubtedly more fundamental than his forms of entertainment.

LITERATURE AND ITS ENVIRONMENT

Social and economic conditions. Literature, like all other human activities, necessarily reflects current social and economic conditions. Class stratification was reflected in literature as soon as it had appeared in life. Among the American Indians, for instance, the chants of the shaman, or medicine man, differ from the secret, personal songs of the individual, and these likewise differ from the group songs of ritual or entertainment sung in community. In the Heroic Age, the epic tales of kings and chiefs that were sung or told in their barbaric courts differed from the folktales that were told in peasant cottages.

Reflection of class distinction in literature

The more cohesive a society, the more the elements—and even attitudes—evolved in the different class strata are interchangeable at all levels. In the tight clan organization that existed in late medieval times at the Scottish border, for example, heroic ballads telling of the deeds of lords and ladies were preserved in the songs of the common people. But where class divisions are unbridgeable, elite literature is liable to be totally separated from popular culture. An extreme example is the classic literature of the Roman Empire. Its forms and its sources were largely Greek—it even adopted its laws of verse patterning from Greek models, even though these were antagonistic to the natural patterns of the Latin language—and most of the sophisticated works of the major Latin authors were com-

pletely closed to the overwhelming majority of people of the Roman Empire.

Printing has made all the difference in the negotiability of ideas. The writings of the 18th-century French writers Voltaire, Rousseau, and Diderot were produced from and for almost as narrow a caste as the Roman elite, but they were printed. Within a generation they had penetrated the entire society and were of vital importance in revolutionizing it.

Class distinctions in the literature of modern times exist more in the works themselves than in their audience. Although Henry James wrote about the upper classes and Émile Zola about workingmen, both were, in fact, members of an elite and were read by members of an elite—moreover, in their day, those who read Zola certainly considered themselves more of an elite than did the readers of Henry James. The ordinary people, if they read at all, preferred sentimental romances and "penny dreadfuls." Popular literature had already become commercially produced entertainment literature, a type which today is also provided by television scripts.

The elite who read serious literature are not necessarily members of a social or economic upper class. It has been said of the most ethereal French poet, Stéphane Mallarmé, that in every French small town there was a youth who carried his poems in his heart. These poems are perhaps the most "elite" product of western European civilization, but the "youths" referred to were hardly the sons of dukes or millionaires. (It is a curious phenomenon that, since the middle of the 18th century in Europe and in the United States, the majority of readers of serious literature—as well as of entertainment literature—have been women. The extent of the influence that this audience has exerted on literature itself must be immense.)

Hippolyte Taine's ecological theory of literature

National and group literature. Hippolyte Taine, the 19th-century French critic, evolved an ecological theory of literature. He looked first and foremost to the national characteristics of western European literatures, and he found the source of these characteristics in the climate and soil of each respective nation. His *History of English Literature* (5 vol., 1863–69) is an extensive elaboration of these ideas. It is doubtful that anyone today would agree with the simplistic terms in which Taine states his thesis. It is obvious that Russian literature differs from English or French from German. English books are written by Englishmen, their scenes are commonly laid in England, they are usually about Englishmen and they are designed to be read by Englishmen—at least in the first instance. But modern civilization becomes more and more a world civilization, wherein works of all peoples flow into a general fund of literature. It is not unusual to read a novel by a Japanese author one week and one by a black writer from West Africa the next. Writers are themselves affected by this cross-fertilization. Certainly, the work of the great 19th-century Russian novelists has had more influence on 20th-century American writers than has the work of their own literary ancestors. Poetry does not circulate so readily, because catching its true significance in translation is so very difficult to accomplish. Nevertheless, for the past 100 years or so, the influence of French poetry upon all the literatures of the civilized world has not just been important, it has been preeminent. The tendentious elements of literature—propaganda for race, nation, or religion—have been more and more eroded in this process of wholesale cultural exchange.

Popular literature on the other hand is habitually tendentious both deliberately and unconsciously. It reflects and stimulates the prejudices and parochialism of its audience. Most of the literary conflicts that have seized the totalitarian countries during the 20th century stem directly from relentless efforts by the state to reduce elite literature to the level of the popular. The great proletarian novels of our time have been produced, not by Russians, but by American blacks, Japanese, Germans, and—most proletarian of all—a German-American living in Mexico, B. Traven. Government control and censorship can inhibit literary development, perhaps deform it a little, and can destroy authors outright; but, whether in the France of Louis XIV or in the Soviet Union of the 20th century, it cannot be said to have a fundamental effect upon the course of literature.

Effect of censorship on literary development

The writer's position in society. A distinguishing characteristic of modern literature is the peculiar elite which it has itself evolved. In earlier cultures the artist, though he may have been neurotic at times, thought of himself as part of his society and shared its values and attitudes. Usually the clerkly caste played a personal, important role in society. In the modern industrial civilization, however, "scribes" became simply a category of skilled hired hands. The writer shared few of the values of the merchant or the entrepreneur or manager. And so the literary and artistic world came to have a subculture of its own. The antagonism between the two resultant sets of values is the source of what we call alienation—among the intellectuals at least (the alienation of the common man in urban, industrial civilization from his work, from himself, and from his fellows is another matter, although its results are reflected and intensified in the alienation of the elite). For about 200 years now, the artistic environment of the writer has not usually been shared with the general populace. The subculture known as bohemia and the literary and artistic movements generated in its little special society have often been more important—at least in the minds of many writers—than the historical, social, and economic movements of the culture as a whole. Even massive historical change is translated into these terms—the Russian Revolution, for instance, into Communist-Futurism, Constructivism, Socialist Realism. Western European literature could be viewed as a parade of movements—Romanticism, Realism, Naturalism, Futurism, Structuralism, and so on indefinitely. Some of the more journalistic critics, indeed, have delighted to regard it in such a way. But after the manifestos have been swept away, the meetings adjourned, the literary cafés of the moment lost their popularity, the turmoil is seen not to have made so very much difference. The Romantic Théophile Gautier (1811–72) and the Naturalist Émile Zola (1840–1902) have more in common than they have differences, and their differences are rather because of changes in society as a whole than because of conflicting literary principles.

Literary subcultures and movements

At first, changes in literary values are appreciated only at the upper levels of the literary elite itself, but often, within a generation, works once thought esoteric are being taught as part of a school syllabus. Most cultivated people once thought James Joyce's *Ulysses* incomprehensible or, where it was not, obscene. Today his methods and subject matter are commonplace in the commercial fiction of the mass culture. A few writers remain confined to the elite. Mallarmé is a good example—but he would have been just as ethereal had he written in the simplest French of direct communication. His subtleties are ultimately grounded in his personality.

Literature and the other arts. Literature has an obvious kinship with the other arts. Presented, a play is drama; read, a play is literature. Most important films have been based upon written literature, usually novels, although all the great epics and most of the great plays have been filmed at some time and thus have stimulated the younger medium's growth. Conversely, the techniques required in writing for film have influenced many writers in structuring their novels and have affected their style. Most popular fiction is written with "movie rights" in mind, and these are certainly a consideration with most modern publishers. Literature provides the libretto for operas, the theme for tone poems—even so anomalous a form as Nietzsche's *Thus Spake Zarathustra* was interpreted in music by Richard Strauss—and of course it provides the lyrics of songs. Many ballets and modern dances are based on stories or poems. Sometimes, music and dance are accompanied by a text read by a speaker or chanted by a chorus. The mid-19th century was the heyday of literary, historical, and anecdotal painting, though, aside from the Surrealists, this sort of thing died out in the 20th century. Cross-fertilization of literature and the arts now takes place more subtly, mostly in the use of parallel techniques—the rational dissociation of the Cubists or the spontaneous action painting of the Abstract Expressionists, for example, which flourished at the same time as

the free-flowing uncorrected narratives of some novelists in the 1950s and '60s.

LITERATURE AS A COLLECTION OF GENRES

Critics have invented a variety of systems for treating literature as a collection of genres. Often these genres are artificial, invented after the fact with the aim of making literature less sprawling, more tidy. Theories of literature must be based upon direct experience of the living texts and so be flexible enough to contain their individuality and variety. Perhaps the best approach is historical, or genetic. What actually happened, and in what way did literature evolve up to the present day?

Oral literature

There is a surprising variety of oral literature among surviving preliterate peoples, and, as the written word emerges in history, the indications are that the important literary genres all existed at the beginning of civilized societies: heroic epic; songs in praise of priests and kings; stories of mystery and the supernatural; love lyrics; personal songs (the result of intense meditation); love stories; tales of adventure and heroism (of common peoples, as distinct from the heroic epics of the upper classes); satire (which was dreaded by barbaric chieftains); satirical combats (in which two poets or two personifications abused one another and praised themselves); ballads and folktales of tragedy and murder; folk stories, such as the tale of the clever boy who performs impossible tasks, outwits all his adversaries, and usually wins the hand of the king's daughter; animal fables like those attributed to Aesop (the special delight of Black Africa and Indian America); riddles, proverbs, and philosophical observations; hymns, incantations, and mysterious songs of priests; and finally actual mythology—stories of the origin of the world and the human race, of the great dead, and of the gods and demigods.

Epic. The true heroic epic never evolved far from its preliterate origins, and it arose only in the Heroic Age which preceded a settled civilization. The conditions reflected in, say, the *Iliad* and *Odyssey* are much the same as those of the Anglo-Saxon *Beowulf,* the German *Nibelungenlied,* or the Irish stories of Cú Chulainn. The literary epic is another matter altogether. Virgil's *Aeneid,* for instance, or Milton's *Paradise Lost* are products of highly sophisticated literary cultures. Many long poems sometimes classified as epic literature are no such thing—Dante's *La divina commedia* (*The Divine Comedy*), for example, is a long theological, philosophical, political, moral, and mystical poem. Dante considered it to be a kind of drama which obeyed the rules of Aristotle's *Poetics.* Goethe's *Faust* is in dramatic form and is sometimes even staged—but it is really a philosophical poetic novel. Modern critics have described long poems such as T.S. Eliot's *Waste Land* and Ezra Pound's *Cantos* as "philosophical epics." There is nothing epic about them; they are reveries, more or less philosophical.

Lyric poetry. Lyric poetry never gets far from its origins, except that some of its finest examples—Medieval Latin, Provençal, Middle High German, Middle French, Renaissance—which today are only read, were actually written to be sung. In the 20th century, however, popular songs of great literary merit have become increasingly common—for example, the songs of Bertolt Brecht and Kurt Weill in German, of Georges Brassens and Anne Sylvestre in French, and of Leonard Cohen, Bob Dylan, and Joni Mitchell. It is interesting to note that, in periods when the culture values artificiality, the lyric becomes stereotyped. Then, after a while, the poets revolt and, usually turning to folk origins, restore to lyric poetry at least the appearance of naturalness and spontaneity.

Diverse forms of satire

Satire. The forms of satire are as manifold as those of literature itself—from those of the mock epic to the biting epigram. A great many social and political novels of today would have been regarded as satire by the ancients. Many of the great works of all time are satires, but in each case they have risen far above their immediate satirical objectives. The 16th-century medieval satire on civilization, the *Gargantua and Pantagruel* of Rabelais, grew under the hand of its author into a great archetypal myth of the lust for life. Cervantes' *Don Quixote,* often called the greatest work of prose fiction in the West, is superficially a satire of the sentimental romance of knightly adventure. But, again, it is an archetypal myth, telling the adventures of the soul of man—of the individual—in the long struggle with what is called the human condition. *The Tale of Genji* by Murasaki Shikibu has sometimes been considered by obtuse critics as no more than a satire on the sexual promiscuity of the Heian court. In fact, it is a profoundly philosophical, religious, and mystical novel.

Prose fiction. Extended prose fiction is the latest of the literary forms to develop. We have romances from classical Greek times that are as long as short novels; but they are really tales of adventure—vastly extended anecdotes.

The earliest prose fiction

The first prose fiction of any psychological depth is the *Satyricon,* almost certainly attributed to Petronius Arbiter (died AD 65/66). Though it survives only in fragments, supposedly one-eleventh of the whole, even these would indicate that it is one of the greatest picaresque novels, composed of loosely connected episodes of robust and often erotic adventure. The other great surviving fiction of classical times is the *Metamorphoses* (known as *The Golden Ass*) by Apuleius (2nd century AD). In addition to being a picaresque adventure story, it is a criticism of Roman society, a celebration of the religion of Isis, and an allegory of the progress of the soul. It contains the justly celebrated story of Cupid and Psyche, a myth retold with psychological subtlety. Style has much to do with the value and hence the survival of these two works. They are written in prose of extraordinary beauty, although it is by no means of "classical" purity. The prose romances of the Middle Ages are closely related to earlier heroic literature. Some, like Sir Thomas Malory's 15th-century *Le Morte Darthur,* are retellings of heroic legend in terms of the romantic chivalry of the early Renaissance, a combination of barbaric, medieval, and Renaissance sensibility which, in the tales of Tristram and Iseult and Launcelot and Guinevere, produced something not unlike modern novels of tragic love.

The Western novel is a product of modern civilization, although in the Far East novels began a separate development as early as the 10th century. Extended prose works of complex interpersonal relations and motivations begin in 17th-century France with *The Princess of Cleves* (1678) by Madame de La Fayette. Eighteenth-century France produced an immense number of novels dealing with love analysis but none to compare with Madame de La Fayette's until Pierre Choderlos de Laclos wrote *Les Liaisons dangereuses* (1782). This was, in form, an exchange of letters between two corrupters of youth; but, in intent, it was a savage satire of the *ancien régime* and a heart-rending psychological study. The English novel of the 18th century was less subtle, more robust—vulgar in the best sense—and is exemplified by Henry Fielding's *Tom Jones* (1749) and Laurence Sterne's *Tristram Shandy.* The 19th century was the golden age of the novel. It became ever more profound, complex, and subtle (or, on the other hand, more popular, eventful, and sentimental). By the beginning of the 20th century it had become the most common form of thoughtful reading matter and had replaced, for most educated people, religious, philosophical, and scientific works as a medium for the interpretation of life. By the late 1920s the novel had begun to show signs of decay as a form, and no works have since been produced to compare with the recent past. This may prove to be a temporarily barren period, or else the novel may be losing its energy as a narrative art form and in this sense giving way to the medium of film.

Current state of the novel

Drama. Like lyric poetry, drama has been an exceptionally stable literary form. Given a little leeway, most plays written by the beginning of the 20th century could be adjusted to the rules of Aristotle's *Poetics.* Before World War I, however, all traditional art forms, led by painting, began to disintegrate, and new forms evolved to take their place. In drama the most radical innovator was August Strindberg (1849–1912), and from that day to this, drama (forced to compete with the cinema) has become ever more experimental, constantly striving for new methods, materials, and, especially, ways to establish a close relationship with the audience. All this activity has profoundly modified drama as literature.

Future developments. In the 20th century the methods of poetry have also changed drastically, although the "innovator" here might be said to have been Baudelaire (1821–67). The disassociation and recombination of ideas of the Cubists, the free association of ideas of the Surrealists, dreams, trance states, the poetry of preliterate people—all have been absorbed into the practice of modern poetry. This proliferation of form is not likely to end. Effort that once was applied to perfecting a single pattern in a single form may in the future be more and more directed toward the elaboration of entirely new "multimedia" forms, employing the resources of all the established arts. At the same time, writers may prefer to simplify and polish the forms of the past with a rigorous, Neoclassicist discipline. In a worldwide urban civilization, which has taken to itself the styles and discoveries of all cultures past and present, the future of literature is quite impossible to determine.

WRITINGS ON LITERATURE

Scholarly research. Research by scholars into the literary past began almost as soon as literature itself—as soon as the documents accumulated—and for many centuries it represents almost all the scholarship that has survived. The most extensive text of the Sumerian *Epic of Gilgamesh,* the first of the world's great classics, is a late Assyrian synthesis that must have required an immense amount of research into clay tablets, written in several languages going back to the beginning of Mesopotamian civilization. Many Egyptian poems and the philosophic creation myth known as the "Memphite Theology" survive in very late texts that carefully reproduce the original language of the first dynasties. Once the function of the scribe was established as essential, he invented literary scholarship, both to secure his position and to occupy his leisure. The great epoch of literary scholarship in ancient times centred on the library (and university) of Alexandria from its foundation in 324 BC to its destruction by the Arabs in AD 640. Hellenistic Greek scholars there developed such an academic and pedantic approach to literary scholarship and scholarly literature that the term Alexandrine remains pejorative to this day. To them, however, is owed the survival of the texts of most of the Greek classics. Roman literary scholarship was rhetorical rather than analytic. With the coming of Islām, there was established across the whole warm temperate zone of the Old World a far-flung community of scholars who were at home in learned circles from India to Spain. Judaism, like Islām, was a religion of the book and of written tradition, so literary scholarship played a central role in each. The same is true of India, China, and later Japan; for sheer bulk, as well as for subtlety and insight, Oriental scholarship has never been surpassed. In a sense, the Renaissance in Europe was a cultural revolution led by literary scholars who discovered, revived, and made relevant again the literary heritage of Greece and Rome. In the 19th century, literary scholarship was dominated by the exhaustive, painstaking German academician, and that Germanic tradition passed to the universities of the United States. The demand that every teacher should write a master's thesis, a doctor's dissertation, and, for the rest of his career, publish with reasonable frequency learned articles and scholarly books, has led to a mass of scholarship of widely varying standards and value. Some is trivial and absurd, but the best has perfected the texts and thoroughly illuminated the significance of nearly all the world's great literature.

Ancient literary scholarship

Literary criticism. Literary criticism, as distinguished from scholarly research, is usually itself considered a form of literature. Some people find great critics as entertaining and stimulating as great poets, and theoretical treatises of literary aesthetics can be as exciting as novels. Aristotle, Longinus, and the Roman rhetorician and critic Quintilian are still read, although Renaissance critics like the once all-powerful Josephus Scaliger are forgotten by all but specialized scholars. Later critics, such as Poe, Sainte-Beuve, Taine, Vissarion Belinsky, Matthew Arnold, Walter Bagehot, Walter Pater, and George Saintsbury, are probably read more for themselves than for their literary judgments and for their general theorizing rather than for their applications (in the case of the first three, for instance, time has confounded almost all the evaluations they made of their contemporaries). The English critics have survived because they largely confined themselves to acknowledged masterpieces and general ideas. Perhaps literary criticism can really be read as a form of autobiography. Aestheticians of literature like I.A. Richards, Sir C.M. Bowra, Paul Valéry, Suzanne Langer, and Ernst Cassirer have had an influence beyond the narrow confines of literary scholarship and have played in our time something approaching the role of general philosophers. This has been true on the popular level as well. The Dane Georg Brandes, the Americans James Gibbons Huneker, H.L. Mencken, and Edmund Wilson—these men have been social forces in their day. Literary criticism can play its role in social change. In Japan, the overthrow of the shogunate, the restoration of the emperor, and the profound change in the Japanese social sensibility begins with the literary criticism of Motoori Norinaga (1730–1801). The 19th-century revolution in theology resulted from the convergence of Darwinian theories of evolution and the technical and historical criticism of the Bible that scholars had undertaken. For many modern intellectuals, the literary quarterlies and weeklies, with their tireless discussions of the spiritual significance and formal characteristics of everything from the greatest masterpiece to the most ephemeral current production, can be said to have filled the place of religion, both as rite and dogma. (K.Re.)

POETRY

The nature of poetry

Poetry is a vast subject, as old as history and older, present wherever religion is present, possibly—under some definitions—the primal and primary form of languages themselves. The present section means only to describe in as general a way as possible certain properties of poetry and of poetic thought regarded as in some sense independent modes of the mind. Naturally, not every tradition nor every local or individual variation can be—or need be—included, but the article illustrates by examples of poetry ranging between nursery rhyme and epic. (For particular information about various types of poetry, see below *Prosody, Ballad,* and *Epic.*)

ATTEMPTS TO DEFINE POETRY

Poetry is the other way of using language. Perhaps in some hypothetical beginning of things it was the only way of using language or simply was language *tout court,* prose being the derivative and younger rival. Both poetry and language are fashionably thought to have belonged to ritual in early agricultural societies; and poetry in particular, it has been claimed, arose at first in the form of magical spells recited to ensure a good harvest. Whatever the truth of this hypothesis, it blurs a useful distinction: by the time there begins to be a separate class of objects called poems, recognizable as such, these objects are no longer much regarded for their possible yam-growing properties, and such magic as they may be thought capable of has retired to do its business upon the human spirit and not directly upon the natural world outside.

Formally, poetry is recognizable by its greater dependence on at least one more parameter, the *line,* than appears in prose composition. This changes its appearance on the page; and it seems clear that people take their cue from this changed appearance, reading poetry aloud in a very different voice from their habitual voice, possibly because, as Ben Jonson said, poetry "speaketh somewhat above a mortal mouth." If, as a test of this description, people are shown poems printed as prose, it most often turns out that they will read the result as prose simply because it looks that way; which is to say that they are no longer

guided in their reading by the balance and shift of the line in relation to the breath as well as the syntax.

That is a minimal definition but perhaps not altogether uninformative. It may be all that ought to be attempted in the way of a definition: Poetry is the way it is because it looks that way, and it looks that way because it sounds that way and vice versa.

POETRY AND PROSE

Distinguishing prose from poetry

People's reason for wanting a definition is to take care of the borderline case, and this is what a definition, as if by definition, will not do. That is, if a man asks for a definition of poetry, it will most certainly not be the case that he has never seen one of the objects called poems that are said to embody poetry; on the contrary, he is already tolerably certain what poetry in the main is, and his reason for wanting a definition is either that his certainty has been challenged by someone else or that he wants to take care of a possible or seeming exception to it: hence the perennial squabble about distinguishing poetry from prose, which is rather like distinguishing rain from snow—everyone is reasonably capable of doing so, and yet there are some weathers that are either-neither.

Sensible things have been said on the question. The poet T.S. Eliot suggested that part of the difficulty lies in the fact that there is the technical term "verse" to go with the term "poetry," while there is no equivalent technical term to distinguish the mechanical part of prose and make the relation symmetrical. The French poet Paul Valéry said that prose was walking, poetry dancing. Indeed, the original two terms, *prosus* and *versus,* meant, respectively, "going straight forth" and "returning"; and that distinction does point up the tendency of poetry to incremental repetition, variation, and the treatment of many matters and different themes in a single recurrent form such as couplet or stanza.

Robert Frost said shrewdly that poetry was what got left behind in translation, which suggests a criterion of almost scientific refinement: when in doubt, translate; whatever comes through is prose, the remainder is poetry. And yet to even so acute a definition the obvious exception is a startling and a formidable one: some of the greatest poetry in the world is in the Authorized Version of the Bible, which is not only a translation but also, as to its appearance in print, identifiable neither with verse nor with prose in English but rather with a cadence owing something to both.

There may be a better way of putting the question by the simple test alluded to above. When people are presented with a series of passages drawn indifferently from poems and stories but all printed as prose, they will show a dominant inclination to identify everything they possibly can as prose. This will be true, surprisingly enough, even if the poem rhymes and will often be true even if the poem in its original typographical arrangement would have been familiar to them. The reason seems to be absurdly plain: the reader recognizes poems by their appearance on the page, and he responds to the convention whereby he recognizes them by reading them aloud in a quite different tone of voice from that which he applies to prose (which, indeed, he scarcely reads aloud at all). It should be added that he makes this distinction also without reading aloud; even in silence he confers upon a piece of poetry an attention that differs from what he gives to prose in two ways especially: in tone and in pace.

In place of further worrying over definitions, it may be both a relief and an illumination to exhibit certain plain and mighty differences between prose and poetry by a comparison. In the following passages a prose writer and a poet are talking about the same subject, growing older.

> Between the ages of 30 and 90, the weight of our muscles falls by 30 percent and the power we can exert likewise . . . . The number of nerve fibres in a nerve trunk falls by a quarter. The weight of our brains falls from an average of 3.03 lb. to 2.27 lb. as cells die and are not replaced . . . . (Gordon Rattray Taylor, *The Biological Time Bomb,* 1968.)

> Let me disclose the gifts reserved for age
> To set a crown upon your lifetime's effort.
> First, the cold friction of expiring sense
> Without enchantment, offering no promise
> But bitter tastelessness of shadow fruit
> As body and soul begin to fall asunder.
> Second, the conscious impotence of rage
> At human folly, and the laceration
> Of laughter at what ceases to amuse.
> And last, the rending pain of re-enactment
> Of all that you have done, and been . . . .
> (T.S. Eliot, *Four Quartets.*)

Before objecting that a simple comparison cannot possibly cover all the possible ranges of poetry and prose compared, the reader should consider for a moment what differences are exhibited. The passages are oddly parallel, hence comparable, even in a formal sense; for both consist of the several items of a catalog under the general title of growing old. The significant differences are of tone, pace, and object of attention. If the prose passage interests itself in the neutral, material, measurable properties of the process, while the poetry interests itself in what the process will signify to someone going through it, that is not accidental but of the essence; if one reads the prose passage with an interest in being informed, noting the parallel constructions without being affected by them either in tone or in pace, while reading the poetry with a sense of considerable gravity and solemnity, that too is of the essence. One might say as tersely as possible that the difference between prose and poetry is most strikingly shown in the two uses of the verb "to fall":

Differences of tone, pace, and object of attention

> The number of nerve fibres in a nerve trunk falls by a quarter
> As body and soul begin to fall asunder

It should be specified here that the important differences exhibited by the comparison belong to the present age. In each period, speaking for poetry in English at any rate, the dividing line will be seen to come at a different place. In Elizabethan times the diction of prose was much closer to that of poetry than it later became, and in the 18th century authors saw nothing strange about writing in couplets about subjects that later would automatically and compulsorily belong to prose—for example, horticulture, botany, even dentistry. Here is not the place for entering into a discussion of so rich a chapter in the history of ideas; but it should be remarked that the changes involved in the relation between poetry and prose are powerfully influenced by the immense growth of science, commerce, and number in man's ways of describing, even of viewing, the world.

Precision of poetic diction

Returning to the comparison, it is observable that though the diction of the poem is well within what could be commanded by a moderately well-educated speaker, it is at the same time well outside the range of terms in fact employed by such a speaker in his daily occasions; it is a diction very conscious, as it were, of its power of choosing terms with an effect of peculiar precision and of combining the terms into phrases with the same effect of peculiar precision and also of combining sounds with the same effect of peculiar precision. Doubtless the precision of the prose passage is greater in the more obvious property of dealing in the measurable; but the poet attempts a precision with respect to what is not in the same sense measurable nor even in the same sense accessible to observation; the distinction is perhaps just that made by the French scientist and philosopher Blaise Pascal in discriminating the spirits of geometry and finesse; and if one speaks of "effects of precision" rather than of precision itself, that serves to distinguish one's sense that the art work is always somewhat removed from what people are pleased to call the real world, operating instead, in Immanuel Kant's shrewd formula, by exhibiting "purposefulness without purpose." To much the same point is what Samuel Taylor Coleridge remembers having learned from his schoolmaster:

> I learnt from him, that Poetry, even that of the loftiest and, seemingly, that of the wildest odes, had a logic of its own, as severe as that of science; and more difficult, because more subtle, more complex, and dependent on more, and more fugitive causes. In the truly great poets, he would say, there is a reason assignable, not only for every word, but for the position of every word. (*Biographia Literaria,* ch. I.)

Perhaps this is a somewhat exaggerated, as it is almost

always an unprovable, claim, illustrating also a propensity for competing with the prestige of science on something like its own terms—but the last remark in particular illuminates the same author's terser formulation: "prose = words in the best order, poetry = the best words in the best order." This attempt at definition, impeccable because uninformative, was derived from Jonathan Swift, who had said, also impeccably and uninformatively, that style in writing was "the best words in the best order." Which may be much to the same effect as Louis Armstrong's saying, on being asked to define jazz, "Baby, if you got to ask the question, you're never going to know the answer." Or the painter Marcel Duchamp's elegant remark on what psychologists call "the problem of perception": "If no solution, then maybe no problem?" This species of gnomic, riddling remark may be determinate for the artistic attitude toward definition of every sort; and its skepticism is not confined to definitions of poetry but extends to definitions of anything whatever, directing one not to dictionaries but to experience and, above all, to use: "Anyone with a watch can tell you what time it is," said Valéry, "but who can tell you what is time?"

Experiencing poetry

Happily, if poetry is almost impossible to define, it is extremely easy to recognize in experience; even untutored children are rarely in doubt about it when it appears:

> Little Jack Jingle,
> He used to live single,
> But when he got tired of this kind of life,
> He left off being single, and liv'd with his wife.

It might be objected that this little verse is not of sufficient import and weight to serve as an exemplar for poetry. It ought to be remembered, though, that it has given people pleasure so that they continued to say it until and after it was written down, nearly two centuries ago. The verse has survived, and its survival has something to do with pleasure, with delight; and while it still lives, how many more imposing works of language—epic poems, books of science, philosophy, theology—have gone down, deservedly or not, into dust and silence. It has, obviously, a form, an arrangement of sounds in relation to thoughts that somehow makes its agreeable nonsense closed, complete, and decisive. But this somewhat muddled matter of form deserves a heading and an instance all to itself.

FORM IN POETRY

People nowadays who speak of form in poetry almost always mean such externals as regular measure and rhyme, and most often they mean to get rid of these in favour of the freedom they suppose must follow upon the absence of form in this limited sense. But in fact a poem having only one form would be of doubtful interest even if it could exist. In this connection, the poet J.V. Cunningham speaks of "a convergence of forms, and forms of disparate orders," adding: "It is the coincidence of forms that locks in the poem." For a poem is composed of internal and intellectual forms as well as forms externally imposed and preexisting any particular instance, and these may be sufficient without regular measure and rhyme; if the intellectual forms are absent, as in greeting-card verse and advertising jingles, no amount of thumping and banging will supply the want.

Form, in effect, is like the doughnut that may be said to be nothing in a circle of something or something around nothing; it is either the outside of an inside, as when people speak of "good form" or "bourgeois formalism," or the inside of an outside, as in the scholastic saying that "the soul is the form of the body." Taking this principle, together with what Cunningham says of the matter, one may now look at a very short and very powerful poem with a view to distinguishing the forms, or schemes, of which it is made. It was written by Rudyard Kipling—a great poet at present somewhat sunken in reputation, probably on account of misinterpretations having to do more with his imputed politics than with his poetry—and its subject, one of a series of epitaphs for the dead of World War I, is a soldier shot by his comrades for cowardice in battle.

> I could not look on Death, which being known,
> Men led me to him, blindfold and alone.

The aim of the following observations and reflections is to distinguish as clearly as possible—distinguish without dividing—the feelings evoked by the subject, so grim, horrifying, tending to helpless sorrow and despair, from the feelings, which might better be thought of as meanings, evoked by careful contemplation of the poem in its manifold and somewhat subtle ways of handling the subject, leading the reader on to a view of the strange delight intrinsic to art, whose mirroring and shielding power allows him to contemplate the world's horrible realities without being turned to stone.

Kinds of poetic form

There is, first, the obvious external form of a rhymed, closed couplet in iambic pentameter (that is, five poetic "feet," each consisting of an unstressed followed by a stressed syllable, per line). There is, second, the obvious external form of a single sentence balanced in four grammatical units with and in counterpoint with the metrical form. There is, third, the conventional form belonging to the epitaph and reflecting back to antiquity; it is terse enough to be cut in stone and tight-lipped also, perhaps for other reasons, such as the speaker's shame. There is, fourth, the fictional form belonging to the epitaph, according to which the dead man is supposed to be saying the words himself. There is, fifth, especially poignant in this instance, the real form behind or within the fictional one, for the reader is aware that in reality it is not the dead man speaking, nor are his feelings the only ones the reader is receiving, but that the comrades who were forced to execute him may themselves have made up these two lines with their incalculably complex and exquisite balance of scorn, awe, guilt, and consideration even to tenderness for the dead soldier. There is, sixth, the metaphorical form, with its many resonances ranging from the tragic through the pathetic to irony and apology: dying in battle is spoken of in language relating it to a social occasion in drawing room or court; the coward's fear is implicitly represented as merely the timorousness and embarassment one might feel about being introduced to a somewhat superior and majestic person, so that the soldiers responsible for killing him are seen as sympathetically helping him through a difficult moment in the realm of manners. In addition, there is, seventh, a linguistic or syntactical form, with at least a couple of tricks to it: the second clause, with its reminiscence of Latin construction, participates in the meaning by conferring a Roman stoicism and archaic gravity on the saying; remembering that the soldiers in the poem had been British schoolboys not long before, the reader might hear the remote resonance of a whole lost world built upon Greek and Roman models; and the last epithets, "blindfold and alone," while in the literal acceptation they clearly refer to the coward, show a distinct tendency to waver over and apply mysteriously to Death as well, sitting there waiting "blindfold and alone." One might add another form, the eighth, composed of the balance of sounds, from the obvious likeness in the rhyme down to subtleties and refinements beneath the ability of coarse analysis to discriminate. And even there one would not be quite at an end; an overall principle remains, the compression of what might have been epic or five-act tragedy into two lines, or the poet's precise election of a single instant to carry what the novelist, if he did his business properly, would have been hundreds of pages arriving at.

The coincidence of poetic forms

It is not at all to be inferred that the poet composed his poem in the manner of the above laborious analysis of its strands; the whole insistence, rather, is that he did not catalog eight or 10 forms and assemble them into a poem; more likely it "just came to him." But the example may serve to indicate how many modes of the mind go together in this articulation of an implied drama and the tension among many possible sentiments that might arise in response to it.

In this way, by the coincidence of forms that locks in the poem, one may see how to answer a question that often arises about poems: though their thoughts are commonplace, they themselves mysteriously are not. One may answer on the basis of the example and the inferences produced from it that a poem is not so much a thought as it is a mind: talk with it, and it will talk back, telling you

many things that you might have thought for yourself but somehow didn't until it brought them together. Doubtless a poem is a much simplified model for the mind. But it might still be one of the best man has available. On this great theme, however, it will be best to proceed not by definition but by parable and interpretation.

POETRY AS A MODE OF THOUGHT: THE PROTEAN ENCOUNTER

In the fourth book of the *Odyssey* Homer tells the following strange tale. After the war at Troy, Menelaus wanted very much to get home but was held up in Egypt for want of a wind because, as he later told Telemachus, he had not sacrificed enough to the gods. "Ever jealous the Gods are," he said, "that we men mind their dues." But because the gods work both ways, it was on the advice of a goddess, Eidothea, that Menelaus went to consult Proteus, the old one of the sea, as one might consult a travel agency.

Proteus was not easy to consult. He was herding seals, and the seals stank even through the ambrosia Eidothea had provided. And when Menelaus crept up close, disguised as a seal, and grabbed him, Proteus turned into a lion, a dragon, a leopard, a boar, a film of water, and a high-branched tree. But Menelaus managed to hang on until Proteus gave up and was himself again; whereupon Menelaus asked him the one great question: How do I get home? And Proteus told him: You had better go back to Egypt and sacrifice to the gods some more.

This story may be taken as a parable about poetry. A man has an urgent question about his way in the world. He already knows the answer, but it fails to satisfy him. So at great inconvenience, hardship, and even peril, he consults a powerful and refractory spirit who tries to evade his question by turning into anything in the world. Then, when the spirit sees he cannot get free of the man, and only then, he answers the man's question, not simply with a commonplace but with the same commonplace the man had been dissatisfied with before. Satisfied or not, however, the man now obeys the advice given him.

A foolish story? All the same, it is to be observed that Menelaus did get home. And it was a heroic thing to have hung onto Proteus through those terrifying changes and compelled him to be himself and answer up. Nor does it matter in the least to the story that Menelaus personally may have been a disagreeable old fool as well as a cuckold.

A poet also has one great and simple question, simple though it may take many forms indeed. Geoffrey Chaucer put it as well as anyone could, and in three lines at that:

> What is this world? what asketh men to have?
> Now with his love, now in his colde grave,
> Allone, with-outen any companye.
>
> ("The Knight's Tale")

And a poet gets the simple answer he might expect, the one the world grudgingly gives to anyone who asks such a question: The world is this way, not that way, and you ask for more than you will be given, which the poet, being scarcely more fool than his fellowmen, knew already. But on the path from question to answer, hanging onto the slippery disguiser and shape-shifter Proteus, he will see many marvels; he will follow the metamorphoses of things in the metamorphoses of their phrases, and he will be so elated and ecstatic in this realm of wonders that the voice in which he speaks these things, down even to the stupid, obvious, and commonplace answer, will be to his hearers a solace and a happiness in the midst of sorrows:

> When I do count the clock that tells the time,
> And see the brave day sunk in hideous night;
> When I behold the violet past prime,
> And sable curls, all silver'd o'er with white;
> When lofty trees I see barren of leaves,
> Which erst from heat did canopy the herd,
> And summer's green all girded up in sheaves,
> Borne on the bier with white and bristly beard,
> Then of thy beauty do I question make,
> That thou among the wastes of time must go,
> Since sweets and beauties must themselves forsake
> And die as fast as they see others grow;
> And nothing 'gainst Time's scythe can make defence
> Save breed, to brave him when he takes thee hence.
>
> (Shakespeare, Sonnet 12.)

Like Menelaus, the poet asks a simple question, to which, moreover, he already knows the unsatisfying answer. Question and answer, one might say, have to be present, although of themselves they seem to do nothing much; but they assert the limits of a journey to be taken. They are the necessary but not sufficient conditions of what really seems to matter here, the Protean encounter itself, the grasping and hanging on to the powerful and refractory spirit in its slippery transformations of a single force flowing through clock, day, violet, graying hair, trees dropping their leaves, the harvest in which, by a peculiarly ceremonial transmutation, the grain man lives by is seen without contradiction as the corpse he comes to. As for the answer to the question, it is not surprising nor meant to be surprising; it is only just.

The poetic question and answer

On this point—that the answer comes as no surprise—poets show an agreement that quite transcends the differences of periods and schools. Alexander Pope's formula, "What oft was thought, but ne'er so well expresst," sometimes considered as the epitome of a shallow and parochial decorum, is not in essence other than this offered by John Keats:

> I think Poetry should surprise by a fine excess, and not by Singularity—it should strike the Reader as a wording of his own highest thoughts, and appear almost a Remembrance. (Letter to John Taylor, 1818.)

In the present century, Robert Frost is strikingly in agreement:

> A word about recognition: In literature it is our business to give people the thing that will make them say, "Oh yes I know what you mean." It is never to tell them something they dont know, but something they know and hadnt thought of saying. It must be something they recognize. (Letter to John Bartlett, in *Modern Poetics,* ed. James Scully, 1965.)

And the poet and critic John Crowe Ransom gives the thought a cryptically and characteristically elegant variation: "Poetry is the kind of knowledge by which we must know that we have arranged that we shall not know otherwise." Perhaps this point about recognition might be carried further, to the extreme at which it would be seen to pose the problem of how poetry, which at its highest has always carried, at least implicitly, a kind of Platonism and claimed to give, if not knowledge itself, what was more important, a "form" to knowledge, can survive the triumph of scientific materialism and a positivism minded to skepticism about everything in the world except its own self (where it turns credulous, extremely). The poet's adjustment, over two or three centuries, to a Newtonian cosmos, Kantian criticism, the spectral universe portrayed by physics has conspicuously not been a happy one and has led alternately or simultaneously to the extremes of rejection of reason and speaking in tongues on the one hand and the hysterical claim that poetry will save the world on the other. But of this let the Protean parable speak as it will.

There is another part to the story of Menelaus and Proteus, for Menelaus asked another question: What happened to my friends who were with me at Troy? Proteus replies, "Son of Atreus, why enquire too closely of me on this? To know or learn what I know about it is not your need: I warn you that when you hear all the truth your tears will not be far behind . . . ." But he tells him all the same: "Of those others many went under; many came through . . . ." And Menelaus does indeed respond with tears of despair, until Proteus advises him to stop crying and get started on the journey home. So it sometimes happens in poetry, too: the sorrowful contemplation of what is, consoles, in the end, and heals, but only after the contemplative process has been gone through and articulated in the detail of its change:

> When to the sessions of sweet silent thought
> I summon up remembrance of things past,
> I sigh the lack of many a thing I sought,
> And with old woes new wail my dear time's waste;
> Then can I drown an eye, unused to flow,
> For precious friends hid in death's dateless night,
> And weep afresh love's long since cancell'd woe,
> And moan the expense of many a vanish'd sight.
> Then can I grieve at grievances foregone,

> And heavily from woe to woe tell o'er
> The sad account of fore-bemoaned moan,
> Which I new pay as if not paid before.
> But if the while I think on thee, dear friend,
> All losses are restor'd and sorrows end.
> (Shakespeare, Sonnet 30)

This poem, acknowledged to be a masterpiece by so many generations of readers, may stand as an epitome and emblem for the art altogether, about which it raises a question that must be put, although it cannot be satisfactorily and unequivocally answered: the question of whether poetry is a sacrament or a confidence game or both or neither. To reply firmly that poetry is not religion and must not promise what religion does is to preserve a useful distinction; nevertheless, the religions of the world, if they have nothing else in common, seem to be based on collections of sacred poems. Nor, at the other extreme, can any guarantee that poetry is not a confidence game be found in the often-heard appeal to the poet's "sincerity." One will never know whether Shakespeare wept all over the page while writing the 30th sonnet, though one inclines to doubt it, nor would it be to his credit if he did, nor to the reader's that he should know it or care to know it.

The nature of poetry

For one thing, the sonnet is obviously artful—that is, full of artifice—and even the artifice degenerates here and there into being artsy. "Then can I drown an eye, unused to flow." Surely that is poesy itself, at or near its worst, where the literal and the conventional, whatever their relations may have been for Shakespeare and the first reader of these sugar'd sonnets among his friends, now live very uncomfortably together (Ben Jonson's "Drink to me only with thine eyes" is a like example of this bathetic crossing of levels), though perhaps it has merely become unattractive as a result of changing fashions in diction.

Moreover, while the whole poem is uniquely Shakespearean, the bits and pieces are many of them common property of the age, what one writer called "joint stock company poetry." And the tricks are terribly visible, too; art is not being used to conceal art in such goings-on as "grieve at grievances" and "fore-bemoaned moan." "He who thus grieves will excite no sympathy," as Samuel Johnson sternly wrote of John Milton's style in the elegy "Lycidas," "he who thus praises will confer no honour."

Nor is that the worst of it. This man who so powerfully works on the reader's sympathies by lamenting what is past contrives to do so by thinking obsessively about litigation and, of all things, money; his hand is ever at his wallet, bidding adieu. He cannot merely "think" sweet silent thoughts about the past; no, he has to turn them into a court in "session," whereto he "summons" the probable culprit "remembrance"; when he "grieves," it is at a "grievance"—in the hands of the law again; finally, as with the sinners in Dante's *Divine Comedy,* his avarice and prodigality occupy two halves of the one circle: he bemoans his expenses while paying double the asking price.

And still, for all that, the poem remains beautiful; it continues to move both the young who come to it still innocent of their dear time's waste and the old who have sorrows to match its sorrows. As between confidence game and sacrament there may be no need to decide, as well as no possibility of deciding: elements of play and artifice, elements of true feeling, elements of convention both in the writing and in one's response to it, all combine to veil the answer. But the poem remains.

Effect of the poem as a whole

If it could be plainly demonstrated by the partisans either of unaided reason or revealed religion that poetry was metaphorical, mythological, and a delusion, while science, say, or religion or politics were real and true, then one might throw poetry away and live honestly though poorly on what was left. But, for better or worse, that is not the condition of man's life in the world; and perhaps men care for poetry so much—if they care at all—because, at last, it is the only one of man's many mythologies to be aware, and to make him aware, that it, and the others, are indeed mythological. The literary critic I.A. Richards, in a deep and searching consideration of this matter, concludes: "It is the privilege of poetry to preserve us from mistaking our notions either for things or for ourselves. Poetry is the completest mode of utterance."

The last thing Proteus says to Menelaus is strange indeed:

> You are not to die in Argos of the fair horse-pastures, not there to encounter death: rather will the Deathless Ones carry you to the Elysian plain, the place beyond the world .... There you will have Helen to yourself and will be deemed of the household of Zeus.

So the greatest of our poets have said, or not so much said, perhaps, as indicated by their fables, though nowadays people mostly sing a different tune. To be as the gods, to be rejoined with the beloved, the world forgotten .... Sacrament or con game? Homer, of course, is only telling an old story and promises mankind nothing; that is left to the priests to do; and in that respect poetry, as one critic puts it, must always be "a ship that is wrecked on entering the harbor." And yet the greatest poetry sings always, at the end, of transcendence; while seeing clearly and saying plainly the wickedness and terror and beauty of the world, it is at the same time humming to itself, so that one overhears rather than hears: All will be well. (H.Ne.)

Prosody

As it has come to be defined in modern criticism, the term prosody encompasses the study of all of the elements of language that contribute toward acoustic and rhythmic effects, chiefly in poetry but also in prose. The term derived from an ancient Greek word that originally meant a song accompanied by music or the particular tone or accent given to an individual syllable. Greek and Latin literary critics generally regarded prosody as part of grammar; it concerned itself with the rules determining the length or shortness of a syllable, with syllabic quantity, and with how the various combinations of short and long syllables formed the metres (*i.e.,* the rhythmic patterns) of Greek and Latin poetry. Prosody was the study of metre and its uses in lyric, epic, and dramatic verse. In sophisticated modern criticism, however, the scope of prosodic study has been expanded until it now concerns itself with what the 20th-century poet Ezra Pound called "the articulation of the total sound of a poem."

Prosody and prose rhythm

Prose as well as verse reveals the use of rhythm and sound effects; however, critics do not speak of "the prosody of prose" but of prose rhythm. The English critic George Saintsbury wrote *A History of English Prosody from the Twelfth Century to the Present* (3 vol., 1906–10), which treats English poetry from its origins to the end of the 19th century; but he dealt with prose rhythm in an entirely separate work, *A History of English Prose Rhythm* (1912). Many prosodic elements such as the rhythmic repetition of consonants (alliteration) or of vowel sounds (assonance) occur in prose; the repetition of syntactical and grammatical patterns also generates rhythmic effect. Traditional rhetoric, the study of how words work, dealt with acoustic and rhythmic techniques in Classical oratory and literary prose. But although prosody and rhetoric intersected, rhetoric dealt more exactly with verbal meaning than with verbal surface. Rhetoric dealt with grammatical and syntactical manipulations and with figures of speech; it categorized the kinds of metaphor. Modern critics, especially those who practice the New Criticism, might be considered rhetoricians in their detailed concern with such devices as irony, paradox, and ambiguity. These subjects are discussed at greater length below in the section *Literary criticism* and in the article RHETORIC.

This section considers prosody chiefly in terms of the English language—the only language that all of the readers of this article may be assumed to know. Some examples are given in other languages to illustrate particular points about the development of prosody in those languages; because these examples are pertinent only for their rhythm and sound, and not at all for their meaning, no translations are given. A further general discussion of the development of prosodic elements will be found in the section above, *Poetry.*

ELEMENTS OF PROSODY

As a part of modern literary criticism, prosody is concerned with the study of rhythm and sound effects as they occur in verse and with the various descriptive, historical,

and theoretical approaches to the study of these structures.

Scansion. The various elements of prosody may be examined in the aesthetic structure of prose. The celebrated opening passage of Charles Dickens' novel *Bleak House* (1853) affords a compelling example of prose made vivid through the devices of rhythm and sound:

> Fog everywhere. Fog up the river, where it flows among green aits and meadows; fog down the river, where it rolls defiled among the tiers of shipping, and the waterside pollutions of a great (and dirty) city. Fog on the Essex Marshes, fog on the Kentish heights. Fog creeping into the cabooses of collier-brigs; fog lying out on the yards, and hovering in the rigging of great ships; fog drooping on the gunwales of barges and small boats. Fog in the eyes and throats of ancient Greenwich pensioners, wheezing by the firesides of their wards; fog in the stem and bowl of the afternoon pipe of the wrathful skipper . . . .

Two phrases of five syllables each ("Fog everywhere"; "Fog up the river") establish a powerful rhythmic expectation that is clinched in repetition:

> . . . fog down the river . . . . Fog on the Essex . . . , fog on the Kentish . . . . Fog creeping into . . . ; . . . fog drooping on the . . . .

This phrase pattern can be scanned; that is, its structure of stressed and unstressed syllables might be translated into visual symbols:

´ ˘ ˘ ´ ˘
Fog down the ri ver.

(This scansion notation uses the following symbols: the acute accent [´] to mark metrically stressed syllables; the breve [˘] to mark metrically weak syllables; a single line [|] to mark the divisions between feet [*i.e.,* basic combinations of stressed and unstressed syllables]; a double line [‖] to mark the caesura, or pause in the line; a rest [^] to mark a syllable metrically expected but not actually occurring.) Such a grouping constitutes a rhythmic constant, or cadence, a pattern binding together the separate sentences and sentence fragments into a long surge of feeling. At one point in the passage, the rhythm sharpens into metre; a pattern of stressed and unstressed syllables falls into a regular sequence:

Cadence and metre

´ ˘ ˘ ´ ˘ ´ ˘ ´ ˘ ˘
Fog on the | Es sex | mar shes, ‖ fog on the |
´ ˘ ´
Ken tish | heights.

The line is a hexameter (*i.e.,* it comprises six feet), and each foot is either a dactyl (´˘˘) or a trochee (´˘).

The passage from Dickens is strongly characterized by alliteration, the repetition of stressed consonantal sounds:

> Fog *c*reeping into the *c*abooses of *c*ollier-brigs;

and by assonance, the patterned repetition of vowel sounds:

> . . . f*o*g d*o*wn the river, where it r*o*lls defiled am*o*ng . . . .

Here the vowel sounds are symmetrically distributed: short, long and long, short. Thus, it is clear that Dickens uses loosely structured rhythms, or cadences, an occasional lapse into metre, and both alliteration and assonance.

The rhythm and sound of all prose are subject to analysis; but compared with even the simplest verse, the "prosodic" structure of prose seems haphazard, unconsidered. The poet organizes his structures of sound and rhythm into rhyme, stanzaic form, and, most importantly, metre. Indeed, the largest part of prosodical study is concerned with the varieties of metre, the nature and function of rhyme, and the ways in which lines of verse fall into regular patterns or stanzas. An analysis of "Vertue" by the 17th-century English poet George Herbert reveals how the elements of prosody combine into a complex organism, a life sustained by the technical means available to the poet. When the metre is scanned with the symbols, it can be seen (and heard) how metre in this poem consists of the regular recurrence of feet, how each foot is a pattern of phonetically stressed and unstressed syllables.

Scansion

˘ ´ ˘ ´ ˘ ´ ˘ ´
1 Sweet day, | so cool, | so calm, | so bright,
˘ ´ ˘ ´ ˘ ´ ˘ ´
2 The bri | dall of | the earth | and skie:
˘ ´ ˘ ´ ˘ ´ ˘ ´
3 The dew | shall weep | thy fall | to-night;
˘ ´ ˘ ´
4 For thou | must die.
˘ ´ ˘ ´ ´ ˘ ˘ ´
5 Sweet rose, | whose hue | an grie | and brave
´ ˘ ˘ ´ ˘ ´ ˘ ´
6 Bids the | rash ga | zer wipe | his eye:
˘ ´ ˘ ´ ˘ ´ ˘ ´
7 Thy root | is ev | er in | its grave,
˘ ´ ˘ ´
8 And thou | must die.
˘ ´ ´ ˘ ˘ ´ ˘ ´ ˘
9 Sweet spring, | full of | sweet dayes | and ro | ses,
˘ ´ ˘ ´ ˘ ´ ˘ ´
10 A box | where sweets | com pac | ted lie;
˘ ´ ˘ ´ ˘ ´ ˘ ´ ˘
11 My mu | sick shows | ye have | your clo | ses,
˘ ´ ˘ ´
12 And all | must die.
´ ˘ ˘ ´ ˘ ´ ˘ ´
13 Onely | a sweet | and ver | tuous soul,
˘ ´ ˘ ´ ˘ ´ ˘ ´
14 Like sea | son'd tim | ber, ne | ver gives;
˘ ´ ˘ ´ ˘ ´ ˘ ´
15 But though | the whole | world turn | to coal,
˘ ´ ˘ ´
16 Then chief | ly lives.

The basic prosodic units are the foot, the line, and the stanza. The recurrence of similar feet in a line determines the metre; here there are three lines consisting of four iambic feet (*i.e.,* of four units in which the common pattern is the iamb—an unstressed syllable followed by a stressed syllable), which are followed by a line consisting of two iambic feet. Thus the stanza or recurring set of lines consists of three iambic tetrameters followed by one iambic dimeter. The stanzaic form is clinched by the use of rhyme; in "Vertue" the first and third and second and fourth lines end with the same sequence of vowels and consonants: bright/night, skie/pie, brave/grave, eye/pie, etc. It should be observed that the iambic pattern (˘´) is not invariable; the third foot of line 5, the first foot of line 6, the second foot of line 9, and the first foot of line 13 are reversals of the iambic foot or trochees (´˘). These reversals are called substitutions; they provide tension between metrical pattern and meaning, as they do in these celebrated examples from Shakespeare:

˘ ´ ˘ ´ ˘ ´ ´ ˘ ˘ ´ ˘
To be, | or not | to be, ‖ *that is* | the ques tion . . . .

Hamlet

˘ ´ ˘ ´ ´ ˘ ˘ ´ ˘ ´
His sil | ver skin | *laced with* | his gol | den blood . . .

Macbeth

Meaning, pace, and sound. Scansion reveals the basic metrical pattern of the poem; it does not, however, tell everything about its prosody. The metre combines with other elements, notably propositional sense or meaning, pace or tempo, and such sound effects as alliteration, assonance, and rhyme. In the fifth line of "Vertue," the reversed third foot occurring at "angry" brings that word into particular prominence; the disturbance of the metre combines with semantic reinforcement to generate a powerful surge of feeling. Thus, the metre here is expressive. The pace of the lines is controlled by the length of number of syllables and feet, line 5 obviously takes longer to read or recite. The line contains more long vowel sounds:

> Sw*ee*t r*o*se, wh*o*se h*ue* angr*ie* and br*a*ve . . .

Vowel length is called quantity. In English verse, quantity cannot by itself form metre although a number of English poets have experimented with quantitative verse. Generally speaking, quantity is a rhythmical but not a metrical feature of English poetry; it can be felt but it cannot be precisely determined. The vowel sounds in "Sweet rose" may be lengthened or shortened at will. No such options are available, however, with the stress patterns of words;

Quantity

/ ⏑ ⏑ /

the word an-gry cannot be read an-gry.

Assonance takes into account the length and distribution of vowel sounds. A variety of vowel sounds can be noted in this line:

Sweet day, so cool, so calm, so bright . . .

To borrow a term from music, the line modulates from ēē, through ā, ōō, ă, to ī. Alliteration takes into account the recurrence and distribution of consonants:

so *c*ool, so *c*alm . . .
*S*weet *s*pring . . .

Rhyme normally occurs at the ends of lines; "Vertue" reveals, however, a notable example of interior rhyme, or rhyme within the line:

My musick *shows* ye have your *closes* . . .

Types of metre. *Syllable-stress metres.* It has been shown that the metre of "Vertue" is determined by a pattern of stressed and unstressed syllables arranged into feet and that a precise number of feet determines the measure of the line. Such verse is called syllable-stress verse (in some terminologies accentual-syllabic) and was the norm for English poetry from the beginning of the 16th century to the end of the 19th century. A line of syllable-stress verse is made up of either two-syllable (disyllabic) or three-syllable (trisyllabic) feet. The disyllabic feet are the iamb and the trochee (noted in the scansion of "Vertue"); the trisyllabic feet are the dactyl (´⏑⏑) and anapest (⏑⏑´).

Following are illustrations of the four principal feet found in English verse:

	⏑ /
iambic	be hold
	/ ⏑
trochaic	ti ger
	/ ⏑ ⏑
dactylic	des per ate
	⏑ ⏑ /
anapestic	un der stand

Some theorists also admit the spondaic foot (´´) and pyrrhic foot (⏑⏑) into their scansions; however, spondees and pyrrhics occur only as substitutions for other feet, never as determinants of a metrical pattern:

/ / / ⏑ ⏑ / / ⏑ /

When to | the ses | sions of | sweet si | lent thought . . .

It has been noted that four feet make up a line of tetrameter verse; a line consisting of one foot is called monometer, of two dimeter, of three trimeter, of five pentameter, of six hexameter, and of seven heptameter. Lines containing more than seven feet rarely occur in English poetry.

The following examples illustrate the principal varieties of syllable-stress metres and their scansions:

Iambic (pentameter)

⏑ / ⏑ / ⏑ / ⏑ / ⏑ /
Then say | not Man's | im per | fect, || Heaven | in fault;
⏑ / ⏑ / ⏑ / ⏑ / ⏑ /
Say ra | ther, || Man's | as per | fect as | he ought:
⏑ / ⏑ / ⏑ / ⏑ / ⏑ /
His know | ledge meas | ured || to | his state | and place,
⏑ / ⏑ / ⏑ / ⏑ / ⏑ /
His time | a mo | ment, || and | a point | his space.

Alexander Pope, *An Essay on Man* (1733–34)

Trochaic (dimeter)

/ ⏑ / ⏑
Could I | catch that
/ ⏑ / ⏑
Nim ble | trai tor
/ ⏑ / ⏑
Scorn ful | Lau ra,
/ ⏑ / ⏑
Swift-foot | Lau ra,
/ ⏑ / ⏑
Soon then | would I
/ ⏑ / ⏑
Seek a | venge ment.

Thomas Campion (1602)

Dactylic (tetrameter)

/ ⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ / ⏑
Af ter the | pangs of a | des per ate | Lover, ∧
/ ⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ /
When day and | night I have | sigh'd all in | vain, ∧ ∧
/ ⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ / ⏑
Ah what a | pleas ure it | is to dis | co ver ∧
/ ⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ /
In her eyes | pi ty, who | cau ses my | pain! ∧ ∧

John Dryden, *An Evening's Love* (1671)

Anapestic (tetrameter)

⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ /
The As syr | ian came down | like a wolf | on the fold,
⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ /
And his co | horts were gleam | ing in pur | ple and gold;
⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ /
And the sheen | of their spears | was like stars | on the sea,
⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ / ⏑ ⏑ /
When the blue | wave rolls night | ly on deep | Ga li lee.

Lord Byron, "The Destruction of Sennacherib"(1815)

Syllable stress became more or less established in the poetry of Geoffrey Chaucer (*c.* 1340–1400). In the century that intervened between Chaucer and the early Tudor poets, syllable-stress metres were either ignored or misconstrued. By the end of the 16th century, however, the now-familiar iambic, trochaic, dactylic, and anapestic metres became the traditional prosody for English verse.

Strong-stress metres. In the middle of the 19th century, with Walt Whitman's free verse and Gerard Manley Hopkins' extensive metrical innovations, the traditional prosody was challenged. Antecedent to the syllable-stress metres was the strong-stress metre of Old English and Middle English poetry. Strong-stress verse is measured by count of stresses alone; the strong stresses are usually constant, but the number of unstressed syllables may vary considerably.

Challenges to traditional prosody in the 19th century

Strong-stress verse survives in nursery rhymes and children's counting songs:

/ / / /
One, two, || buckle my shoe;
/ / / /
Three, four, || knock at the door;
/ / / /
Five, six, || pick up sticks . . .

The systematic employment of strong-stress metre can be observed in the Old English epic poem *Beowulf* (*c.* 1000) and in William Langland's vision-poem, *Piers Plowman* ('A' Text, *c.* 1362):

/ / / /
In a somer sesun || whon softe was the sonne,
/ / / /
I schop me in-to a schroud || a scheep as I were;
/ / / /
In habite of an hermite || un-holy of werkes,
/ / / /
Wende I wydene in this world || wondres to here.

These lines illustrate the structural pattern of strong-stress metre. Each line divides sharply at the caesura (||), or medial pause; on each side of the caesura are two stressed syllables strongly marked by alliteration.

Strong-stress verse is indigenous to the Germanic languages with their wide-ranging levels of stressed syllables and opportunities for alliteration. Strong-stress metre was normative to Old English and Old Germanic heroic poetry, as well as to Old English lyric poetry. With the rising influence of French literature in the 12th and 13th centuries, rhyme replaced alliteration and stanzaic forms replaced the four-stress lines. But the strong-stress rhythm persisted; it can be felt in the anonymous love lyrics of the 14th century and in the popular ballads of the 15th century.

"Lord Randal" can be comfortably scanned to show a line of mixed iambic and anapestic feet; it clearly reveals, however, a four-stress structure:

/ / / /
'O where ha' you been, || Lord Randal, my son?

´ ´ ´ ´
And where ha' you been, || my handsome young man?'
´ ´ ´ ´
'I ha' been at the greenwood; || mother, mak my bed soon,
´ ´ ´ ´
For I'm wearied wi' huntin', || and fain wad lie down.'

A number of 20th-century poets, including Ezra Pound, T.S. Eliot, and W.H. Auden, have revived strong-stress metre. The versification of Pound's *Cantos* and Eliot's *Four Quartets* (1943) shows the vitality of the strong-stress, or, as they are often called, "native," metres.

Syllabic metres. Most of English poetry is carried by the strong-stress and syllable-stress metres. Two other kinds of metres must be mentioned: the purely syllabic metres and the quantitative metres. The count of syllables determines the metres of French, Italian, and Spanish verse. In French poetry the alexandrine, or 12-syllabled line, is a dominant metrical form:

The dominant alexandrine

O toi, qui vois la honte où je suis descendue,
Implacable Vénus, suis-je assez confondue?
Tu ne saurais plus loin pousser ta cruauté.
Ton triomphe est parfait; tous tes traits ont porté.
Racine, *Phèdre* (1677)

Stress and pause in these lines are variable; only the count of syllables is fixed. English poets have experimented with syllabic metres; the Tudor poet Thomas Wyat's translations from Petrarch's Italian poems of the 14th century attempted to establish a metrical form based on a decasyllabic or 10-syllabled line:

The long love that in my thought doth harbor,
And in my heart doth keep his residence,
Into my face presseth with bold pretense
And there encampeth, spreading his banner.
"The Lover for Shamefastness Hideth . . ." (1557)

Most ears can detect that these lines waver between syllabic and syllable-stress metre; the second line falls into a pattern of iambic feet. Most ears also discover that the count of syllables alone does not produce any pronounced rhythmic interest; syllabic metres in English generate a prosody more interesting to the eye than to the ear.

Quantitative metres. Quantitative metres determine the prosody of Greek and Latin verse. Renaissance theorists and critics initiated a confused and complicated argument that tried to explain European poetry by the rules of Classical prosody and to draft laws of quantity by which European verse might move in the hexameters of the ancient Roman poets Virgil or Horace. Confusion was compounded because both poets and theorists used the traditional terminology of Greek and Latin prosody to describe the elements of the already existing syllable-stress metres; iambic, trochaic, dactylic, and anapestic originally named the strictly quantitative feet of Greek and Latin poetry. Poets themselves adapted the metres and stanzas of Classical poetry to their own languages; whereas it is not possible here to trace the history of Classical metres in European poetry, it is instructive to analyze some attempts to make English and German syllables move to Greek and Latin music. Because neither English nor German has fixed rules of quantity, the poets were forced to revise the formal schemes of the Classical paradigms in accordance with the phonetic structure of their own language.

A metrical paradigm much used by both Greek and Latin poets was the so-called Sapphic stanza. It consisted of three quantitative lines that scanned

– ∪ – – – ∪ ∪ – ∪ – ∪,

followed by a shorter line, called an Adonic,

– ∪ ∪ – – .

Sapphics in English and German

"Sapphics" by the 19th-century English poet Algernon Charles Swinburne shows the Sapphic metre and stanza in English:

All the night sleep came not upon my eyelids,
Shed not dew, nor shook nor unclosed a feather,
Yet with lips shut close and with eyes of iron
Stood and beheld me . . .
Saw the white implacable Aphrodite,
Saw the hair unbound and the feet unsandalled
Shine as fire of sunset on western waters;
Saw the reluctant . . . .

The same metre and stanza in German are found in "Sapphische Ode," by the 19th-century poet Hans Schmidt, which was beautifully set to music by Johannes Brahms (Opus 94, No. 4):

Rosen brach ich nachts mir am dunklen Hage;
süsser hauchten Duft sie, als je am Tage;
doch verstreuten reich die bewegten Äste
Tau den mich nässte.
Auch der Küsse Duft mich wie nie berückte,
die ich nachts vom Strauch deiner Lippen pflückte:
doch auch dir, bewegt im Gemüt gleich jenem,
tauten die Tränen.

Quantitative metres originated in Greek, a language in which the parts of speech appear in a variety of inflected forms (*i.e.,* changes of form to indicate distinctions in case, tense, mood, number, voice, and others). Complicated metrical patterns and long, slow-paced lines developed because the language was hospitable to polysyllabic metrical feet and to the alternation of the longer vowels characterizing the root syllables and the shorter vowels characterizing the inflected case-endings. The Classical metres can be more successfully adapted to German than to English because English lost most of its inflected forms in the 15th century, while German is still a highly inflected language. Thus Swinburne's "Sapphics" does not move as gracefully, as "naturally" as Schmidt's. A number of German poets, notably Goethe and Friedrich Hölderlin, both of the early 19th century, made highly successful use of the Classical metres. English poets, however, have never been able to make English syllables move in the ancient metres with any degree of comfort or with any sense of vital rhythmic force.

The American poet Henry Wadsworth Longfellow adapted the Classical hexameter for his *Evangeline* (1847):

´ ∪ ∪ ´ ∪ ∪ ´ ∪ ∪ ´ ∪ ∪ ´ ∪
This is the | for est pri | me val. The | mur mu ring | pines and
∪ ´ ∪
the | hem locks . . .

In Virgil's *Aeneid,* Longfellow's Classical model, the opening line scans:

– ∪ ∪ – ∪ ∪ – – – ∪ – ∪ ∪ – –
Ar ma vi rum que ca no, Troj ae qui pri mus ab or is

The rules determining length of syllable in Classical Greek and Latin poetry are numerous and complicated; they were established by precise grammatical and phonetic conventions. No such rules and conventions obtain in English; Robert Bridges, the British poet laureate and an authority on prosody, remarked in his *Poetical Works* (1912) that the difficulty of adapting English syllables to the Greek rules is "very great, and even deterrent." Longfellow's hexameter is in reality a syllable-stress line of five dactyls and a final trochee; syllabic quantity plays no part in determining the metre.

PROSODIC STYLE

The analysis of prosodic style begins with recognizing the metrical form the poet uses. Is he writing syllable-stress, strong-stress, syllabic, or quantitative metre? Or is he using a nonmetrical prosody? Again, some theorists would not allow that poetry can be written without metre; the examples of Whitman and many 20th-century innovators, however, have convinced most modern critics that a nonmetrical prosody is not a contradiction in terms but an obvious feature of modern poetry. Metre has not disappeared as an important element of prosody; indeed, some of the greatest poets of the modern period—William Butler Yeats, T.S. Eliot, Ezra Pound, Wallace Stevens—revealed themselves as masters of the traditional metres. They also experimented with newer prosodies based on prose cadences, on expansions of the blank-verse line, and revivals of old forms—such as strong-stress and ballad metres. Also noteworthy are the "visual" prosodies fostered by the poets of the Imagist movement and by such experimenters as e.e. cummings. Cummings revived the practice of certain 17th-century poets (notably George Herbert) of "shaping" the poem by typographic arrangements.

Non-metrical prosody

The prosodic practice of poets has varied enormously with the historical period, the poetic genre, and the poet's

individual style. In English poetry, for example, during the Old English period (to 1100), the strong-stress metres carried both lyric and narrative verse. In the Middle English period (from *c.* 1100 to *c.* 1500), stanzaic forms developed for both lyric and narrative verse. The influence of French syllable counting pushed the older stress lines into newer rhythms; Chaucer developed for *The Canterbury Tales* a line of 10 syllables with alternating accent and regular end rhyme—an ancestor of the heroic couplet. The period of the English Renaissance (from *c.* 1500 to 1660) marks the fixing of syllable-stress metre as normative for English poetry. Iambic metre carried three major prosodic forms: the sonnet, the rhyming couplet, and blank verse. The sonnet was the most important of the fixed stanzaic forms. The iambic pentameter rhyming couplet (later known as the heroic couplet) was used by Christopher Marlowe for his narrative poem *Hero and Leander* (1598); by John Donne in the early 17th century for his satires, his elegies, and his longer meditative poems. Blank verse (unrhymed iambic pentameter), first introduced into English in a translation by Henry Howard, earl of Surrey, published in 1557, became the metrical norm for Elizabethan drama. The period of the Renaissance also saw the refinement of a host of lyric and song forms; the rapid development of English music during the second half of the 16th century had a salutary effect on the expressive capabilities of poetic rhythms.

The personal element. A poet's choice of a prosody obviously depends on what his language and tradition afford; these are primary considerations. The anonymous author of the Old English poem *Deor* used the conventional four-stress metric available to him; but he punctuated groups of lines with a refrain:

þaes ofereode: þisses swa maeg!
(that passed away: this also may!)

The refrain adds something to the prosodic conventions of regulated stress, alliteration, and medial pause: a sense of a smaller and sharper rhythmic unit within the larger rhythms of the given metre. While the poet accepts from history his language and from poetic convention the structure of his metre, he shapes his own style through individual modifications of the carrying rhythms. When critics speak of a poet's "voice," his personal tone, they are also speaking of his prosodic style.

Style through tension

Prosodic style must be achieved through a sense of tension; it is no accident that the great masters of poetic rhythm work against the discipline of a given metrical form. In his sonnets, Shakespeare may proceed in solemn iambic regularity, creating an effect of measured progression through time and its legacy of suffering and despair:

No longer mourn for me when I am dead
Than you shall hear the surly sullen bell
Give warning to the world that I am fled . . . .
"Sonnet 71"

Or he may wrench the metre and allow the reader to feel the sudden violence of his feelings, the power of a conviction raised to a command:

′ ∪ ′ ∪ ∪ ′ ∪ ∪ ′ ′
Let me | not to | the mar | riage of | true minds
∪ ′ ∪ ′ ∪ ′ ′ ∪ ∪ ′
Ad mit | im pe | di ments. | Love is | not love . . . .
"Sonnet 116"

The first two feet of the first line are trochaic reversals; the last two feet comprise a characteristic pyrrhic-spondaic formation. A trochaic substitution is quite normal in the first foot of an iambic pentameter line; a trochaic substitution in the second foot, however, creates a marked disturbance in the rhythm. There is only one "normal" iambic foot in the first line; this line runs over (or is enjambed) to the second line with its three consecutive iambic feet followed by a strong caesura and reversed fourth foot. These lines are, in Gerard Manley Hopkins' terms, metrically "counter-pointed"; trochees, spondees, and pyrrhics are heard against a ground rhythm of regular iambics. Without the ground rhythm, Shakespeare's expressive departures would not be possible.

A poet's prosodic style may show all of the earmarks of revolt against prevailing metrical practice. Whitman's celebrated "free verse" marks a dramatic break with the syllable-stress tradition; he normally does not count syllables, stresses, or feet in his long sweeping lines. Much of his prosody is rhetorical; that is, Whitman urges his language into rhythm by such means as anaphora (*i.e.,* repetition at the beginning of successive verses) and the repetition of syntactical units. He derives many of his techniques from the example of biblical verses, with their line of various types of parallelism. But he often moves toward traditional rhythms; lines fall into conventional parameters:

Whitman's innovations

O past! O happy life! O songs of joy!
"Out of the Cradle Endlessly Rocking" (1859)

Or they fall more often into dissyllabic hexameters:

Borne through the smoke of the battles and pierc'd
with missiles I saw them . . . .
"When Lilacs Last in the Dooryard Bloom'd" (1865–66)

Despite the frequent appearance of regular metrical sequences, Whitman's lines cannot be scanned by the usual graphic method of marking syllables and feet; his prosody, however, is fully available to analysis. The shape on the page of the lines below (they comprise a single strophe or verse unit) should be noted, specifically the gradual elongation and sudden diminution of line length. Equally noteworthy are the repetition of the key word *carols,* the alliteration of the *s* sounds, and the use of words in falling (trochaic) rhythm, "lagging," "yellow," "waning":

Shake out carols!
Solitary here, the night's carols!
Carols of lonesome love! death's carols!
Carols under that lagging, yellow, waning moon!
O under that moon where she droops almost down into
the sea!
O reckless despairing carols.
"Out of the Cradle"

No regular metre moves these lines; but a clearly articulated rhythm—produced by shape, thematic repetitions, sound effects, and patterns of stress and pause—defines a prosody.

Whitman's prosody marks a clear break with previous metrical practices. Often a new prosody modifies an existing metrical form or revives an obsolete one. In "Gerontion" (1920), T.S. Eliot adjusted the blank-verse line to the emotionally charged, prophetic utterance of his persona, a spiritually arid old man:

After such knowledge, what forgiveness? Think now
History has many cunning passages, contrived corridors
And issues, deceives with whispering ambitions,
Guides us by vanities. Think now . . .
(From T.S. Eliot, *Collected Poems 1909–1962,*
Harcourt Brace Jovanovich, Inc.)

The first three lines expand the pentameter line beyond its normal complement of stressed and unstressed syllables; the fourth line contracts, intensifying the arc of feeling. Both Pound and Eliot used stress prosodies. Pound counted out four strong beats and used alliteration in his brilliant adaptation of the old English poem "The Seafarer" (1912):

The stress prosodies of Pound and Eliot

Chill its chains are; chafing sighs
Hew my heart round and hunger begot
Mere-weary mood. Lest man known not
That he on dry land loveliest liveth . . .
(From Ezra Pound, *Personae,* Copyright 1926 by
Ezra Pound. Reprinted by permission of New Directions Publishing Corporation.)

He uses a similar metric for the energetic opening of his "Canto I." Eliot mutes the obvious elements of the form in the celebrated opening of *The Waste Land* (1922):

April is the cruellest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
(From T.S. Eliot, *Collected Poems 1909–1962,*
Harcourt Brace Jovanovich, Inc.)

Here is the "native metre" with its falling rhythm, elegiac tone, strong pauses, and variably placed stresses. If this is free verse, its freedoms are most carefully controlled. "No verse is free," said Eliot, "for the man who wants to do a good job."

The prosodic styles of Whitman, Pound, and Eliot—

though clearly linked to various historical antecedents—are innovative expressions of their individual talents. In a sense, the prosody of every poet of genius is unique; rhythm is perhaps the most personal element of the poet's expressive equipment. Alfred Lord Tennyson and Robert Browning, English poets who shared the intellectual and spiritual concerns of the Victorian age, are miles apart in their prosodies. Both used blank verse for their dramatic lyrics, poems that purport to render the accents of real men speaking. The blank verse of Tennyson's "Ulysses" (1842) offers smoothly modulated vowel music, carefully spaced spondaic substitutions, and unambiguous pentameter regularity:

The prosody of Tennyson and Browning

> The long day wanes; the slow moon climbs; the deep
> Moans round with many voices. Come, my friends,
> 'Tis not too late to seek a newer world.

Browning's blank verse aims at colloquial vigour; its "irregularity" is a function not of any gross metrical violation—it always obeys the letter of the metrical law—but of the adjustment of abstract metrical pattern to the rhythms of dramatic speech. If Tennyson's ultimate model is Milton's Baroque prosody with its oratorical rhythms, Browning's model was the quick and nervous blank verse of the later Elizabethan dramatists. Characteristic of Browning's blank verse are the strong accents, involuted syntax, pregnant caesuras, and headlong energy in "The Bishop Orders His Tomb at St. Praxed's Church" (1845):

> Vanity, saith the preacher, vanity!
> Draw round my bed: is Anselm keeping back?
> Nephews—sons mine . . . ah God, I know not! Well—
> She, men would have to be your mother once,
> Old Gandolf envied me, so fair she was!

Influence of period and genre. In the lyric genres, the rhythms of the individual poet—or, in the words of the 20th-century American poet Robert Lowell, "the person himself"—can be heard in the prosody. In the long poem, the dramatic, narrative, and didactic genres, a period style is more likely to be heard in prosody. The blank-verse tragedy of the Elizabethan and Jacobean dramatists, the blank verse of Milton's *Paradise Lost* (1667) and its imitators in the 18th century (James Thomson and William Cowper), and the heroic couplet of Neoclassical satiric and didactic verse, each, in different ways, defines the age in which these prosodies flourished. The flexibility and energy of the dramatic verse of Marlowe, Shakespeare, and John Webster reflect the later Renaissance with its nervous open-mindedness, its obsessions with power and domination, and its lapses into despair. Miltonic blank verse, based on Latin syntax and adaptations of the rules of Latin prosody, moved away from the looseness of the Elizabethans and Jacobeans toward a more ceremonial style. It is a Baroque style in that it exploits the musical qualities of sounds for their ornamental values. The heroic couplet, dominating the poetry of the entire 18th century, was unequivocally a prosodic period style; its elegance and epigrammatic precision entirely suited an age that valued critical judgment, satiric wit, and the powers of rationality.

Prosody in dramatic verse

It is in dramatic verse, perhaps, that a prosody shows its greatest vitality and clarity. Dramatic verse must make a direct impression not on an individual reader able to reconsider and meditate on what he has read but on an audience that must immediately respond to a declaiming actor or a singing chorus. The ancient Greek dramatists developed two distinct kinds of metres: "stichic" forms (*i.e.*, consisting of "stichs," or lines, as metrical units) such as the iambic trimeter for the spoken dialogues; and lyric, or strophic, forms (*i.e.*, consisting of stanzas), of great metrical intricacy, for the singing and chanting of choruses. Certain of the Greek metres developed a particular ethos; characters of low social standing never were assigned metres of the lyric variety. Similar distinctions obtained in Elizabethan drama. Shakespeare's kings and noblemen speak blank verse; comic characters, servants, and country bumpkins discourse in prose; clowns, romantic heroines, and supernatural creatures sing songs. In the early tragedy *Romeo and Juliet,* the chorus speaks in "excellent conceited" sonnets: in what was one of the most popular and easily recognized lyric forms of the period.

The metrical forms used by ancient and Renaissance dramatists were determined by principles of decorum. The use or non-use of a metrical form (or the use of prose) was a matter of propriety; it was important that the metre be suitable to the social status and ethos of the individual character as well as be suitable to the emotional intensity of the particular situation. Decorum, in turn, was a function of the dominant Classical and Neoclassical theories of imitation.

THEORIES OF PROSODY

Ancient critics like Aristotle and Horace insisted that certain metres were natural to the specific poetic genres; thus, Aristotle (in the *Poetics*) noted that "Nature herself, as we have said, teaches the choice of the proper measure." In epic verse the poet should use the heroic measure (dactylic hexameter) because this metre most effectively represents or imitates such qualities as grandeur, dignity, and high passion. Horace narrowed the theory of metrical decorum, making the choice of metre prescriptive; only an ill-bred and ignorant poet would treat comic material in metres appropriate to tragedy. Horace prepared the way for the legalisms of the Renaissance theorists who were quite willing to inform practicing poets that they used "feete without joyntes," in the words of Roger Ascham, Queen Elizabeth's tutor, and should use the quantitative metres of Classical prosody.

The Middle Ages. During the Middle Ages little of importance was added to actual prosodic theory; in poetic practice, however, crucial developments were to have important ramifications for later theorists. From about the second half of the 6th century to the end of the 8th century, Latin verse was written that no longer observed the rules of quantity but was clearly structured on accentual and syllabic bases. This change was aided by the invention of the musical sequence; it became necessary to fit a musical phrase to a fixed number of syllables, and the older, highly complex system of quantitative prosody could not be adapted to simple melodies that must be sung in sequential patterns. In the musical sequence lies the origin of the modern lyric form.

The musical sequence

The 9th-century hymn "Ave maris stella" is a striking instance of the change from quantitative to accentual syllabic prosody; each line contains three trochaic feet determined not by length of syllable but by syllabic intensity or stress:

> Ave maris stella
> Dei mater alma
> atque semper virgo,
> felix caeli porta.
>
> Sumens illud *Ave*
> Gabrielis ore,
> funda nos in pace,
> mutans nomen Evae.

The rules of quantity have been disregarded or forgotten; rhyme and stanza and a strongly felt stress rhythm have taken their place. In the subsequent emergence of the European vernacular literatures, poetic forms follow the example of the later Latin hymns. The earliest art lyrics, those of the Provençal troubadours of the 12th and 13th centuries, show the most intricate and ingenious stanzaic forms. Similarly, the Goliardic songs of the *Carmina Burana* (13th century) reveal a rich variety of prosodic techniques; this "Spring-song" embodies varying lines of trochees and iambs and an ababcdccd rhyme scheme:

> Ver redit optatum
> Cum gaudio,
> Flore decoratum
> Purpureo;
> Aves edunt cantus
> Quam dulciter!
> Revirescit nemus,
> Cantus est amoenus
> Totaliter.

The Renaissance. Renaissance prosodic theory had to face the fact of an accomplished poetry in the vernacular that was not written in metres determined by "rules" handed down from the practice of Homer and Virgil. Nevertheless, the classicizing theorists of the 16th century made a determined attempt to explain existing poetry

Rules of "trewe versifying"

by the rules of short and long and to draft "laws" by which modern verse might move in Classical metres. Roger Ascham, in *The Scholemaster* (1570), attacked "the Gothic . . . barbarous and rude Ryming" of the early Tudor poets. He admitted that Henry Howard, earl of Surrey, did passably well as a poet but complained that Surrey did not understand "perfite and trewe versifying"; that is, Surrey did not compose his English verses according to the principles of Latin and Greek quantitative prosody.

Ascham instigated a lengthy argument, continued by succeeding theorists and poets, on the nature of English prosody. Sir Philip Sidney, Gabriel Harvey, Edmund Spenser, and Thomas Campion all (to use Saintsbury's phrase) committed whoredom with the enchantress of quantitative metric. While this hanky-panky had no adverse effect on poetry itself (English poets went on writing verses in syllable-stress, the prosody most suitable to the language), it produced misbegotten twins of confusion and discord, whose heirs, however named, are still apparent today. Thus, those who still talk about "long and short" (instead of stressed and unstressed), those who perpetuate a punitive prosodic legalism, and those who regard prosody as an account of what poets should have done and did not, trace their ancestry back to Elizabethan dalliance and illicit classicizing.

Although Renaissance prosodic theory produced scarcely anything of value to either literary criticism or poetic technique—indeed, it did not even develop a rational scheme for scanning existing poetry—it raised a number of important questions. What were the structural principles animating the metres of English verse? What were the aesthetic nature of prosody and the functions of metre? What were the connections between poetry and music? Was poetry an art of imitation (as Aristotle and all of the Neoclassical theorists had maintained), and was its sister art painting; or was poetry (as Romantic theory maintained) an art of expression, and prosody the element that produced (in Coleridge's words) the sense of musical delight originating (in T.S. Eliot's words) in the auditory imagination?

Pope's doctrine of imitation

The 18th century. Early in the 18th century, Pope affirmed, in his *Essay on Criticism* (1711), the classic doctrine of imitation. Prosody was to be more nearly onomatopoetic; the movement of sound and metre should represent the actions they carry:

> 'Tis not enough no harshness gives offence,
> The sound must seem an Echo to the sense:
> Soft is the strain when Zephyr gently blows,
> And the smooth stream in smoother numbers flows;
> But when loud surges lash the sounding shoar,
> The hoarse, rough verse should like the torrent roar.
> When Ajax strives some rock's vast weight to throw,
> The line too labours, and the words move slow;
> Not so, when swift Camilla scours the plain,
> Flies o'er th' unbending corn, and skims along the main.

In 18th-century theory the doctrine of imitation was joined to numerous strictures on "smoothness," or metrical regularity. Theorists advocated a rigid regularity; minor poets composed in a strictly regular syllable-stress verse devoid of expressive variations. This regularity itself expressed the rationalism of the period. The prevailing dogmas on regularity made it impossible for Samuel Johnson to hear the beauties of Milton's versification; he characterized the metrically subtle lines of "Lycidas" as "harsh" and without concern for "numbers." Certain crosscurrents of metrical opinion in the 18th century, however, moved toward new theoretical stances. Joshua Steele's *Prosodia Rationalis* (1779) is an early attempt to scan English verse by means of musical notation. (A later attempt was made by the American poet Sidney Lanier in his *Science of English Verse,* 1880.) Steele's method is highly personal, depending on an idiosyncratic assigning of such musical qualities as pitch and duration to syllabic values; but he recognized that a prosodic theory must take into account not merely metre but "all properties or accidents belonging to language." His work foreshadows the current concerns of the structural linguists who attempt an analysis of the entire range of acoustic elements contributing to prosodic effect. Steele is also the first "timer" among metrists; that is, he bases his scansions on musical pulse and claims that English verse moves in either common or triple time. Modern critics of musical scanners have pointed out that musical scansion constitutes a performance, not an analysis of the metre, that it allows arbitrary readings, and that it levels out distinctions between poets and schools of poetry.

The 19th century. With the Romantic movement and its revolutionary shift in literary sensibility, prosodic theory became deeply influenced by early 19th-century speculation on the nature of imagination, on poetry as expression—"the spontaneous overflow of powerful feelings," in Wordsworth's famous phrase—and on the concept of the poem as organic form. The discussion between Wordsworth and Coleridge on the nature and function of metre illuminates the crucial transition from Neoclassical to modern theories. Wordsworth (in his "Preface" to the *Lyrical Ballads,* 1800) followed 18th-century theory and saw metre as "superadded" to poetry; its function is more nearly ornamental, a grace of style and not an essential quality. Coleridge saw metre as being organic; it functions together with all of the other parts of a poem and is not merely an echo to the sense or an artifice of style. Coleridge also examined the psychologic effects of metre, the way it sets up patterns of expectation that are either fulfilled or disappointed:

The poem as organic form

> As far as metre acts in and for itself, it tends to increase the vivacity and susceptibility both of the general feelings and of the attention. This effect it produces by the continued excitement of surprize, and by the quick reciprocations of curiosity still gratified and still re-excited, which are too slight indeed to be at any one moment objects of distinct consciousness, yet become considerable in their aggregate influence. As a medicated atmosphere, or as wine during animated conversation; they act powerfully, though themselves unnoticed. Where, therefore, correspondent food and appropriate matter are not provided for the attention and feelings thus roused, there must needs be a disappointment felt; like that of leaping in the dark from the last step of a staircase, when we had prepared our muscles for a leap of three or four.
>
> *Biographia Literaria,* XVIII (1817)

Romantic literary theory, although vastly influential in poetic practice, had little to say about actual metrical structure. Coleridge described the subtle relationships between metre and meaning and the effects of metre on the reader's unconscious mind; he devoted little attention to metrical analysis. Two developments in 19th-century poetic techniques, however, had greater impact than any prosodic theory formulated during the period. Walt Whitman's nonmetrical prosody and Gerard Manley Hopkins' far-ranging metrical experiments mounted an assault on the traditional syllable-stress metric. Both Whitman and Hopkins were at first bitterly denounced, but, as is often the case, the heresies of a previous age become the orthodoxies of the next. Hopkins' "sprung rhythm"—a rhythm imitating natural speech, using mixed types of feet and counterpointed verse—emerged as viable techniques in the poetry of Dylan Thomas and W.H. Auden. It is virtually impossible to assess Whitman's influence on the various prosodies of modern poetry. Such American poets as Hart Crane, William Carlos Williams, and Theodore Roethke all have used Whitman's long line, extended rhythms, and "shaped" strophes.

The 20th century. Since 1900 the study of prosody has emerged as an important and respectable part of literary study. George Saintsbury published his great *History of English Prosody* during the years 1906–10. Sometime later, a number of linguists and aestheticians turned their attention to prosodic structure and the nature of poetic rhythm. Graphic prosody (the traditional syllable and foot scansion of syllable-stress metre) was placed on a securer theoretical footing. A number of prosodists, taking their lead from the work of Joshua Steele and Sidney Lanier, have recently attempted to use musical notation to scan English verse. For the convenience of synoptic discussion, modern prosodic theorists may be divided into four groups: the linguists who examine verse rhythm as a function of phonetic structures; the aestheticians who examine the psychologic effects, the formal properties, and the phenomenology of rhythm; the musical scanners, or "timers," who try to adapt the procedures of musical notation to

Linguists, aestheticians, timers, and traditionalists

metrical analysis; and the traditionalists who rely on the graphic description of syllable and stress to uncover metrical paradigms. It is necessary to point out that only the traditionalists concern themselves specifically with metrical form; aestheticians, linguists, and timers all examine prosody in its larger dimensions.

Modern structural linguistics has placed the study of language on a solid scientific basis. Linguists have measured the varied intensities of syllabic stress and pitch and the durations of junctures or the pauses between syllables. These techniques of objective measurement have been applied to prosodic study. The Danish philologist Otto Jespersen's early essay "Notes on Metre" (1900) made a number of significant discoveries. He established the principles of English metre on a demonstrably accurate structural basis; he recognized metre as a gestalt phenomenon (*i.e.,* with emphasis on the configurational whole); he saw metrics as descriptive science rather than proscriptive regulation. Jespersen's essay was written before the burgeoning interest in linguistics, but since World War II numerous attempts have been made to formulate a descriptive science of metrics.

It has been noted that Coleridge defined metrical form as a pattern of expectation, fulfillment, and surprise. Taking his cue from Coleridge, the British aesthetician I.A. Richards in *Principles of Literary Criticism* (1924) developed a closely reasoned theory of the mind's response to rhythm and metre. His theory is organic and contextual; the sound effects of prosody have little psychologic effect by themselves. It is prosody in conjunction with "its contemporaneous other effects"—chiefly meaning or propositional sense—that produces its characteristic impact on our neural structures. Richards insists that everything that happens in a poem depends on the organic environment; in his *Practical Criticism* (1929) he constructed a celebrated "metrical dummy" to "support [an] argument against anyone who affirms that the mere sound of verse has *independently* any considerable aesthetic virtue." For Richards the most important function of metre is to provide aesthetic framing and control; metre makes possible, by its stimulation and release of tensions, "the most difficult and delicate utterances."

Rhythmic structure as a symbolic form

Other critics, following the Neo-Kantian theories of the philosophers Ernst Cassirer and Susanne Langer, have suggested that rhythmic structure is a species of symbolic form. Harvey Gross in *Sound and Form in Modern Poetry* (1964) saw rhythmic structure as a symbolic form, signifying ways of experiencing organic processes and the phenomena of nature. The function of prosody, in his view, is to image life in a rich and complex way. Gross's theory is also expressive; prosody articulates the movement of feeling in a poem. The unproved assumption behind Gross's expressive and symbolic theory is that rhythm is in some way iconic to human feeling: that a particular rhythm or metre symbolizes, as a map locates the features of an actual terrain, a particular kind of feeling.

The most sophisticated argument for musical scansion is given by Northrop Frye in his influential *Anatomy of Criticism* (1957). He differentiates between verse that shows unmistakable musical quality and verse written according to the imitative doctrines current in the Renaissance and Neoclassic periods. All of the poetry written in the older strong-stress metric, or poetry showing its basic structure, is musical poetry, and its structure resembles the music contemporary with it.

The most convincing case for traditional "graphic prosody" has been made by the American critics W.K. Wimsatt and Monroe C. Beardsley. Their essay "The Concept of Meter" (1965) argues that both the linguists and musical scanners do not analyze the abstract metrical pattern of poems but only interpret an individual performance of the poem. Poetic metre is not generated by any combination of stresses and pauses capable of precise scientific measurement; rather, metre is generated by an abstract pattern of syllables standing in positions of relative stress to each other. In a line of iambic pentameter

Preserved in Milton's or in Shakespeare's name . . .

the "or" of the third foot is only slightly stronger than the preceding syllable "-ton's," but this very slight difference makes the line recognizable as iambic metre. Wimsatt and Beardsley underline the paradigmatic nature of metre; as an element in poetic structure, it is capable of exact abstraction.

Non-Western theories. The metres of the verse of ancient India were constructed on a quantitative basis. A system of long and short syllables, as in Greek, determined the variety of complicated metrical forms that are found in poetry of post-Vedic times—that is, after the 5th century BC.

Chinese prosody is based on the intricate tonal system of the language. In the T'ang dynasty (AD 618–907) the metrical system for classical verse was fixed. The various tones of the language were subsumed under two large groups, even tones and oblique tones. Patterned arrangements of tones and the use of pauses, or caesuras, along with rhyme determine the Chinese prosodic forms.

Japanese prosody

Japanese poetry is without rhyme or marked metrical structure; it is purely syllabic. The two main forms of syllabic verses are the tanka and the haiku. Tanka is written in a stanza of 31 syllables that are divided into alternating lines of five and seven syllables. Haiku is an extremely concentrated form of only 17 syllables. Longer poems of 40 to 50 lines are also written; however, alternate lines must contain either five or seven syllables. The haiku form has been adapted to English verse and has become in recent years a popular form. Other experimenters in English syllabic verse show the influence of Japanese prosody. Syllabic metre in English, however, is limited in its rhythmic effects; it is incapable of expressing the range of feeling that is available in the traditional stress and syllable-stress metres. (Ha.G.)

NARRATIVE FICTION

Epic

An ambiguous term, "epic" is used most often to designate a long narrative poem recounting heroic deeds, though it has also been loosely used to describe novels, such as Tolstoy's *War and Peace,* and motion pictures, such as Eisenstein's *Ivan the Terrible.* In literary usage, the term encompasses both oral and written compositions. The prime examples of the oral epic are Homer's *Iliad* and *Odyssey.* Outstanding examples of the written epic include Virgil's *Aeneid* and Lucan's *Pharsalia* in Latin; *Chanson de Roland* in medieval French; Ariosto's *Orlando Furioso* and Tasso's *Gerusalemme liberata* in Italian; *Poema* (or *Cantar*) *de mio Cid* in Spanish; and Milton's *Paradise Lost* and Spenser's *Faerie Queene* in English. There are also seriocomic epics, such as the *Morgante* of a 15th-century Italian poet, Pulci, and the pseudo-Homeric *Battle of the Frogs and Mice.* Another distinct group is made up of the so-called beast epics—narrative poems written in Latin in the Middle Ages and dealing with the struggle between a cunning fox and a cruel and stupid wolf. Underlying all of the written forms is some trace of an oral character, partly because of the monumental persuasiveness of Homer's example but more largely because the epic was, in fact, born of an oral tradition. It is on the oral tradition of the epic form that this article will focus.

GENERAL CHARACTERISTICS

The epic as a genre of heroic ages

An epic may deal with such various subjects as myths, heroic legends, histories, edifying religious tales, animal stories, or philosophical or moral theories. Epic poetry has been used by peoples all over the world and in different ages to transmit their traditions from one generation to another, without the aid of writing. These traditions frequently consist of legendary narratives about the glorious deeds of their national heroes. Thus scholars have often

identified "epic" with a certain kind of heroic oral poetry, which comes into existence in so-called heroic ages. Such ages have been experienced by many nations, usually at a stage of development in which they have had to struggle for a national identity. This effort, combined with such other conditions as an adequate material culture and a sufficiently productive economy, tend to produce a society dominated by a powerful and warlike nobility, constantly occupied with martial activities, whose individual members seek, above all, everlasting fame for themselves and for their lineages.

Uses of the epic. The main function of poetry in heroic-age society, appears to be to stir the spirit of the warriors to heroic actions by praising their exploits and those of their illustrious ancestors, by assuring a long and glorious recollection of their fame, and by supplying them with models of ideal heroic behaviour. One of the favorite pastimes of the nobility in heroic ages in different times and places has been to gather in banquet halls to hear heroic songs, in praise of famous deeds, sung by professional singers as well as by the warriors themselves. Heroic songs also were often sung before a battle, and such recitations had tremendous effect on the morale of the combatants. Among the Negro tribe of the Fulbe in The Sudan, for instance, whose epic poetry has been recorded, a nobleman customarily set out in quest of adventures accompanied by a singer (*mabo*), who also served as his shield bearer. The singer was thus the witness of the heroic deeds of his lord, which he celebrated in an epic poem called *baudi*.

The aristocratic warriors of the heroic ages were, thus, members of an illustrious family, a link in a long chain of glorious heroes. And the chain could snap if the warrior failed to preserve the honour of the family, whereas, by earning fame through his own heroism, he could give it new lustre. Epic traditions were to a large extent the traditions of the aristocratic families: the Old French word *geste*, used for a form of epic that flourished in the Middle Ages, means not only a story of famous deeds, but also a genealogy.

The passing of a heroic age does not necessarily mean the end of its heroic oral poetry. An oral-epic tradition usually continues for as long as the nation remains largely illiterate. Usually it is after the heroic age has passed that the narratives about its legendary heroes are fully elaborated. Even when the nobility that originally created the heroic epic perishes or loses interest, the old songs can persist as entertainments among the people. Court singers, then, are replaced by popular singers, who recite at public gatherings. This popular tradition, however, must be distinguished from a tradition that still forms an integral part of the culture of a nobility. For when a heroic epic loses its contact with the banquet halls of the princes and noblemen, it cannot preserve for long its power of renewal. Soon it enters what has been called the reproductive stage in the life cycle of an oral tradition, in which the bards become noncreative reproducers of songs learned from older singers. Popular oral singers, like the *guslari* of Yugoslavia, no doubt vary their songs to a certain extent each time they recite them, but they do so mainly by transposing language and minor episodes from one acquired song to another. Such variations must not be confounded with the real enrichment of the tradition by succeeding generations of genuine oral poets of the creative stage. The spread of literacy, which has a disastrous effect on the oral singer, brings about a quick corruption of the tradition. At this degenerate stage, the oral epic soon dies out if it is not written down or recorded.

The popular tradition of the epic

The ancient Greek epic exemplifies the cycle of an oral tradition. Originating in the late Mycenaean period, the Greek epic outlasted the downfall of the typically heroic age culture (*c.* 1100 BC), and maintained itself through the "Dark Age" to reach a climax in the Homeric poems by the close of the Geometric Period (900–750 BC). After Homer, the activity of the *aoidoi*, who sang their own epic songs at the courts of the nobility, slowly declined. During the first half of the 7th century, the *aoidoi* produced such new poems as those of Hesiod and some of the earlier poems of what was to become known as the Epic Cycle. Between 625 and 575 BC, the *aoidoi* gave way to oral reciters of a new type called rhapsodes or "stitchers of songs," who declaimed for large audiences the already famous works of Homer, while holding in their hand, a staff (*rhabdos*), which they used to give emphasis to their words. It seems probable that these rhapsodes, who played a crucial role in the transmission of the Homeric epic, were using some sort of written aids to memory before Homeric recitations were adopted in 6th-century Athens as part of the Panathenaic festivals held each year in honour of the goddess Athena.

Verbal formulas. To compose and to memorize long narrative poems like the *Iliad* and the *Odyssey*, oral poets used a highly elaborate technical language with a large store of traditional verbal formulas, which could describe recurring ideas and situations in ways that suited the requirements of metre. So long as an oral epic tradition remains in its creative period, its language will be continually refined by each generation of poets in opposite directions, refinements that are called scope and economy. Scope is the addition of new phrases to express a larger number of recurrent concepts in varying metrical values fitting the possible positions in a verse. Economy is the elimination of redundancies that arise as gifted poets invent new set phrases that duplicate, both in a general sense and in metrical value, the formulas that already exist in the traditional stock.

Nowhere has this refinement proceeded any nearer to perfection than in the language of the Homeric epic. As has been shown by statistical analysis, it exhibits a remarkable efficiency, both in the rareness of unnecessarily duplicative variants and in the coverage of each common concept by the metrical alternatives useful in the composition of the six-foot metric line the Greeks used for epic poetry.

Thus, for example, if the idea of a ship has to be expressed at the end of a line of verse, the ship may be described as "well-trimmed" (*nēos eisēs*), "curved" (*nēos amphielissa*), or "dark-prowed" (*nēos kyanoprōiros*), depending entirely on the number of feet that remain to be filled by the phrase in the hexameter; if the phrase has to cover the two final feet of the verse and the words have to be put in the dative case, the formula "of a well-trimmed ship" will be replaced by "to a black ship" (*nēi melainē*). The sole occurrence of "Zeus who gathers lightning" (*steropēgereta Zeus*), which is an exact metrical equivalent of the more common "Zeus who delights in thunder" (*Zeus terpikeraunos*), constitutes one of the very few actual duplications of such formulas found in Homer.

Finally, some of the typical scenes in the heroic life, such as the preparation of a meal or sacrifice, or the launching or beaching of a ship, contain set descriptions comprising several lines that are used by rote each time the events are narrated.

This highly formalized language was elaborated by generations of oral poets to minimize the conscious effort needed to compose new poems and memorize existing ones. Because of it, an exceptionally gifted *aoidos*, working just prior to the corruption of the genre, could orally create long and finely structured poems like the *Iliad* and the *Odyssey* and those poems could then be transmitted accurately by the following generations of rhapsodes until complete, written texts were produced.

BASES

Oral heroic poetry, at its origin, usually deals with outstanding deeds of kings and warriors who lived in the heroic age of the nation. Since the primary function of this poetry is to educate rather than to record, however, the personages are necessarily transformed into ideal heroes and their acts into ideal heroic deeds that conform to mythological or ideological patterns. Some of these patterns are archetypes found all over the world, while others are peculiar to a specific nation or culture. Thus, in many epic traditions, heroes are born as a result of an illegitimate union of a maiden mother with a divine or supernatural being; they are exposed at birth, fed by an animal, and brought up by humble foster parents in a rustic milieu; they grow up with marvellous speed, fight a dragon—in their first combat—to rescue a maiden whom

Heroic patterns and deeds

they marry, and die young in circumstances as fabulous as those that surrounded their birth.

In the traditions of Indo-European peoples a hero is often a twin, who acquires soon after his supernatural birth an invulnerability that has one defect, generally of his heel or of some other part of his foot, which ultimately causes his death. He is educated by a blacksmith, disguises himself as a woman at some time in his youth, and conquers a three-headed dragon, or some other kind of triple opponent, in his first battle. He then begets, by a foreign or supernatural woman, a child who, reared by his mother in her country, becomes a warrior as brave as his father. When this child meets his unknown father, the latter fails to recognize him, so that the father kills his own child after a long and fierce single combat. The hero, himself, usually dies after committing the third of three sins.

In Japan, to take another example, renowned members of the warrior aristocracy of the past, who have acquired the status of popular heroes, are in many cases supplied in their legend with four exceptionally brave and faithful retainers called their *shi-tennō*, the guardians of the four cardinal points; these form the closest entourage of their lord—who is usually depicted as excelling in command but not in physical strength—and defend him from dangers. The retainers reflect a mythological model, taken from Buddhism, of four *deva* kings, who guard the teaching of the Buddha against the attack of the devils.

A striking pattern for a number of epic traditions has been found in a so-called "tripartite ideology" or "trifunctional system" of the Indo-Europeans. The concept was based on the discovery of the remarkable philosophy of a prehistoric nation that survived as a system of thought in the historic Indo-European civilizations and even in the subconsciousness of the modern speakers of Indo-European tongues.

This philosophy sees in the universe three basic principles that are realized by three categories of people: priests, warriors, and producers of riches. In conformity with this philosophy, most Indo-European epics have as their central themes interaction among these three principles or functions which are: (1) religion and kingship; (2) physical strength; (3) fecundity, health, riches, beauty, and so forth. In the long Indian epic the *Mahābhārata,* for example, the central figures, the Pāṇḍava brothers, together with their father Paṇḍu, their two uncles Dhṛtarāṣṭra and Vidura, and their common wife, Draupadī, correspond to traditional deities presiding over the three functions of the Indo-European ideology.

Central themes of Indo-European epics

During the first part of their earthly career, the Pāṇḍava suffer constantly from the persistent enmity and jealousy of their cousins, Duryodhana and his 99 brothers, who, in reality, are incarnations of the demons Kālī and the Paulastya. The demons at first succeeded in snatching the Kingdom from the Pāṇḍava and in exiling them. The conflict ends in a devastating war, in which all the renowned heroes of the time take part. The Pāṇḍava survive the massacre, and establish on earth a peaceful and prosperous reign, in which Dhṛtarāṣṭra and Vidura also participate.

This whole story, it has been shown, is a transposition to the heroic level of an Indo-European myth about the incessant struggle between the gods and the demons since the beginning of the world. Eventually, it results in a bloody eschatological battle, in which the gods and the devils exterminate each other. The destruction of the former world order, however, prepares for a new and better world, exempt from evil influences, over which reign a few divine survivors of the catastrophe.

EARLY PATTERNS OF DEVELOPMENT

In the ancient Middle East. The earliest-known epic poetry is that of the Sumerians. Its origin has been traced to a preliterate Heroic Age, not later than 3000 BC, when the Sumerians had to fight, under the direction of a warlike aristocracy, for possession of this fertile Mesopotamian land. Among the extant literature of this highly gifted people are fragments of narrative poems recounting the heroic deeds of their early kings: Enmerkar, Lugalbanda, and Gilgamesh. By far the most important in the development of Mesopotamian literature are the five poems of the Gilgamesh cycle. This epic tells the odyssey of a king, Gilgamesh, part human and part divine, who seeks immortality. A god who dislikes his rule, fashions a wild man, Enkidu, to challenge him. Enkidu first lives among wild animals, then goes to the capital and engages in a trial of strength with Gilgamesh, who emerges victorious. The two, now friends, set out on various adventures, in one of which they kill a wild bull that the goddess of love had sent to destroy Gilgamesh because he spurned her marriage proposal. Enkidu dreams the gods have decided he must die for the death of the bull, and, upon awakening, he does fall ill and die. Gilgamesh searches for a survivor of the Babylonian flood to learn how to escape death. The survivor shows him where to find a plant that renews youth, but after Gilgamesh gets the plant it is snatched away by a serpent. Gilgamesh returns, saddened, to his capital.

The legend of Gilgamesh was taken over by the Babylonians, who developed it into a long and beautiful poem, one of the masterpieces of mankind.

The legend of Gilgamesh

Another Babylonian epic, composed around 2000 BC, is called in Akkadian *Enuma elish,* after its opening words, meaning "When on high." Its subject is not heroic but mythological. It recounts events from the beginning of the world to the establishment of the power of Marduk, the great god of Babylon. The outline of a Babylonian poem narrating the adventure of a hero named Adapa ("Man") can be reconstructed from four fragmentary accounts. It shares with the *Epic of Gilgamesh* the theme of man's loss of an opportunity for immortality.

Among clay tablets of the 14th century BC, covered with inscriptions in an old Phoenician cuneiform alphabet, from Ras Shamra (the site of ancient Ugarit), in northern Syria, there are important fragments of three narrative poems. One of these is mythological and recounts the career of the god Baal, which seems to coincide with the yearly cycle of vegetation on earth. As was usual with the death of gods in the ancient Mediterranean world, Baal's end brings about a drought that ceases only with his resurrection. Another fragment, about a hero named Aqhat, is perhaps a transposition of this myth of Baal to the human level. Just as the death of Baal is avenged on his slayer by Baal's sister Anath, so is the murder of Aqhat, which also causes a drought, revenged by his sister Paghat. Since the end of the poem is missing, however, it is not known whether Paghat, like Anath, succeeded in bringing her brother back to life.

The third fragment, the Ugaritic epic of Keret, has been interpreted as a Phoenician version of the Indo-European theme of the siege of an enemy city for the recovery of an abducted woman. This theme is also the subject of the Greek legend of the Trojan War and of the Indian epic *Rāmāyaṇa.* The fragmentary text does not reveal, however, whether the expedition of Keret, like that of the Achaean army against Troy, was meant to regain the hero's wife or to acquire for him a new bride.

The Greek epic. In its originative stage, especially, the Greek epic may have been strongly influenced by these oriental traditions. The Greek world in the late Bronze Age was related to the Middle East by so many close ties that it formed an integral part of the Levant. At Ugarit a large quarter of the city was occupied by Greek merchants, whose presence is also attested, among other places, at the gate of Mesopotamia, at Alalakh, in what is now Turkey. Thus, it is no surprise that, for example, the Greek myth about the succession of the divine kingship told in the *Theogony* of Hesiod and elsewhere is paralleled in a Hittite version of a Hurrian myth. In it, Anu, Kumarbi, and the storm god respectively, parallel Uranus, Cronos, and Zeus in the *Theogony.* The Hittites had continuous diplomatic relations with the Achaeans of Greece, whose princes went to the royal court at Hattusa to perfect their skill with the chariot. The Greeks, therefore, had ample opportunity to become familiar with Hittite myths.

Eastern influences on the Greek epic

The *Epic of Gilgamesh* was then well-known in the Levant, as is indicated by discoveries of copies of it throughout this wide area. Many parallels with the *Epic of Gilgamesh* have been pointed out in the *Odyssey;* the encounters of Odysseus with Circe and Calypso on their

mythical isles, for instance, closely resemble the visit by Gilgamesh to a divine woman named Siduri, who keeps an inn in a marvellous garden of the sun god near the shores of ocean. Like the two Greek goddesses, Siduri tries to dissuade Gilgamesh from the pursuit of his journey by representing the pleasures of life, but the firm resolution of the hero obliges her finally to help him cross the waters of death. In the *Iliad,* Patroclus, who dies as a substitute for his king and dearest friend, Achilles, and then gives Achilles a description of the miserable condition of man after his death, bears striking similarities to the friend of Gilgamesh, Enkidu.

If these are indeed borrowings, it is all the more remarkable that they are used in Homer to express a view of life and a heroic temper radically different from those of the Sumerian epic of Mesopotamia. Gilgamesh persists in his quest of immortality even when Siduri shows him the vanity of such an ambition, but Odysseus shuns a goddess's offer of everlasting life, preferring to bear his human condition to the end. The loss of a beloved friend does not make Achilles seek desperately to escape from death; instead he rushes into combat to revenge Patroclus, although he knows that he is condemning himself to an early death, and that the existence of a king in Hades will be incomparably less enviable than that of a slave on earth. The Mesopotamian mind never tires of expressing man's deep regret at not being immortal through stories about ancient heroes who, despite their superhuman strength and wisdom, and their intimacy with gods, failed to escape from death. A decisively different idea, however, is fundamental to the Greek heroic view of life. It has been demonstrated that the Greek view is derived from an Indo-European notion of justice—that each being has a fate (*moira*) assigned to him and marked clearly by boundaries that should never be crossed. Man's energy and courage should, accordingly, be spent not in exceeding the proper limits of his human condition but in bearing it with style, pride, and dignity, gaining as much fame as he can within the boundaries of his *moira.* If he is induced by Folly (Ate, personified as a goddess of mischief) to commit an excess (*hybris*) with regard to his *moira,* he will be punished without fail by the divine vengeance personified as Nemesis.

The Greek heroic view of life

At the beginning of the *Iliad,* a plague decimates the Achaean army because its commander in chief, Agamemnon, refuses to return a captive, Chryseis, to her father, a priest of Apollo who offers a generous ransom. By unjustly insulting Achilles, Agamemnon commits another excess that causes the defeat of his army. Achilles, in the meantime, lets Ate take possession of his mind and refuses, to the point of excess, to resume his fight. He thus brings about a great misfortune, the loss of his dearest companion, Patroclus. Patroclus, however, also contributes to his own death by his *hybris* in pushing his triumph too far, ignoring Achilles' order to come back as soon as he has repulsed the enemy far from the Greek ships. The death of Hector also results from his *hybris* in rejecting the counsel of Polydamas and maintaining his army on the plain after the return of Achilles to combat. After so many disasters caused by the mischievous action of Ate among men, the last book of the *Iliad* presents a noble picture of Priam and Achilles, who submit piously to the orders of Zeus, enduring with admirable courage and moderation their respective fates.

On the other hand, at the beginning of the *Odyssey,* Zeus evokes the ruin that Aegisthus will have to suffer for having acted "beyond his due share" by marrying Clytemnestra and murdering Agamemnon. This sets an antithesis to the story of the wise Odysseus, who, to accomplish his destiny as a mortal hero, never changes his purpose trying always to make the best of his countless misfortunes. He earns by this the favour of Athena and succeeds eventually in regaining Ithaca and punishing the wooers of Penelope for their *hybris* during his long absence. Present scholarship inclines to the view that such admirably well-structured poems as the *Iliad* and the *Odyssey* could have been created only by a single highly gifted poet whose name was Homer. This position contrasts with the extreme skepticism that marked all phases of Homeric criticism during the previous century. Yet the personality of Homer remains unknown and nothing certain is known about his life.

In comparison, information derived from his own works is fairly plentiful about the other great epic poet of Greece, Hesiod. He produced them presumably around 700 BC, while tilling a farm in Askra, a small village of Boeotia. The social and geographical background of his poems, called didactic because of their occasionally moral and instructive tone, differs from the aristocratic society of Ionian Asia Minor that Homer addressed. Despite their different style, subjects, and view of life, however, Hesiod's *Theogony* and the *Works and Days* illustrate the same basic conception of justice as the Homeric epic. The *Theogony* describes a long sequence of primordial events that resulted in the present world order, in which man's inescapable lot is assigned to him by Zeus. The *Works and Days* explains, through a series of three myths, why the lot of man is to work hard to produce riches. Man has to shut his ears to the goddess who causes wars and lawsuits, listening only to the goddess who urges him to toil more laboriously than his neighbour to become richer. Pain and suffering have become unavoidable since Pandora opened the fatal jar containing all the ills of mankind at Prometheus' house in conformity with the will of Zeus. Moreover, the age of the race of iron has arrived when the fate of human beings is not to pass their lives in perpetual banquets or warfare, as did the preceding races, but to suffer constantly the fatigue and misery of labour. As long as the goddesses Aidos (a personification of the sense of shame) and Nemesis (a personification of divine retribution) stay with mankind, however, helping people observe their *moira* without committing excesses, man can still gain riches, merits, and glory by the sweat of his brow. Only if he knows how to avoid all faults in doing his daily work will he not offend Justice (Dikē), the sensitive virgin daughter of Zeus. This is why it is so vitally important for a farmer to know all the rules listed in the rest of the poem about seemingly trivial details of his work.

Hesiod's contribution

LATER VARIATIONS

The Latin epic. Latin epic poetry was initiated in the 3rd century BC by Livius Andronicus, who translated the *Odyssey* into the traditional metre of Saturnian verse. It was not until the 1st century BC, however, that Rome possessed a truly national epic in the unfinished *Aeneid* of Virgil (70–19 BC), who used Homer as his model. The story of Aeneas' journey, recounted in the first six books, is patterned after the *Odyssey,* with many imitative passages and even direct translations, while the description of the war in the last six books abounds with incidents modelled after those from the *Iliad.* More basically, however, Virgil made use of another model, Rome's own national legend about the war fought under Romulus against the Sabines. This legend preserves, in a historical disguise, an original Indo-European myth about a primitive conflict between the gods of sovereignty and war and the gods of fecundity, ending with the unification of the two divine races. In the development of this theme by Virgil, Aeneas and the Etruscans can be seen as representing the gods of sovereignty and war, and the Latins representing the gods of fecundity. Aeneas, who has brought the Trojan gods to Rome, is forced to fight with the help of the Etruscans against the Latins. It is the destiny of Aeneas to rule, and it is the fate of the Latins to share their land and women with the invaders and to accept Aeneas as their king. This resembles the unification of the warring races that climaxes the Indo-European myth.

The power exercised by the Indo-European ideological pattern on the Roman mind even under the empire is seen in the *Pharsalia* of Lucan (AD 39–65). In this historical epic, Cato, Caesar, and Pompey are depicted respectively as moral, warlike, and popular in a way that gives the story a clear trifunctional structure.

The persistence of the Indo-European pattern

Germanic epics. A typical Heroic Age occurred during the wanderings of the Germanic tribes from the 3rd to the 6th centuries AD. Out of this, too, came a rich oral tradition, from which developed in the Middle Ages many epic poems. One of the greatest of these is the Old English

Beowulf, written down in the 8th century. Archetypal Indo-European themes also reappear in these epics. For example, the theme of the fatal fight between father and son is recounted in the German *Hildebrandslied,* of which a 67-line fragment is extant. Again, an heroic version of the Indo-European myth about the rescue of the Sun Maiden from her captivity by the Divine Twins, which also provided the basic plot of the Greek Trojan cycle and the Indian *Rāmāyaṇa,* is found in the German *Gudrun* (*c.* 1230).

Chansons de geste. The French chansons de geste are epic poems whose action takes place during the reign of Charlemagne and his immediate successors. The *Chanson de Roland,* probably written down about the end of the 11th century, is by far the most refined of the group. The story of the poem had developed from a historical event, the annihilation of the rear guard of Charlemagne's army at Roncesvalles in the Pyrenees in 778 by Basque mountaineers. The Basques, however, are transformed in the epic to the Saracens, who to a later generation typified France's enemies in Spain. The other chansons de geste, none of which is comparable to *Roland* as a literary work, have been classified into three main cycles. The cycle of Guillaume d'Orange forms a biography of William (probably a historical, count of William of Toulouse, who had, like the hero of the epic, a wife called Guibourg and a nephew, Vivien, and who became a monk in 806). Guibourg, the most faithful of wives, and the noble Vivien take prominent roles in the epic. The so-called Cycle of the Revolted Knights groups those poems that tell of revolts of feudal subjects against the emperor (Charlemagne or, more usually, his son, Louis). The Cycle of the King consists of the songs in which Charlemagne himself is a principal figure.

Arthurian Romance. The Arthurian Romance seems to have developed first in the British Isles, before being taken to the Continent by Bretons, who migrated to Brittany in the 6th and 7th centuries. The core of the legend about Arthur and his knights derives from lost Celtic mythology. Many of the incidents in the former parallel the deeds of such legendary Irish characters as Cú Chulainn, an Ulster warrior said to have been fathered by the god Lug, and Finn, hero of the Fenian cycle about a band of warriors defending Ireland, both of whom are gods transformed into human heroes. The earliest extant works on Arthurian themes are four poems of Chrétien de Troyes, written in French between 1155 and 1185: *Erec, Yvain, Le Chevalier de la Charette* (left unfinished by Chrétien and completed by Godefroy de Lagny), and an unfinished *Perceval.* In German, after 1188, Hartmann von Aue (who also wrote two legendary poems not belonging to the Arthurian cycle, *Gregorius* and *Poor Henry*) modelled his *Erec* and *Iwein* on those of Chrétien. The story of Perceval was given a full account by Wolfram von Eschenbach (*c.* 1170–1220) in his *Parzival* and in the unfinished *Titurel.* Another incomplete work of Wolfram, *Willehalm,* deals with the legend of William of Orange. *Tristan* of Gottfried von Strassburg is based directly on the older French version of Thomas of Britain (*c.* 1170–80). The romance proper, however, although it has similarities to the epic, differs in its lack of high purpose: fictions are told for their entertainment value rather than as models for national heroism. Developed in France in the Middle Ages, the romance is usually an adventure story with a strong love interest, intimately associated with the "courtly love" tradition of that time. (For further treatment, see below *Romance.*)

Distinctions between the epic and the romance

The epic in Japan. In Japan, there were in ancient times families of reciters (*katari-be*) whose duty was to hand down myths and legends by word of mouth and to narrate them during official ceremonies and banquets. After the introduction of Chinese letters, however, from the 4th century AD onward, these traditional tales were put in writing and the *katari-be* professional gradually died out. By the end of the 7th century, each clan of the ruling aristocracy seems to have possessed a written document that recounted the mythology and legendary history of Japan in a form biassed in favour of the clan concerned. These family documents were collected at the command of the emperor Temmu (672–686) and were used as basic materials for the compilation of the first national chronicles of Japan, the *Koji-ki* (712) and the *Nihon shoki* (720). The myths and legends that are contained in the earlier parts of these two books derive, therefore, from the oral tradition of *katari-be.* Although no document preserves those narrations in their primitive form, it is generally assumed that they were originally in the form of poems. Many scholars believe that they were genuine epic poems, which were produced during a period of incessant warfare around the 4th century. At that time mounted aristocratic warriors of the future imperial family struggled to extend its power over the larger part of Japan. Exploits of warriors, such as the emperor Jimmu or Prince Yamato-Takeru, in the earliest extant texts—the *Koji-ki* and *Nihon shoki* of the 8th century—probably derive from a heroic epic about the wars of conquest of the first emperors, whose legendary feats were transformed into those of a few idealized heroic figures.

The middle of the Heian period (794–1185) saw the emergence of a new class of warrior known as samurai. They attached a greater importance to fame than to life. The battles they fought became the subject of epic narratives that were recited by itinerant blind priests to the accompaniment of a lute-like instrument called a *biwa.*

Tales of the samurai

In the early part of the 13th century, tales about the wars of the preceding century, fought between the two strongest families of samurai, the Genji, or Minamoto, and the Heike, or Taira, were compiled in three significant war chronicles. The *Hōgen monogatari* and the *Heiji monogatari* deal with two small wars, the Hōgen (1156) and Heiji (1159), in which the Genji and Heike warriors fought for opposing court factions. The structure of the two works is roughly the same. Each celebrates the extraordinary prowess of a young Genji warrior, Minamoto Tametomo in the *Hōgen monogatari* and Minamoto Yoshihira in the *Heiji monogatari;* each hero fights to the finish in exemplary manner not so much to win, for from the beginning each foresees the defeat of his own side, as for the sake of fame; and the consummate courage of the two heroes forms a striking contrast to the cowardice of court aristocrats. The bitterly fought Gempei War (1180–85), in which survivors of the Genji family challenged and defeated the Heike, is recounted in detail in the *Heike monogatari,* the greatest epic of Japanese literature. The sudden decline and ultimate extinction of the proud Heike, whose members had held the highest offices of the imperial court, illustrates the Buddhist philosophy of the transitory nature of all things; it invites the readers to seek deliverance from the world of sufferings through a faith that will take them to a land of eternal felicity at the moment of their death. The work is filled with tales of heroic actions of brave warriors. The most conspicuous is Minamoto Yoshitsune, one of the chief commanders of the Genji army: the legend of this man of military genius continued to develop in later literature, so that he has become the most popular hero of Japanese legend.

The later written epic. The vitality of the written epic is manifested by such masterworks as the Italian *Divine Comedy* of Dante (1265–1321) and the great Portuguese patriotic poem *Os Lusíadas* of Luiz de Camões (1524–80), which celebrates the voyage of Vasco da Gama to India. In more recent times, novels and long narrative poems written by such major authors as Scott, Byron, Tennyson, William Morris, and Melville were patterned, to some extent, on the epic. Their fidelity to the genre, however, is found primarily in their large scope and their roots in a national soil; their distance from the traditional oral epic tends to be considerable.

Among the epics written in modern times, the Finnish *Kalevala* (first ed. 1835; enlarged ed. 1849) occupies a very special position. This is because its author, the Finnish poet-scholar Elias Lönnrot (1802–84), who composed this masterpiece by combining short popular songs (*runot*) collected by himself among the Finns, had absorbed his material so well and identified himself so completely with the *runo* singers. He thus came close to showing what the oral epic, which he could study only at its degenerative stage, might have been at its creative stage, on the lips of an exceptionally gifted singer. (A.Y.)

Fable, parable, and allegory

Fables, parables, and allegories are forms of imaginative literature or spoken utterance constructed in such a way that their readers or listeners are encouraged to look for meanings hidden beneath the literal surface of the fiction. A story is told or perhaps enacted whose details—when interpreted—are found to correspond to the details of some other system of relations (its hidden, allegorical sense). The poet, for example, may describe the ascent of a hill in such a way that each physical step corresponds to a new stage in the soul's progress toward a higher level of existence.

Many forms of literature elicit this kind of searching interpretation, and the generic term for the cluster is allegory; under it may be grouped fables, parables, and other symbolic shapings. Allegory may involve either a creative or an interpretive process: either the act of building up the allegorical structure and giving "body" to the surface narrative or the act of breaking down this structure to see what themes or ideas run parallel to it.

NATURE AND OBJECTIVES

Allegory and myth. The fate of allegory, in all its many variations, is tied to the development of myth and mythology. Every culture embodies its basic assumptions in stories whose mythic structures reflect the society's prevailing attitudes toward life. If the attitudes are disengaged from the structure, then the allegorical meaning implicit in the structure is revealed. The systematic discipline of interpreting the real meaning of a text (called the hermeneutic process) plays a major role in the teaching and defense of sacred wisdom, since religions have traditionally preserved and handed down the old beliefs by telling exemplary stories; these sometimes appear to conflict with a system of morality that has in the meantime developed, and so their "correct" meaning can only be something other than the literal narration of events. Every culture puts pressure on its authors to assert its central beliefs, which are often reflected in literature without the author's necessarily being aware that he is an allegorist. Equally, determined critics may sometimes find allegorical meaning in texts with less than total justification—instances might include the Hebraic–Christian mystical interpretation of the Old Testament's Song of Solomon, an erotic marriage poem, or the frequent allegorizing of classical and modern literature in the light of Freud's psychoanalytic discoveries. Some awareness of the author's intention seems necessary in order to curb unduly fanciful commentary.

The allegorical mode. The range of allegorical literature is so wide that to consider allegory as a fixed literary genre is less useful than to regard it as a dimension, or mode, of controlled indirectness and double meaning (which, in fact, all literature possesses to some degree). Critics usually reserve the term allegory itself for works of considerable length, complexity, or unique shape. Thus, the following varied works might be called allegories: the biblical parable of the sower; *Everyman,* the medieval morality play; *The Pilgrim's Progress,* by John Bunyan; Jonathan Swift's *Gulliver's Travels; The Scarlet Letter,* by Nathaniel Hawthorne; William Wordsworth's "Ode: Intimations of Immortality"; Nikolay Gogol's *Dead Souls; The Picture of Dorian Gray,* by Oscar Wilde; and the plays *Six Characters in Search of an Author,* by Luigi Pirandello; *Waiting for Godot,* by Samuel Beckett; and *Who's Afraid of Virginia Woolf?,* by Edward Albee. No one genre can take in such modal range.

Fable. Fable and parable are short, simple forms of naïve allegory. The fable is usually a tale about animals who are personified and behave as though they were humans. The device of personification is also extended to trees, winds, streams, stones, and other natural objects. The earliest of these tales also included humans and gods as characters, but fable tends to concentrate on animating the inanimate. A feature that isolates fable from the ordinary folktale, which it resembles, is that a moral—a rule of behaviour—is woven into the story.

Personification of animals

Parable. Like fable, the parable also tells a simple story. But, whereas fables tend to personify animal characters—often giving the same impression as does an animated cartoon—the typical parable uses human agents. Parables generally show less interest in the storytelling and more in the analogy they draw between a particular instance of human behaviour (the true neighbourly kindness shown by the good Samaritan in the Bible story, for example) and human behaviour at large. Parable and fable have their roots in preliterate oral cultures, and both are means of handing down traditional folk wisdom. Their styles differ, however. Fables tend toward detailed, sharply observed social realism (which eventually leads to satire), while the simpler narrative surface of parables gives them a mysterious tone and makes them especially useful for teaching spiritual values.

Derivation of the terms. The original meanings of these critical terms themselves suggest the direction of their development. Fable (from the Latin *fabula,* "a telling") puts the emphasis on narrative (and in the medieval and Renaissance periods was often used when speaking of "the plot" of a narrative). Parable (from Greek *parabolē,* a "setting beside") suggests a juxtaposition that compares and contrasts this story with that idea. Allegory (from Greek *allos* and *agoreuein,* an "other-speaking") suggests a more expanded use of deceptive and oblique language. (In early Greek, though, the term allegory itself was not used. Instead, the idea of a hidden, underlying meaning is indicated by the word *hyponoia*—literally, "underthought"—and this term is used of the allegorical interpretation of the Greek poet Homer.)

Diverse objectives. *Fable.* Fables teach a general principle of conduct by presenting a specific example of behaviour. Thus, to define the moral that "People who rush into things without using judgment run into strange and unexpected dangers," Aesop—the traditional "father" of the fable form—told the following story:

The moral objective of fables

> There was a dog who was fond of eating eggs. Mistaking a shell-fish for an egg one day, he opened his mouth wide and swallowed it down in one gulp. The weight of it in his stomach caused him intense pain. "Serve me right," he said, "for thinking that anything round must be an egg."

By a slight change of emphasis, the fabulist could have been able to draw a moral about the dangerous effects of gluttony.

Because the moral is embodied in the plot of the fable, an explicit statement of the moral need not be given, though it usually is. Many of these moral tag lines have taken on the status of proverb because they so clearly express commonly held social attitudes.

The Aesopian fables emphasize the social interactions of human beings, and the morals they draw tend to embody advice on the best way to deal with the competitive realities of life. With some irony, fables view the world in terms of its power structures. One of the shortest Aesopian fables says: "A vixen sneered at a lioness because she never bore more than one cub. 'Only one,' she replied, 'but a lion.' " Foxes and wolves, which the poet Samuel Taylor Coleridge called "Everyman's metaphor" for cunning and cruelty, appear often as characters in fables chiefly because, in the human world, such predatory cunning and cruelty are able to get around restraints of justice and authority. The mere fact that fables unmask the "beast in me," as James Thurber, the 20th-century American humorist and fabulist, put it, suggests their satirical force. Subversive topical satire in tsarist and Soviet Russia is often called "Aesopism"; all comic strips that project a message (such as the Charles Schulz creation "Peanuts" and Walt Kelly's "Pogo") have affinities with Aesop's method.

Parable. Parables do not analyze social systems so much as they remind the listener of his beliefs. The moral and spiritual stress of the form falls upon memory rather than on the critical faculty. The audience hearing the parable is assumed to share a communal truth but perhaps to have set it aside or forgotten it. The rhetorical appeal of a parable is directed primarily toward an elite, in that a final core of its truth is known only to an inner circle, however simple its narrative may appear on the surface (a number of the parables that Christ used for teaching, for example, conveyed figuratively the meaning of the elusive concept Kingdom of Heaven).

Allegory. Allegory, as the basic process of arousing in the reader or listener a response to levels of meaning, provides writers with the structure of fables, parables, and other related forms. By awakening the impulse to question appearances and by bringing order to mythological interpretation, allegory imparts cultural values. A measure of allegory is present in literature whenever it emphasizes thematic content, ideas rather than events. Generally, the allegorical mode flourishes under authoritarian conditions. Thus it found sustenance during the age of medieval Christendom, when Christian dogma sought universal sway over the mind of Western man. As such, allegory was a means of freedom under conditions of strong restraint. In general, realism, mimetic playfulness, and the resistance to authority tend to counteract the allegorical process, by loosening its stratified forms. This unbinding of symbolic hierarchies has forced allegory to seek new structures in the modern period. Nevertheless, through allegorical understanding, the great myths continue to be reread and reinterpreted, as the human significance of the new interpretations is passed down from one generation to the next. The abiding impression left by the allegorical mode is one of indirect, ambiguous, even enigmatic symbolism, which inevitably calls for interpretation.

Diversity of forms. Since an allegorical purpose can inform works of literature in a wide range of genres, it is not surprising to find that the largest allegories are epic in scope. A quest forms the narrative thread of both the Greek epic *Odyssey* and the Latin, *Aeneid,* and it is an allegory of the quest for heroic perfection; thus, allegory is aligned with the epic form. Romances, both prose and verse, are inevitably allegorical, although their forms vary in detail with the prevailing cultural ideals of the age. By comparison, the forms of fable and parable are relatively stable—yet even they may play down the moral idea or the mysterious element and emphasize instead the narrative interest, which then results in an elaboration of the form. (Such an elaboration may be seen in a given tale, as told by successive fabulists, such as a fable of the town mouse and the country mouse; with each retelling, the story is absorbed into a new matrix of interpretation.)

Shifts from naïve to sophisticated intent are accompanied by shifts in form. The early authors of fable, following Aesop, wrote in verse; but in the 10th century there appeared collected fables, entitled *Romulus,* written in prose (and books such as this brought down into the medieval and modern era a rich tradition of prose fables). This collection in turn was converted back into elegiac verse. Later masters of fable wrote in verse, but modern favourites—such as Joel Chandler Harris, author of "Uncle Remus" stories, Beatrix Potter, creator of Peter Rabbit, or James Thurber in *Fables for Our Time* —employ their own distinctive prose. Again, while for parables prose narrative may be the norm, they have also been told in verse (as in the emblematic poetry of the 17th-century English Metaphysical poets such as George Herbert, Francis Quarles, and Henry Vaughan).

Forms that mix prose and poetry

Loosening the allegorical forms further, some authors have combined prose with verse. Boethius' *Consolation of Philosophy* (*c.* AD 524) and Dante's *The New Life* (*c.* 1293) interrupt the prose discourse with short poems. Verse and prose then interact to give a new thematic perspective. A related mixing of elements appears in Menippean satire (those writings deriving from the 3rd-century-BC Cynic philosopher Menippus of Gadara), as exemplified in Swift's *Tale of a Tub.* There a relatively simple allegory of Reformation history (the *Tale* proper) is interrupted by a series of digressions that comment allegorically on the story into which they break.

Even the lyric poem can be adapted to yield allegorical themes and was made to do so, for example, in the visionary and rhapsodic odes written during the high Romantic period after the late 18th century throughout Europe.

The lesson seems to be that every literary genre is adaptable to the allegorical search for multiplicity of meaning.

Diversity of media. In the broadest sense, allegorical procedures are linguistic. Allegory is a manipulation of the language of symbols. Verbally, this mode underwent a major shift in medium along with the shift from oral to written literature: allegories that had initially been delivered in oral form (Christ's parables, for example) were written down by scribes and then transcribed by subsequent generations. Much more remarkable transformations, however, take place when the verbal medium is replaced by nonverbal or partially verbal media.

The drama is the chief of such replacements. The enactment of myth in the beginning had close ties with religious ritual, and in the drama of classical Greece both comedy and tragedy, by preserving ritual forms, lean toward allegory. Old Comedy, as represented by the majority of plays by Aristophanes, contains a curious blending of elements—allusions to men of the day, stories suggesting ideas other than the obvious literal sense, religious ceremony, parodies of the graver mysteries, personified abstractions, and stock types of character. Aeschylus' *Prometheus Bound* uses allegory for tragic ends, while Euripides' tragedies make a continuous interpretive commentary on the hidden meaning of the basic myths. Allegory is simplified in Roman drama, submitting heroic deeds to the control of the fickle, often malignant goddess Fortuna. Christian symbolism is responsible for the structure of the medieval morality plays, in which human dilemmas are presented through the conflicts of personified abstractions such as the "Virtues" and their "Vice" opponents. The allegory in Renaissance drama is often more atmospheric than structural—though even Shakespeare writes allegorical romances, such as *Cymbeline, Pericles,* and *The Winter's Tale* (and allowed his tragedy of *Coriolanus* to grow out of the "fable of the belly," which embodies a commonplace of Renaissance political wisdom and is recounted by one of the characters in the play). In 1598 Ben Jonson introduced the comedy of humours, which was dependent on the biological theory that the humours of the body (blood, phlegm, black bile, yellow bile) affect personality: in Jonson's play *Epicoene, or The Silent Woman* (1609), the character Morose is possessed by the demon of ill humour. Comic allegory of this kind evolved into the Restoration comedy of manners, and through that channel entered modern drama, with Wilde, Shaw, and Pirandello. Ibsen, the master of realistic drama, himself used a free-style allegory in *Peer Gynt,* while the surrealism of modern dramatists such as Ionesco, Genet, and Beckett serves to reinforce the real meaning of their plays.

The morality play

The degree to which the cinema has been allegorical in its methods has never been surveyed in detail. Any such survey would certainly reveal that a number of basic techniques in film montage builds up multiple layers of meaning. (Animated cartoons, too, continue the tradition of Aesopian fable.)

From time immemorial men have carved religious monuments and have drawn and painted sacred icons. Triumphal arches and chariots have symbolized glory and victory. Religious art makes wide use of allegory, both in its subject matter and in its imagery (such as the cross, the fish, the lamb). Even in poetry there can be an interaction of visual and verbal levels, sometimes achieved by patterning the stanza form. George Herbert's "Easter Wings," for instance, has two stanzas set out by the typographer to resemble the shape of a dove's wings. Such devices belong to the Renaissance tradition of the "emblem," which combines a motto with a simple symbolic picture (often a woodcut or engraving) and a concise explanation of the picture motto.

While allegory thrives on the visual, it has also been well able to embrace the empty form of pure mathematics. Number symbolism is very old: early Christian systems of cosmology were often based on the number three, referring to the doctrine of the Trinity (and in fact recalling earlier Hebraic and even Hellenic numerology). Musical symbolism has been discovered in the compositions of the 18th-century Baroque composers such as Johann Sebastian Bach. The most evanescent form of allegory, musical imagery and patterns, is also the closest to pure religious vision, since it merges the physical aspects of harmony (based on number) with the sublime and metaphysical effect on its hearers. The final extension of media occurs in the combination of spectacle, drama, dance, and music that is achieved by grand opera, which is at its most

The use of musical symbolism

allegorical in the total artwork of Richard Wagner in the second half of the 19th century. His *Ring* cycle of operas is a complete mythography and allegory, with words and music making two levels of meaning and the whole unified by a type of musical emblem, which Wagner called the leitmotif.

Allegory and cosmology. The allegorical mode has been of major importance in representing the cosmos: the earliest Greek philosophers, for example, speculated on the nature of the universe in allegorical terms; in the Old Testament's oblique interpretation of the universe, too, the world is seen as a symbolic system. The symbolic stories that explain the cosmos are ritualized to ensure that they encode a message. Held together by a system of magical causality, events in allegories are often surrounded by an occult atmosphere of charms, spells, talismans, genies, and magic rites. Science becomes science fiction or a fantastic setting blurs reality so that objects and events become metamorphically unstable. Allegorical fictions are often psychological dramas whose scene is the mind; then their protagonists are personified mental drives. Symbolic climate is most prominent in romance, whose heroic quests project an aura of erotic mysticism, perfect courtesy, and moral fervour that creates a sublime heightening of tone and a picturesque sense of good order.

The cosmic and demonic character of allegorical thinking is most fully reflected in the period of its greatest vogue, the High Middle Ages. During this period poets and priests alike were able to read with increasingly elaborate allegorical technique until their methods perhaps overgrew themselves. A belief had been inherited in the "Great Chain of Being," the Platonic principle of cosmic unity and fullness, according to which the lowest forms of being were linked with the highest in an ascending order. On the basis of this ladderlike conception were built systems of rising transcendency, starting from a material basis and rising to a spiritual pinnacle. The early Church Fathers sometimes used a threefold method of interpreting texts, encompassing literal, moral, and spiritual meanings. This was refined and commonly believed to have achieved its final form in the medieval allegorist's "fourfold theory of interpretation." This method also began every reading with a search for the literal sense of the passage. It moved up to a level of ideal interpretation in general, which was the allegorical level proper. (This was an affirmation that the true Christian believer was right to go beyond literal truth.) Still higher above the literal and the allegorical levels, the reader came to the tropological level, which told him where his moral duty lay. Finally, since Christian thought was apocalyptic and visionary, the fourfold method reached its apogee at the anagogic level, at which the reader was led to meditate on the final cosmic destiny of all Christians and of himself as a Christian hoping for eternal salvation.

The "fourfold theory of interpretation"

While modern scholars have shown that such thinking played its part in the poetry of the Middle Ages and while the Italian poet Dante himself discussed the theological relations between his poems and such a method of exegesis, the main arena for the extreme elaboration of this allegory was in the discussion and the teaching of sacred Scriptures. As such, the fourfold method is of highest import, and it should be observed that it did not need to be applied in a rigid four-stage way. It could be reduced, and commonly was reduced, to a two-stage method of interpretation. Then the reader sought simply a literal and a spiritual meaning. But it could also be expanded. The passion for numerology, combined with the inner drive of allegory toward infinite extension, led to a proliferation of levels. If four levels were good, then five or eight or nine might be better.

HISTORICAL DEVELOPMENT IN WESTERN CULTURE

Fable. The origins of fable are lost in the mists of time. Fables appear independently in ancient Indian and Mediterranean cultures. The Western tradition begins effectively with Aesop (6th century BC), of whom little or nothing is known for certain; but before him the Greek poet Hesiod (8th century BC) recounts the fable of the hawk and the nightingale, while fragments of similar tales survive in Archilochus, the 7th-century-BC warrior-poet. Within 100 years of the first Aesopian inventions the name of Aesop was firmly identified with the genre, as if he, not a collective folk, were its originator. Like the Greek philosopher Socrates, Aesop was reputed to have been ugly but wise. Legend connected him with the island of Samos; the historian Herodotus believed him to have been a slave.

Modern editions list approximately 200 "Aesop" fables, but there is no way of knowing who invented which tales or what their original occasions might have been. Aesop had already receded into legend when Demetrius of Phaleron, a rhetorician, compiled an edition of Aesop's fables in the 4th century BC. The poetic resources of the form developed slowly. A versified Latin collection made by Phaedrus, a freed slave in the house of the Roman emperor Augustus, included fables invented by the poet, along with the traditional favourites, which he retold with many elaborations and considerable grace. (Phaedrus may also have been the first to write topically allusive fables, satirizing Roman politics.) A similar extension of range marks the work of the Hellenized Roman Babrius, writing in the 2nd century AD. Among the classical authors who advanced upon Aesopian formulas may be named the Roman poet Horace, the Greek biographer Plutarch, and the great satirist Lucian of Samosata.

Beast epic. In the Middle Ages, along with every other type of allegory, fable flourished. Toward the end of the 12th century, Marie de France made a collection of over 100 tales, mingling beast fables with stories of Greek and Roman worthies. In another compilation, Christine de Pisan's Othéa manuscript illuminations provide keys to the interpretation of the stories and support the appended moral tag line. Expanded, the form of the fable could grow into what is called the beast epic, a lengthy, episodic animal story, replete with hero, villain, victim, and endless epic endeavour. (One motive for thus enlarging upon fable was the desire to parody epic grandeur: the beast epic mocks its own genre.) Most famous of these works is a 12th-century collection of related satirical tales called *Roman de Renart,* whose hero is a fox, symbolizing cunning man. The *Roman* includes the story of the fox and Chantecler (Chanticleer), a cock, a tale soon afterward told in German, Dutch, and English versions (in *The Canterbury Tales,* Geoffrey Chaucer took it as the basis for his "Nun's Priest's Tale"). Soon the *Roman* had achieved universal favour throughout Europe. The Renaissance poet Edmund Spenser also made use of this kind of material; in his "Mother Hubberd's Tale," published in 1591, a fox and an ape go off to visit the court, only to discover that life is no better there than in the provinces. More sage and serious, John Dryden's poem of *The Hind and Panther* (1687) revived the beast epic as a framework for theological debate. Bernard de Mandeville's *Fable of the Bees* (first published 1705 as *The Grumbling Hive, or Knaves Turn'd Honest*) illustrated the rapacious nature of humans in society through the age-old metaphor of the kingdom of the bees. In modern times, children's literature has made use of animal fable but often trivialized it. But the form has been taken seriously, as, for example, by the political satirist George Orwell, who, in his novel *Animal Farm* (1945), used it to attack Stalinist Communism.

Tales of the cunning fox

Influence of Jean de La Fontaine. The fable has normally been of limited length, however, and the form reached its zenith in 17th-century France, at the court of Louis XIV, especially in the work of Jean de La Fontaine. He published his *Fables* in two segments: the first, his initial volume of 1668, and the second, an accretion of "Books" of fables appearing over the next 25 years. The 1668 *Fables* follow the Aesopian pattern, but the later ones branch out to satirize the court, the bureaucrats attending it, the church, the rising bourgeoisie—indeed the whole human scene. La Fontaine's great theme was the folly of human vanity. He was a skeptic, not unkind but full of the sense of human frailty and ambition. His satiric themes permitted him an enlargement of poetic diction; he could be eloquent in mocking eloquence or in contrast use a severely simple style. (His range of tone and style has been admirably reflected in a version of his works made by a

20th-century American poet, Marianne Moore.) La Fontaine's example gave new impetus to the genre throughout Europe, and during the Romantic period a vogue for Aesopian fable spread to Russia, where its great practitioner was Ivan Andreyevich Krylov. The 19th century saw the rise of literature written specifically for children, in whom fable found a new audience. Among the most celebrated authors who wrote for them are Lewis Carroll, Charles Kingsley, Rudyard Kipling, Kenneth Grahame, Hilaire Belloc, and Beatrix Potter. There is no clear division between such authors and the "adult" fabulist, such as Hans Christian Andersen, Lewis Carroll, Oscar Wilde, Saint-Exupéry, or J.R.R. Tolkien. In the 20th century there are the outstanding *Fables for Our Time,* written by James Thurber and apparently directed toward an adult audience (although a sardonic parent might well read the *Fables* to his children).

Parable. In the West, the conventions of parable were largely established by the teachings of Christ. The New Testament records a sufficient number of his parables, with their occasions, to show that to some extent his disciples were chosen as his initiates and followers because they "had ears to hear" the true meaning of his parables. (It has already been noted that the parable can be fully understood only by an elite, made up of those who can decipher its inner core of truth.) Despite a bias toward simplicity and away from rhetorical elaboration, the parable loses little in the way of allegorical richness: the speaker can exploit an enigmatic brevity that is akin to the style of presenting a complex riddle. Parable is thus an immensely useful preaching device: while theologians in the period of the early Christian Church were developing glosses on Christ's enigmatic stories, preachers were inventing their own to drive home straightforward lessons in good Christian conduct. For centuries, therefore, the model of parable that had been laid down by Christ flourished on Sundays in churches all over the Western world. Pious tales were collected in handbooks: the *Gesta Romanorum,* the *Alphabet of Thales,* the *Book of the Knight of La Tour Landry,* and many more. Infinitely varied in subject matter, these exemplary tales used a plain but lively style, presenting stories of magicians, necromancers, prophets, chivalrous knights and ladies, great emperors—a combination bound to appeal to congregations, if not to theologians. An important offshoot of the parable and exemplary tale was the saint's life. Here, too, massive compilations were possible; the most celebrated was *The Golden Legend* of the 15th century, which included approximately 200 stories of saintly virtue and martyrdom.

20th-century parables

As long as preaching remained a major religious activity, the tradition of parable preserved its strong didactic strain. Its more paradoxical aspect gained renewed lustre in theological and literary spheres when the 19th-century Danish philosopher Søren Kierkegaard began to use parables in his treatises on Christian faith and action. In *Fear and Trembling* he retold the story of Abraham and Isaac; in *Repetition* he treated episodes in his own life in the manner of parable. Such usage led to strange new literary forms of discourse, and his writing influenced, among others, the Austro-Czech novelist Franz Kafka and the French "absurdist" philosopher, novelist, and playwright Albert Camus. Kafka's parables, full of doubt and anxiety, mediate on the infinite chasm between man and God and on the intermediate role played by the law. His vision, powerfully expressed in parables of novel length (*The Castle, The Trial, Amerika*), is one of the most enigmatic in modern literature.

Allegory. The early history of Western allegory is intricate and encompasses an interplay between the two prevailing world views—the Hellenic and the Hebraic-Christian—as theologians and philosophers attempted to extract a higher meaning from these two bodies of traditional myth.

In terms of allegory, the Greco-Roman and Hebraic-Christian cultures both have a common starting point: a creation myth. The Old Testament's book of Genesis roughly parallels the story of the creation as told by the Greek poet Hesiod in his *Theogony* (and the later Roman version of the same event given in Ovid's *Metamorphoses*). The two traditions thus start with an adequate source of cosmic imagery, and both envisage a universe full of mysterious signs and symbolic strata. But thereafter the two cultures diverge. This is most apparent in the way that the style of the body of poetry attributed to Homer—the ancient Greek "Bible"—differs from the Old Testament narrative. The Greek poet presented his heroes against an articulated narrative scene, a context full enough for the listener (and, later, the reader) to ignore secondary levels of significance. By contrast, the Jewish authors of the Old Testament generally emptied the narrative foreground, leaving the reader to fill the scenic vacuum with a deepening, thickening allegorical interpretation.

Old Testament. The Old Testament, including its prophetic books, has a core of historical record focussing on the trials of the tribes of Israel. In their own view an elect nation, the Israelites believe their history spells out a providential design. The prophets understand the earliest texts, Genesis and Exodus, in terms of this providential scheme. Hebraic texts are interpreted as typological: that is, they view serious myth as a theoretical history in which all events are types—portents, foreshadowing the destiny of the chosen people. Christian exegesis (the critical interpretation of Scripture) inherits the same approach.

Typological allegory

Typological allegory looks for hidden meaning in the lives of actual men who, as types or figures of later historical persons, serve a prophetic function by prefiguring those later persons. Adam, for example (regarded as a historical person), is thought to prefigure Christ in his human aspect, Joshua to prefigure the victorious militant Christ. This critical approach to Scripture is helped by the fact of monotheism, which makes it easier to detect the workings of a divine plan. The splendours of nature hymned in the Psalms provide a gloss upon the "glory of God." The Law (the Torah) structures the social aspect of sacred history and, as reformulated by Christ, provides the chief link between Old and New Testaments. Christ appeals to the authority of "the Law and the Prophets" but assumes the ultimate prophetic role himself, creating the New Law and the New Covenant—or Testament—with the same one God of old.

The Greeks. Hellenic tradition after Homer stands in sharp contrast to this concentration on the fulfilling of a divine plan. The analytic, essentially scientific histories of Herodotus and Thucydides precluded much confident belief in visionary providence. The Greeks rather believed history to be structured in cycles, as distinct from the more purposive linearity of Hebraic historicism.

Nevertheless, allegory did find a place in the Hellenic world. Its main arena was in philosophic speculation, centring on the interpretation of Homer. Some philosophers attacked and others defended the Homeric mythology. A pious defense argued that the stories—about the monstrous love affairs of the supreme god Zeus, quarrels of the other Olympian gods, scurrility of the heroes, and the like—implied something beyond their literal sense. The defense sometimes took a scientific, physical form; in this case, Homeric turmoil was seen as reflecting the conflict between the elements. Or Homer was moralized; the goddess Pallas Athene, for example, who in physical allegory stood for the ether, in moral allegory was taken to represent reflective wisdom because she was born out of the forehead of her father, Zeus. Moral and physical interpretation is often intermingled.

Plato, the Idealist philosopher, occupies a central position with regard to Greek allegory. His own myths imply that our world is a mere shadow of the ideal and eternal world of forms (the Platonic ideas), which has real, independent existence, and that the true philosopher must therefore be an allegorist in reverse. He must regard phenomena—things and events—as a text to be interpreted upward, giving them final value only insofar as they reveal, however obscurely, their ideal reality in the world of forms. Using this inverted allegorical mode, Plato attacked Homeric narrative, whose beauty beguiles men into looking away from the truly philosophic life. Plato went further. He attacked other fashionable philosophic allegorists because they did not lead up to the reality but limited speculation to the sphere of moral and physical necessity. Platonic

allegory envisaged the system of the universe as an ascending ladder of forms, a "Great Chain of Being," and was summed up in terms of myth in his *Timaeus.* Plato and Platonic thought became, through the influence of this and other texts on Plotinus (died 269/270) and through him on Porphyry (died *c.* 304), a pagan mainstay of later Christian allegory. Medieval translations of Dionysius the Areopagite (before 6th century AD) were equally influential descendants of Platonic vision.

Stoic allegory

A second and equally influential Hellenic tradition of allegory was created by the Stoic philosophers, who held that the local gods of the Mediterranean peoples were signs of a divinely ordered natural destiny. Stoic allegory thus emphasized the role of fate, which, because all men were subject to it, could become a common bond between peoples of different nations. A later aspect of moral exegesis in the Stoic manner was the notion that myths of the gods really represent, in elevated form, the actions of great men. In the 2nd century BC, under Stoic influence, the Sicilian writer Euhemerus argued that theology had an earthly source. His allegory of history was the converse of Hebraic typology—which found the origin of the divine in the omnipotence of the One God—for Euhemerus found the origin of mythological gods in human kings and heroes, divinized by their peoples. His theories enjoyed at least an aesthetic revival during the Renaissance.

Blending of rival systems: the Middle Ages. At the time of the birth of Christ, ideological conditions within the Mediterranean world accelerated the mingling of Hellenic and Hebraic traditions. Philo Judaeus laid the groundwork; Clement of Alexandria and Origen followed him. The craft of allegorical syncretism—that is, making rival systems accommodate one another through the transformation of their disparate elements—was already a developed art by the time St. Paul and the author of the Gospel According to St. John wove the complex strands of the Hebraic–Christian synthesis. Over centuries of quarrelling, the timeless philosophy of the Greek allegorists was accommodated to the time-laden typology of the Hebrew prophets and their Christian successors and at length achieved a hybrid unity that permitted great allegories of Western Christendom to be written.

As a hybrid method, allegory could draw on two archetypal story lines: the war and the quest of Homer's *Iliad* and *Odyssey,* which was paralleled by the struggles and wanderings of the children of Israel. Throughout the Middle Ages the figure of the wandering Aeneas (who, in the second half of Virgil's Latin epic, *Aeneid,* fought bloody battles) was seen as a type in a system of hidden Christianity. Virgil's fourth *Eclogue,* a prophetic vision of the birth of a child who would usher in the "golden age," was read as a prophecy of the birth of Christ. Seen by many Christian commentators as the ideal allegorist, Virgil himself was hailed as a proto-Christian prophet. The blending of rival systems of allegory from widely assorted cultures became the rule for later allegory. Adapting the Latin writer Apuleius' fable of Cupid and Psyche, Edmund Spenser combined its elements with ancient Middle Eastern lore, Egyptian wisdom, and dashes of Old Testament critical interpretation to convert the enclosed garden of the biblical Song of Solomon into the gardens of Adonis in *The Faerie Queene,* Book III. The pagan gods survived unharmed throughout the Middle Ages if wearing Christian costumes, because Christians were taught that pagan worthies could be read as figures of Christian rulers. The labours of Hercules, for instance, stood for the wanderings and trials of all Christian men; the Hellenic theme of heroic warfare took a Christianized form, available to allegory, when in the 4th century the poet and hymn writer Prudentius internalized war as the inner struggle of Christian man, suspended between virtue and vice. For complete triumph in explaining the significance of the world, Christianity needed one further element: a world-historical theory large enough to contain all other theories of meaning. This it found in the belief that God was the author of the world. His creation wrote the world. The world, read as a text, provided a platform for transforming the piecemeal, postclassical syncretism into some semblance of order. Firmly established in the West, Christianity, for all its strains of discord, slowly achieved a measure of coherence. St. Thomas Aquinas could write its *Summa.* Theocentric, authoritarian, spiritualist, and word oriented, the medieval model of allegory lent itself to the creation of the most wonderful of all allegorical poems, Dante's *Divine Comedy,* completed shortly before his death in 1321.

The importance of the Christian world view in allegory

Before this could happen, however, the Christian world view had been subjected to an important pressure during the 12th century. It may be called the pressure to externalize. Alanus de Insulis (Alain de Lille), Bernard of Sylvestris, John of Salisbury, and other forerunners of the movement known as European Humanism "discovered" nature. Delighting in the wonders of God's cosmic text, they brought theological speculation down to earth. Romances of love and chivalry placed heroes and heroines against the freshness of spring. Everywhere nature shone, sparkling with the beauty of earthly life. The externalization and naturalizing of Christian belief flowers most obviously in *The Romance of the Rose,* begun in the 13th century by Guillaume de Lorris and completed, in vastly complicated form, by Jean de Meung. The *Romance* personifies the experiences of courtly love, recounting the pursuit of an ideal lady by an ideal knight, set in an enclosed garden and castle, which permits Guillaume to dwell on the beauty of nature. With Jean de Meung the interest in nature is made explicit, and the poem ends in a series of lengthy digressive discourses, several of them spoken by Dame Nature herself. In medieval English poetry this same love of spring and seasonal pleasures is apparent everywhere—certainly in the poems of Geoffrey Chaucer, who, besides creating several allegories of his own, translated *The Romance of the Rose* into English.

Dante's *Divine Comedy* has physical immediacy and contains an immense amount of historical detail. He anchors his poem in a real world, accepting Christian typology as historical fact and adopting an ordered system of cosmology (based on the number three, proceeding from the Trinity). Dante's passion for numerology does not, however, block a closeness to nature that had perhaps not been equalled in poetry since Homer. He enfolds classical thought into his epic by making Virgil one of its main protagonists—again to prefigure Christian heroism. Perhaps only William Langland, the author of *The Vision of Piers Plowman,* could be said to rival Dante's cosmic range. *Piers Plowman* is a simpler apocalyptic vision than the *Comedy,* but it has an existential immediacy, arising from its concern for the poor, which gives it great natural power.

Renaissance. Romance and romantic forms provide the main vehicle for the entrance of allegory into the literature of the Renaissance period. The old Arthurian legends carry a new sophistication and polish in the epics of the Italians Boiardo, Ariosto, and Tasso and in the work of Edmund Spenser. By interlacing several simultaneous stories in one larger narrative, the literary technique known as *entrelacement* allowed digression—yet kept an ebbing, flowing kind of unity—while presenting opportunities for moral and ironic commentary. But although the forms and themes of romance were medieval in origin, the new age was forced to accommodate altered values. The Middle Ages had externalized the Christian model; the Renaissance now internalized it, largely by emphasizing the centrality of human understanding. This process of internalization had begun slowly. In rough outline it can be discerned in the belief that biological humours affected personality, in the adaptations of Platonic Idealism from which arose a new emphasis on imagination, in the rise of an introspective, soliloquizing drama in England. It can further be discerned in the gradual adoption of more self-conscious theories of being: Shakespeare's Hamlet, finding himself by thinking out his situation, prefigures the first modern philosopher, René Descartes, whose starting point for argument was "I think, therefore I am." Christopher Marlowe's characterization of Dr. Faustus epitomizes the new age. Pursuing power in the form of knowledge, he is led to discover the demons of allegory within himself. He is an essential figure for later European literature, archetypal in Germany for both Johann Wolfgang von Goethe and Thomas Mann and influential everywhere.

Modern period. With the Baroque and Neoclassical periods, allegory began to turn away from cosmology and toward rhetorical ambiguity. John Milton allegorized sin and death in his epic poem *Paradise Lost,* but allegory for him seems chiefly to lie in the ambiguous diction and syntax employed in the poem. Instead of flashing allegorical emblems before the reader, Milton generates a questioning attitude that searches out allegory more as a mysterious form than a visible content. His central allegorical theme is perhaps the analogy he draws between poetry, music, and ideas of cosmic order. This theme, which generates allegory at once, recurs in later English poetry right up to modern times with T.S. Eliot's *Four Quartets.*

Milton's allegory: *Paradise Lost*

The social and religious attitudes of the 18th-century age of Enlightenment could be expressed coolly and without ambiguity—and thus there was little need for spiritual allegory in the period's literature. Oblique symbolism was used mainly for satirical purposes. John Dryden and Alexander Pope were masters of verse satire, Jonathan Swift of prose satire. Voltaire and the French writers of the Enlightenment similarly employed a wit whose aim was to cast doubt on inherited pieties and attitudes. A new vogue for the encyclopaedia allowed a close, critical commentary on the ancient myths, but the criticism was rationalist and opposed to demonology. Under such conditions the allegorical mode might have dried up entirely. Yet the new Romantic age of the late 18th and early 19th centuries revived the old cosmologies once more, and poetic forms quickly reflected the change, with the Romantic poets and their precursors (Blake, William Collins, Edward Young, Thomas Gray, and others) managing to reinstate the high destiny of the allegorical imagination. The Romantics went back to nature. Poets took note of exactly what they saw when they went out walking, and their awareness of nature and its manifestations found its way into their poetry. Appropriate poetic forms for expressing this sensibility tended to be open, rhapsodic, and autobiographical—qualities notably present in William Wordsworth and in Samuel Taylor Coleridge, for example. Percy Bysshe Shelley is the most strikingly allegorical of English Romantics; he not only followed the Platonic tradition of Spenser and the Renaissance—with ode, elegy, and brief romance—but he also invented forms of his own, such as *Epipsychidion,* a rhapsodic meditation, and he was working on a great Dantesque vision, *The Triumph of Life,* when he died. Visionary masterpieces came from Germany, where Novalis and Friedrich Hölderlin hymned the powers of nature in odes of mythic overtone and resonance. French Romanticism, merging gradually with the theory and practice of the Symbolist movement (dealing in impressions and intuitions rather than in descriptions), in turn followed the same path. The pantheist cosmology of Victor Hugo, the central writer of the somewhat delayed French Romantic movement, created an allegory of occult forces and demonic hero worship. It is fair to say that, in its most flexible and visionary forms, allegory flourished throughout the Romantic period.

There also developed a novelistic mode of allegory by which prose authors brought fate, necessity, the demonic, and the cosmological into their narratives. Émile Zola used a theory of genetics, Charles Dickens the idea of ecological doom, Leo Tolstoy the belief in historical destiny, and Fyodor Dostoyevsky the fatalism of madness and neurosis. Nikolay Gogol revived the art of the grotesque, picturing absurdities in the scene of tsarist Russia. Even the arch-naturalist playwright, Anton Chekhov, made an emblem of the cherry orchard and the sea gull in his plays of those titles. However its dates are established, the modern period is exceedingly complex in its mythmaking. Psychoanalytic theory has been both a critical and a creative resource; modern allegory has remained internalized in the Renaissance tradition. But Marxist social realism has kept to the externals of dialectical materialism, though without notable aesthetic success. In the free play of American letters, where Nathaniel Hawthorne, Herman Melville, Edgar Allen Poe, and Henry James (particularly in his later novels) had essayed an allegorical mode, the future of its use is uncertain. T.S. Eliot's enigmatic style in a long poem, "Ash Wednesday," may be related to his search for a Dantesque dramatic style, for which he also tried in plays, most obviously *Murder in the Cathedral* (a morality) and *The Cocktail Party* (a philosophic farce). More clearly popular authors such as George Orwell and William Golding have used the most familiar allegorical conventions. D.H. Lawrence shaped novels such as *The Plumed Serpent* to project a thematic, cultural polemic. W.H. Auden's operatic librettos reflect once more the allegorical potential of this mixture of media.

Complexity of mythmaking in the 20th century

Modern allegory has in fact no set pattern, or model, although Surrealism has provided a dominant style of discontinuous fragmentary expression. The only rule seems to be that there is no rule. Science fiction, an ancient field dating back at least to the earliest philosophers of Greece, has set no limits on the speculations it will entertain. The allegorical author now even questions the allegorical process itself, criticizing the very notions of cosmos, demon, and magic. It may be that modern allegory has completed a vast circle begun by the first conflict between ways of interpreting myth, as revealed in Homer and the Hebraic prophets.

(A.S.F.)

ALLEGORICAL LITERATURE IN THE EAST

India. Fables appeared early in India, but it is impossible to determine whether they are older or later than the Greek. Undoubtedly there was mutual influence from very early times, for indirect contacts between Greece and India (by trade routes) had existed long before the time of Alexander the Great. In the form in which they are now known the Greek fables are the older, but this may be an accident of transmission.

The fable was apparently first used in India as a vehicle of Buddhist instruction. Some of the *Jātakas,* birth stories of the Buddha, which relate some of his experiences in previous animal incarnations, resemble Greek fables and are used to point a moral. They may date from as far back as the 5th century BC, though the written records are much later. The most important compilation is the *Pañcatantra,* a Sanskrit collection of beast fables. The original has not survived, but it has been transmitted (via a lost Pahlavi version) as the mid-8th-century Arabic *Kalīlah wa Dimnah.* Kalilah and Dimnah are two jackals, counselors to the lion king, and the work is a frame story containing numerous fables designed to teach political wisdom or cunning. From the Arabic this was translated into many languages, including Hebrew, which version John of Capua used to make a Latin version in the 13th century. This, the *Directorium humanae vitae* ("Guide for Human Life"), was the chief means by which oriental fables became current in Europe. In the fables of Bidpai, animals act as men in animal form, and little attention is paid to their supposed animal characteristics. It is in this respect that they differ most from the fables of Aesop, in which animals behave as animals.

(Ed.)

China. Chinese philosophers from the Ch'in dynasty (221–206 BC) onward often used extended metaphors (from which fable is the logical development) to make their points. This is believed to reflect the fact that, as "realistic" thinkers, the Chinese generally did not favour more abstract argument. Thus simple allegory helped to stimulate audience interest and to increase the force of an argument. A century earlier, Mencius, a Confucian philosopher, had used the following little allegory in illustrating his theory that an effort has to be made if man's natural goodness is to be recovered: "A man will begin searching when his dog or chicken is missing; but he does not go searching for the good character he was born with after it is lost. Is this not regrettable?" The same writer also used a parable to bring home his point that mental training could not be hurried, but was a gradual process: "A man in Sung sowed seeds in a field. The seedlings grew so slowly, however, that one day he took a walk through the field pulling at each one of the seedlings. On returning home he announced that he was exhausted, but that he had helped the seedlings' growth. His son, hurrying to the field, found the seedlings dead."

Tales such as this were often borrowed from folklore,

but others were probably original creations, including a striking story that opens the *Chuang-tzu,* a summa of Taoist thought. It makes the point that ordinary people frequently deplore the actions of a man of genius because they are unable to understand his vision, which is not answerable to the laws of "common sense": "A giant fish, living at the northern end of the world, transformed itself into a bird so that it could make the arduous flight to the southernmost sea. Smaller birds, measuring his ambition against their own capabilities, laughed at the impossibility of it."

But the full development of fable, as it is understood in the West, was hindered by the fact that Chinese ways of thinking prohibited them from accepting the notion of animals that thought and behaved as humans. Actual events from the past were thought to be more instructive than fictitious stories, and this led to the development of a large body of legendary tales and supernatural stories. Between the 4th and 6th centuries, however, Chinese Buddhists adapted fables from Buddhist India in a work known as *Po-Yü ching,* and they also began to make use of traditional Chinese stories that could further understanding of Buddhist doctrines. (Na.Mo.)

Japan. In Japan, the *Koji-ki* (712; "Records of Ancient Matters") and the *Nihon-shoki* (8th century, "Chronicles of Japan"), both of them official histories of Japan, were studded with fables, many of them on the theme of a small intelligent animal getting the better of a large stupid one. The same is true of the *fudoki* (local gazetteers dating from 713 and later). The form reached its height in the Kamakura period (1192–1333). Toward the end of the Muromachi period (1338–1573) Jesuit missionaries introduced the fables of Aesop to Japan, and the influence of these can be traced in stories written between then and the 19th century. (T.Io.)

Ballad

The ballad is a short narrative folk song whose distinctive style crystallized in Europe in the late Middle Ages and persists to the present day in communities where literacy, urban contacts, and mass media have not yet affected the habit of folk singing. France, Denmark, Germany, Russia, Greece, and Spain, as well as England and Scotland, possess impressive ballad collections. At least one-third of the 300 extant English and Scottish ballads have counterparts in one or several of these continental balladries, particularly those of Scandinavia. In no two language areas, however, are the formal characteristics of the ballad identical. For example, British and American ballads are invariably rhymed and strophic (*i.e.,* divided into stanzas); the Russian ballads known as *byliny* and almost all Balkan ballads are unrhymed and unstrophic; and, though the *romances* of Spain, as their ballads are called, and the Danish *viser* are alike in using assonance instead of rhyme, the Spanish ballads are generally unstrophic while the Danish are strophic, parcelled into either quatrains or couplets.

ELEMENTS

Narrative basis. Typically, the folk ballad tells a compact little story that begins eruptively at the moment when the narrative has turned decisively toward its catastrophe or resolution. Focussing on a single, climactic situation, the ballad leaves the inception of the conflict and the setting to be inferred or sketches them in hurriedly. Characterization is minimal, the characters revealing themselves in their actions or speeches; overt moral comment on the characters' behaviour is suppressed and their motivation seldom explicitly detailed. Whatever description occurs in ballads is brief and conventional; transitions between scenes are abrupt and time shifts are only vaguely indicated; crucial events and emotions are conveyed in crisp, poignant dialogue. In short, the ballad method of narration is directed toward achieving a bold, sensational, dramatic effect with purposeful starkness and abruptness. But despite the rigid economy of ballad narratives, a repertory of rhetorical devices is employed for prolonging highly charged moments in the story and thus thickening the emotional atmosphere. In the most famous of such devices, incremental repetition, a phrase or stanza is repeated several times with a slight but significant substitution at the same critical point. Suspense accumulates with each substitution, until at last the final and revelatory substitution bursts the pattern, achieving a climax and with it a release of powerful tensions. The following stanza is a typical example:

Rhetorical devices

Then out and came the thick, thick, blood,
Then out and came the thin,
Then out and came the bonny heart's blood,
Where all the life lay in.

Oral transmission. Since ballads thrive among unlettered people and are freshly created from memory at each separate performance, they are subject to constant variation in both text and tune. Where tradition is healthy and not highly influenced by literary or other outside cultural influences, these variations keep the ballad alive by gradually bringing it into line with the style of life, beliefs, and emotional needs of the immediate folk audience. Ballad tradition, however, like all folk arts, is basically conservative, a trait that explains the references in several ballads to obsolete implements and customs, as well as the appearance of words and phrases that are so badly garbled as to indicate that the singer does not understand their meaning though he takes pleasure in their sound and respects their traditional right to a place in his version of the song. The new versions of ballads that arise as the result of cumulative variations are no less authentic than their antecedents. A poem is fixed in its final form when published, but the printed or taped record of a ballad is representative only of its appearance in one place, in one line of tradition, and at one moment in its protean history. The first record of a ballad is not its original form but merely its earliest recorded form, and the recording of a ballad does not inhibit tradition from varying it subsequently into other shapes, because tradition preserves by re-creating rather than by exact reproduction.

COMPOSITION

Theories. How ballads are composed and set afloat in tradition has been the subject of bitter quarrels among scholars. The so-called communal school, which was led by two American scholars F.B. Gummere (1855–1919) and G.L. Kittredge (1860–1941), argued at first that ballads were composed collectively during the excitement of dance and song festivals. Under attack the communalists retreated to the position that although none of the extant ballads had been communally composed, the prototypical ballads that determined the style of the ballads had originated in this communal fashion. Their opponents were the individualists, who included the British men of letters W.J. Courthope (1842–1917) and Andrew Lang (1844–1912) and the American linguist Louise Pound (1872–1958). They held that each ballad was the work of an individual composer, who was not necessarily a folk singer, tradition serving simply as the vehicle for the oral perpetuation of the creation. According to the widely accepted communal re-creation theory, put forward by the American collector Phillips Barry (1880–1937) and the scholar G.H. Gerould (1877–1953), the ballad is conceded to be an individual composition originally. This fact is considered of little importance because the singer is not expressing himself individually, but serving as the deputy of the public voice, and because a ballad does not become a ballad until it has been accepted by the folk community and been remolded by the inevitable variations of tradition into a communal product. Ballads have also been thought to derive from art songs, intended for sophisticated audiences, which happened to filter down to a folk level and become folk song. This view, though plausible in the case of certain folk lyrics, is inapplicable to the ballads, for if the ballads were simply miscellaneous castoffs, it would not be possible to discern so clearly in them a style that is unlike anything in sophisticated verse.

Technique and form. Ballads are normally composed in two kinds of stanzas; the first consists of a couplet of lines each with four stressed syllables, and with an interwoven refrain:

But it would have made your heart right sair,
With a hey ho and a lillie gay
To see the bridegroom rive his haire.
As the primrose spreads so sweetly

the second a stanza of alternating lines of four stresses and three stresses, the second and fourth lines rhyming:

There lived a wife at Usher's Well,
And a wealthy wife was she;
She had three stout and stalwart sons,
And sent them o'er the sea.

Reference to the tunes show that the three-stress lines actually end in an implied fourth stress to match the pause in the musical phrase at these points. The interwoven refrain is a concession to the musical dimension of the ballad; it may be a set of nonsense syllables (Dillum down dillum, Fa la la la) or irrelevant rigmaroles of flowers or herbs. A few ballads have stanza-length burdens interspersed between the narrative stanzas, a technique borrowed from the medieval carols. The lyrical and incantatory effect of refrains during the ballad performance is very appealing, but in cold print they often look ridiculous, which is perhaps why early collectors failed to note them. In the first example above, it will be noted that the gaiety of the refrain is at odds with the mood of the meaningful lines. Not infrequently the ballad stanza satisfies the music's insistence on lyrical flourishes by repeating textual phrases and lines:

So he ordered the grave to be opened wide,
And the shroud to be turned down;
And there he kissed her clay cold lips
Till the tears came trickling down, down, down,
Till the tears came trickling down

Kinds of repetition used in ballads

The refrain is just one of the many kinds of repetition employed in ballads. Incremental repetition, already discussed, is the structural principle on which whole ballads ("The Maid Freed from the Gallows," "Lord Randal") are organized, and many other ballads contain long exchanges of similarly patterned phrases building cumulatively toward the denouement:

"Oh what will you leave to your father dear?"
"The silver-shod steed that brought me here."
"What will you leave to your mother dear?"
"My velvet pall and my silken gear."
"What will you leave to your brother John?"
"The gallows-tree to hang him on."

Any compressed narrative of sensational happenings told at a high pitch of feeling is bound to repeat words and phrases in order to accommodate the emotion that cannot be exhausted in one saying, a tendency that accounts for such stanzas as:

Then He says to His mother, "Oh: the withy [willow], oh:
the withy,
The bitter withy that causes me to smart, to
smart,
Oh: the withy, it shall be the very first tree
That perishes at the heart."

Much repetition in ballads is mnemonic as well as dramatic. Since ballads are performed orally, the hearer cannot turn back a page to recover a vital detail that slipped by in a moment of inattention. Crucial facts in narrative, therefore, are incised in the memory by skillful repetition; instructions given in a speech are exactly repeated when the singer reports the complying action; answers follow the form of the questions that elicited them.

Conventional imagery

The exigencies of oral performance also account for the conventional stereotyped imagery of the ballads. For unlike the poet, who reaches for the individualistic, arresting figure of speech, the ballad singer seldom ventures beyond a limited stock of images and descriptive adjectives. Knights are always gallant, swords royal, water wan, and ladies gay. Whatever is red is as red as blood, roses, coral, rubies, or cherries; white is stereotyped as snow white, lily white, or milk white. Such conventions fall into place almost by reflex action, easing the strain on the singer's memory and allowing him to give his full attention to the manipulation of the story. The resulting bareness of verbal texture, however, is more than compensated for by the dramatic rhetoric through which the narrative is projected. In any case, complex syntax and richness of language are forbidden to texts meant to be sung, for music engages too much of the hearer's attention for him to untangle an ambitious construction or relish an original image. Originality indeed, like anything else that exalts the singer, violates ballad decorum, which insists that the singer remain impersonal.

Music. A ballad is not technically a ballad unless it is sung; but though tunes and texts are dynamically interdependent, it is not unusual to find the same version of a ballad being sung to a variety of tunes of suitable rhythm and metre or to find the same tune being used for several different ballads. And just as there are clusters of versions for most ballads, so a given ballad may have associated with it a family of tunes whose members appear to be versions of a single prototypical form.

Ballad tunes are based on the modes rather than on the diatonic and chromatic scales that are used in modern music. Where chromaticism is detected in American folk music, the inflected tones are derived from black folk practice or from learned music. Of the six modes, the preponderance of folk tunes are Ionian, Dorian, or Mixolydian; Lydian and Phrygian tunes are rare. The folk music least affected by sophisticated conditioning does not avail itself of the full seven tones that compose each of the modal scales. Instead, it exhibits gapped scales, omitting either one of the tones (hexatonic) or two of them (pentatonic). Modulation sometimes occurs in a ballad from one mode to an adjacent mode.

Most tunes consist of 16 bars with duple rhythm, or two beats per measure, prevailing slightly over triple rhythm. The tune, commensurate with the ballad stanza, is repeated as many times as there are stanzas. Unlike the "through-composed" art song, where the music is given nuances to correspond to the varying emotional colour of the content, the folk song affords little opportunity to inflect the contours of the melody. This limitation partly explains the impassive style of folk singing, Musical variation, however, is hardly less frequent than textual variation; indeed, it is almost impossible for a singer to perform a ballad exactly the same way twice. The stablest part of the tune occurs at the mid-cadence (the end of the second text line) and the final cadence (the end of the fourth line). The third phrase of the tune, corresponding to the third line of the stanza, proves statistically the most variable. Significantly, these notes happen to coincide with the rhyming words. The last note of the tune, the point of resolution and final repose, usually falls on the fundamental tone (*i.e.*, keynote) of the scale; the mid-cadence falling normally a perfect fifth above the tonic or a perfect fourth below it. To make for singability, the intervals in the melodic progression seldom involve more than three degrees. And since the singer performs solo or plays the accompanying instrument himself, he need not keep rigidly to set duration or stress but may introduce grace notes to accommodate hypermetric syllables and lengthen notes for emphasis.

Impassive style of folk singing

TYPES OF BALLADRY

The traditional folk ballad, sometimes called the Child ballad in deference to Francis Child, the scholar who compiled the definitive English collection, is the standard kind of folk ballad in English and is the type of balladry that this section is mainly concerned with. But there are peripheral kinds of ballads that must also be noticed in order to give a survey of balladry.

Minstrel ballad. Minstrels, the professional entertainers of nobles, squires, rich burghers, and clerics until the 17th century, should properly have had nothing to do with folk ballads, the self-created entertainment of the peasantry. Minstrels sometimes, however, affected the manner of folk song or remodelled established folk ballads. Child included many minstrel ballads in his collection on the ground that fragments of traditional balladry were embedded in them. The blatant style of minstrelsy marks these ballads off sharply from folk creations. In violation of the strict impersonality of the folk ballads, minstrels constantly intrude into their narratives with moralizing comments and fervent assurances that they are not lying at the very moment when they are most fabulous. The

Difference between minstrel and folk ballads

ministrels manipulate the story with coarse explicitness, begging for attention in a servile way, predicting future events in the story and promising that it will be interesting and instructive, shifting scenes obtrusively, reflecting on the characters' motives with partisan prejudice. Often their elaborate performances are parcelled out in clear-cut divisions, usually called fits or cantos, in order to forestall tedium and build up suspense by delays and piecemeal revelations. Several of the surviving minstrel pieces are poems in praise of such noble houses as the Armstrongs ("Johnie Armstrong"), the Stanleys ("The Rose of England"), and the Percys ("The Battle of Otterburn," "The Hunting of the Cheviot," "The Earl of Westmoreland"), doubtless the work of propagandists in the employ of these families. The older Robin Hood ballads are also minstrel propaganda, glorifying the virtues of the yeomanry, the small independent landowners of preindustrial England. The longer, more elaborate minstrel ballads were patently meant to be recited rather than sung.

Broadside ballad. Among the earliest products of the printing press were broadsheets about the size of handbills on which were printed the text of ballads. A crude woodcut often headed the sheet, and under the title it was specified that the ballad was to be sung to the tune of some popular air. Musical notation seldom appeared on the broadsides; those who sold the ballads in the streets and at country fairs sang their wares so that anyone unfamiliar with the tune could learn it by listening a few times to the balladmonger's rendition. From the 16th century until the end of the 19th century, broadsides, known also as street ballads, stall ballads, or slip songs, were a lively commodity, providing employment for a troop of hack poets. Before the advent of newspapers, the rhymed accounts of current events provided by the broadside ballads were the chief source of spectacular news. Every sensational public happening was immediately clapped into rhyme and sold on broadsheets. Few of the topical pieces long survived the events that gave them birth, but a good number of pathetic tragedies, such as "The Children in the Wood" and broadsides about Robin Hood, Guy of Warwick, and other national heroes, remained perennial favourites. Although the broadside ballad represents the adaptation of the folk ballad to the urban scene and middle class sensibilities, the general style more closely resembles minstrelsy, only with a generous admixture of vulgarized traits borrowed from book poetry. A few folk ballads appeared on broadsheets; many ballads, however, were originally broadside ballads the folk adapted.

Broadside ballad as a source of news

Literary ballads. The earliest literary imitations of ballads were modelled on broadsides, rather than on folk ballads. In the early part of the 18th century, Jonathan Swift, who had written political broadsides in earnest, adapted the style for several jocular bagatelles. Poets such as Swift, Matthew Prior, and William Cowper in the 18th century and Thomas Hood, W.M. Thackeray, and Lewis Carroll in the 19th century made effective use of the jingling metres, forced rhymes, and unbuttoned style for humorous purposes. Lady Wardlaw's "Hardyknute" (1719), perhaps the earliest literary attempt at a folk ballad, was dishonestly passed off as a genuine product of tradition. After the publication of Thomas Percy's ballad compilation *Reliques of Ancient English Poetry* in 1765, ballad imitation enjoyed a considerable vogue, which properly belongs in the history of poetry rather than balladry.

SUBJECT MATTER

The supernatural. The finest of the ballads are deeply saturated in a mystical atmosphere imparted by the presence of magical appearances and apparatus. "The Wife of Usher's Well" laments the death of her children so unconsolably that they return to her from the dead as revenants; "Willie's Lady" cannot be delivered of her child because of her wicked mother-in-law's spells, an enchantment broken by a beneficent household spirit; "The Great Silkie of Sule Skerry" begets upon an "earthly" woman a son, who, on attaining maturity, joins his seal father in the sea, there shortly to be killed by his mother's human husband; "Kemp Owyne" disenchants a bespelled maiden by kissing her despite her bad breath and savage looks. An encounter between a demon and a maiden occurs in "Lady Isabel and the Elf-Knight," the English counterpart of the ballads known to the Dutch-Flemish as "Herr Halewijn," to Germans as "Ulinger," to Scandinavians as "Kvindemorderen" and to the French as "Renaud le Tueur de Femme." In "The House Carpenter," a former lover (a demon in disguise) persuades a wife to forsake husband and children and come away with him, a fatal decision as it turns out. In American and in late British tradition the supernatural tends to get worked out of the ballads by being rationalized: instead of the ghost of his jilted sweetheart appearing to Sweet William of "Fair Margaret and Sweet William" as he lies in bed with his bride, it is rather the dead girl's image in a dream that kindles his fatal remorse. In addition to those ballads that turn on a supernatural occurrence, casual supernatural elements are found all through balladry.

Romantic tragedies. The separation of lovers through a misunderstanding or the opposition of relatives is perhaps the commonest ballad story. "Barbara Allen" is typical: Barbara cruelly spurns her lover because of an unintentional slight; he dies of lovesickness, she of remorse. The Freudian paradigm operates rigidly in ballads: fathers oppose the suitors of their daughters, mothers the sweethearts of their sons. Thus "The Douglas Tragedy"—the Danish "Ribold and Guldborg"—occurs when an eloping couple is overtaken by the girl's father and brothers or "Lady Maisry," pregnant by an English lord, is burned by her fanatically Scottish brother. Incest, frequent in ballads recorded before 1800 ("Lizie Wan," "The Bonny Hind"), is shunned by modern tradition.

The Freudian paradigm

Romantic comedies. The outcome of a ballad love affair is not always, though usually, tragic. But even when true love is eventually rewarded, such ballad heroines as "The Maid Freed from the Gallows" and "Fair Annie," among others, win through to happiness after such bitter trials that the price they pay seems too great. The course of romance runs hardly more smoothly in the many ballads, influenced by the cheap optimism of broadsides, where separated lovers meet without recognizing each other: the girl is told by the "stranger" of her lover's defection or death: her ensuing grief convinces him of her sincere love: he proves his identity and takes the joyful girl to wife. "The Bailiff's Daughter of Islington" is a classic of the type. Later tradition occasionally foists happy endings upon romantic tragedies: in the American "Douglas Tragedy" the lover is not slain but instead gets the irate father at his mercy and extorts a dowry from him. With marriage a consummation so eagerly sought in ballads, it is ironical that the bulk of humorous ballads deal with shrewish wives ("The Wife Wrapped in Wether's Skin") or gullible cuckolds ("Our Goodman").

Crime. Crime, and its punishment, is the theme of innumerable ballads: his sweetheart poisons "Lord Randal"; "Little Musgrave" is killed by Lord Barnard when he is discovered in bed with Lady Barnard, and the lady, too, is gorily dispatched. The murders of "Jim Fisk," Johnny of "Frankie and Johnny," and many other ballad victims are prompted by sexual jealousy. One particular variety of crime ballad, the "last goodnight," represents itself falsely to be the contrite speech of a criminal as he mounts the scaffold to be executed. A version of "Mary Hamilton" takes this form, which was a broadside device widely adopted by the folk. "Tom Dooley" and "Charles Guiteau," the scaffold confession of the assassin of Pres. James A. Garfield, are the best known American examples.

Medieval romance. Perhaps a dozen or so ballads derive from medieval romances. As in "Hind Horn" and "Thomas Rymer," only the climactic scene is excerpted for the ballad. In general, ballads from romances have not worn well in tradition because of their unpalatable fabulous elements, which the modern folk apparently regard as childish. Thus "Sir Lionel" becomes in America "Bangum and the Boar," a humorous piece to amuse children. Heterodox apocryphal legends that circulated widely in the Middle Ages are the source of almost all religious ballads, notable "Judas," "The Cherry-Tree Carol," and "The Bitter Withy." The distortion of biblical narrative is not peculiarly British: among others, the Russian bal-

lads of Samson and Solomon, the Spanish "Pilgrim to Compostela" and the French and Catalonian ballads on the penance of Mary Magdalence reshape canonical stories radically.

Historical ballads. Historical ballads date mainly from the period 1550–750, though a few, like "The Battle of Otterburn," celebrate events of an earlier date, in this case 1388. "The Hunting of the Cheviot," recorded about the same time and dealing with the same campaign, is better known in a late broadside version called "Chevy Chase." The details in historical ballads are usually incorrect as to fact because of faulty memory or partisan alterations, but they are valuable in reflecting folk attitudes toward the events they imperfectly report. For example, neither "The Death of Queen Jane," about one of the wives of Henry VIII, nor "The Bonny Earl of Murray" is correct in key details, but they accurately express the popular mourning for these figures. By far the largest number of ballads that can be traced to historical occurrences have to do with local skirmishes and matters of regional rather than national importance. The troubled border between England and Scotland in the 16th and early 17th century furnished opportunities for intrepid displays of loyalty, courage, and cruelty that are chronicled in such dramatic ballads as "Edom o Gordon," "The Fire of Frendraught," "Johnny Cock," "Johnie Armstrong," and "Hobie Noble." Closely analogous to these are Spanish *romances* such as "The Seven Princes of Lara," on wars between Moors and Christians.

Disaster. Sensational shipwrecks, plagues, train wrecks, mine explosions—all kinds of shocking acts of God and man—were regularly chronicled in ballads, a few of which remained in tradition, probably because of some special charm in the language or the music. The shipwreck that lies in the background of one of the most poetic of all ballads "Sir Patrick Spens" cannot be fixed, but "The Titanic," "Casey Jones," "The Wreck on the C & O," and "The Johnstown Flood" are all circumstantially based on actual events.

Epic and saga heroes

Outlaws and badmen. Epic and saga heroes figure prominently in Continental balladries, notable examples being the Russian Vladimir, the Spanish Cid Campeador, the Greek Digenes Akritas, and the Danish Tord of Havsgaard and Diderik. This kind of hero never appears in English and Scottish ballads. But the outlaw hero of the type of the Serbian Makro Kraljević or the Danish Marsk Stig is exactly matched by the English Robin Hood, who is the hero of some 40 ballads, most of them of minstrel or broadside provenance. His chivalrous style and generosity to the poor was imitated by later ballad highwaymen in "Dick Turpin," "Brennan on the Moor," and "Jesse James." "Henry Martyn" and "Captain Kidd" were popular pirate ballads, but the most widely sung was "The Flying Cloud," a contrite "goodnight" warning young men to avoid the curse of piracy. The fact that so many folk heroes are sadistic bullies ("Stagolee"), robbers ("Dupree"), or pathological killers ("Sam Bass," "Billy the Kid") comments on the folk's hostile attitude toward the church, constabulary, banks, and railroads. The kindly, law-abiding, devout, enduring steel driver "John Henry" is a rarity among ballad heroes.

Occupational ballads. A large section of balladry, especially American, deals with the hazards of such occupations as seafaring ("The Greenland Whale Fishery"), lumbering ("The Jam on Gerry's Rock"), mining ("The Avondale Mine Disaster"), herding cattle ("Little Joe the Wrangler"), and the hardships of frontier life ("The Arkansaw Traveler"). But men in these occupations sang ballads also that had nothing to do with their proper work: "The Streets of Laredo," for example, is known in lumberjack and soldier versions as well as the usual cowboy lament version, and the pirate ballad "The Flying Cloud" was much more popular in lumbermen's shanties than in forecastles.

CHRONOLOGY

Singing stories in song, either stories composed for the occasion out of a repertory of traditional motifs or phrases or stories preserved by memory and handed down orally, is found in most primitive cultures. The ballad habit thus is unquestionably very ancient. But the ballad genre itself could not have existed in anything like its present form before about 1100. "Judas," the oldest example found in Francis James Child's exhaustive collection, *The English and Scottish Popular Ballads* (1882–98), dates from 1300, but until the 17th century ballad records are sparse indeed. As an oral art, the ballad does not need to be written down to be performed or preserved; in any case, many of the carriers of the ballad tradition are illiterate and could not make use of a written and notated ballad. The few early ballads' records survived accidentally, due to some monk's, minstrel's, or antiquary's fascination with rustic pastimes.

Precise dating almost impossible

The precise date of a ballad, therefore, or even any particular version of a ballad, is almost impossible to determine. In fact, to ask for the date of a folk ballad is to show that one misunderstands the peculiar nature of balladry. As remarked earlier, the first recording of a ballad must not be assumed to be the ballad's original form; behind each recorded ballad can be one detected the working of tradition upon some earlier form, since a ballad does not become a ballad until it has run a course in tradition. Historical ballads would seem on the surface to be easily datable, but their origins are usually quite uncertain. The ballad could have arisen long after the events it describes, basing itself, as do the Russian ballads of the Kievan cycle and the Spanish ballads about the Cid, on chronicles or popular legends. It is also likely that many historical ballads developed from the revamping of earlier ballads on similar themes through the alteration of names, places, and local details. (A.B.F.)

Romance

Medieval romance developed in western Europe from the 12th century onward, and was extended up to and after the Renaissance, and reemerged in the 18th century.

The Old French word *romanz* originally meant "the speech of the people," or "the vulgar tongue," from a popular Latin word, *Romanice,* meaning written in the vernacular, in contrast with the written form of literary Latin. Its meaning then shifted from the language in which the work was written to the work itself. Thus, an adaptation of Geoffrey of Monmouth's *Historia regum Britanniae* (*c.* 1137), made by Wace of Jersey in 1155, was known as *Li Romanz de Brut,* while an anonymous adaptation (of slightly later date) of Virgil's *Aeneid* was known as *Li Romanz d'Enéas;* it is difficult to tell whether in such cases *li romanz* still meant "the French version" or had already come to mean "the story." It soon specialized in the latter sense, however, and was applied to narrative compositions similar in character to those imitated from Latin sources but totally different in origin; and, as the nature of these compositions changed, the word itself acquired an increasingly wide spectrum of meanings. In modern French a *roman* is just a novel, whatever its content and structure; while in modern English the word "romance" (derived from Old French *romanz*) can mean either a medieval narrative composition or a love affair, or, again, a story about a love affair, generally one of a rather idyllic or idealized type, sometimes marked by strange or unexpected incidents and developments; and "to romance" has come to mean "to make up a story that has no connection with reality."

For a proper understanding of these changes it is essential to know something of the history of the literary form to which, since the Middle Ages, the term has been applied. The account that follows is intended to elucidate historically some of the ways in which the word is used in English and in other European languages.

THE COMPONENT ELEMENTS

Greek romances

The romances of love, chivalry, and adventure produced in 12th-century France have analogues elsewhere, notably in what are sometimes known as the Greek romances—narrative works in prose by Greek writers from the 1st century BC to the 3rd century AD. The first known, the fragmentary Ninus romance, in telling the story of the love of Ninus, mythical founder of Nineveh, anticipates the

medieval *roman d'antiquité.* A number of works by writers of the 2nd and 3rd centuries AD—Chariton, Xenophon of Ephesus, Heliodorus, Achilles Tatius, and Longus—introduce a theme that was to reappear in the *roman d'aventure:* that of faithful lovers parted by accident or design and reunited only after numerous adventures. Direct connection, however, can be proved only in the case of the tale of *Apollonius of Tyre,* presumably deriving from a lost Greek original but known through a 3rd- or 4th-century Latin version. This too is a story of separation, adventure, and reunion, and, like the others (except for Longus' pastoral *Daphnis and Chloë*), it has a quasi-historical setting. It became one of the most popular and widespread stories in European literature during the Middle Ages and later provided Shakespeare with the theme of *Pericles.*

Style and subject matter. But the real debt of 12th-century romance to classical antiquity was incurred in a sphere outside that of subject matter. During the present century, scholars have laid ever-increasing emphasis on the impact of late classical antiquity upon the culture of medieval Europe, especially on that of medieval France. In particular, it is necessary to note the place that rhetoric (the systematic study of oratory) had assumed in the educational system of the late Roman Empire. Originally conceived as part of the training for public speaking, essential for the lawyer and politician, it had by this time become a literary exercise, the art of adorning or expanding a set theme: combined with grammar and enshrined in the educational system inherited by the Christian Church, rhetoric became an important factor in the birth of romance. Twelfth-century romance was, at the outset, the creation of "clerks"—professional writers who had been trained in grammar (that is to say, the study of the Latin language and the interpretation of Latin authors) and in rhetoric in the cathedral schools. They were skilled in the art of exposition, by which a subject matter was not only developed systematically but also given such meaning as the author thought appropriate. The "romance style" was, apparently, first used by the authors of three *romans d'antiquité,* all composed in the period 1150–65: *Roman de Thebes,* an adaptation of the epic *Thebaïs* by the late Latin poet Statius; *Roman d'Enéas,* adapted from Virgil's *Aeneid;* and *Roman de Troie,* a retelling by Benoît de Sainte-Maure of the tale of Troy, based not on Homer (who was not known in western Europe, where Greek was not normally read) but on 4th- and 5th-century Latin versions. In all three, style and subject matter are closely interconnected; elaborate set descriptions, in which the various features of what is described are gone through, item by item, and eulogized, result in the action's taking place in lavish surroundings, resplendent with gold, silver, marble, fine textiles, and precious stones. To these embellishments are added astonishing works of architecture and quaint technological marvels, that recall the Seven Wonders of the World and the reputed glories of Byzantium. *Troie* and *Enéas* have, moreover, a strong love interest, inspired by the Roman poet Ovid's conception of love as a restless malady. This concept produced the first portrayal in Western literature of the doubts, hesitations, and self-torment of young lovers, as exemplified in the Achilles–Polyxena story in *Troie* and in the Aeneas–Lavinia story in *Enéas.* Yet even more important is the way in which this new theme is introduced: the rhetorical devices appropriate to expounding an argument are here employed to allow a character in love to explore his own feelings, to describe his attitude to the loved one, and to explain whatever action he is about to take.

Earliest works in the "romance style"

Developing psychological awareness. As W.P. Ker, a pioneer in the study of medieval epic and romance, observed in his *Epic and Romance* (1897), the advent of romance is "something as momentous and as far-reaching as that to which the name Renaissance is generally applied." The Old French poets who composed the chansons de geste (as the Old French epics are called) had been content to tell a story; they were concerned with statement, not with motivation, and their characters could act without explicitly justifying their actions. Thus, in what is one of the earliest and certainly the finest of the chansons de geste, the *Chanson de Roland* (*c.* 1100), the hero's decision to fight on against odds—to let the rear guard of Charlemagne's army be destroyed by the Saracen hordes in the hopeless and heroic Battle of Roncesvalles rather than sound his horn to call back Charlemagne—is not treated as a matter for discussion and analysis: the anonymous poet seems to take it for granted that the reader is not primarily concerned with the reason why things happened as they did. The new techniques of elucidating and elaborating material, developed by romance writers in the 12th century, produced a method whereby actions, motives, states of mind, were scrutinized and debated. The story of how Troilus fell in love with Briseïs and how, when taken to the Grecian camp, she deserted him for Diomedes (as related, and presumably invented, by Benoît de Sainte-Maure in his *Roman de Troie*) is not one of marvellous adventures in some exotic fairyland setting: it is clearly a theme of considerable psychological interest, and it was for this reason that it attracted three of the greatest writers of all time: Boccaccio in his *Filostrato* (*c.* 1338), Chaucer in his *Troilus and Criseyde* (before 1385), and Shakespeare in his *Troilus and Cressida* (*c.* 1601–02). With the 12th-century pioneers of what came to be called romance, the beginnings of the analytical method found in the modern novel can easily be recognized.

The development of more sophisticated techniques

Sources and parallels. Where exactly medieval romance writers found their material when they were not simply copying classical or pseudo-classical models is still a highly controversial issue. Parallels to certain famous stories, such as that of Tristan and Iseult, have been found in regions as wide apart as Persia and Ireland: in the mid-11th-century Persian epic of *Wis and Ramin* and in the Old Irish *Diarmaid and Gráinne;* but while in the latter case it is possible to argue in favour of a genetic link between the two traditions, the former is more likely to be a case of parallel development due, on the one hand, to the inner logic of the theme and, on the other, to certain similarities in the ideological and social background of the two works. Failure to maintain the essential distinction between source and parallel has greatly hindered the understanding of the true nature of medieval romance and has led to the production of a vast critical literature the relevance of which to the study of the genre is at best questionable.

The marvellous. The marvellous is by no means an essential ingredient of "romance" in the sense in which it has been defined. Yet to most English readers the term romance does carry implications of the wonderful, the miraculous, the exaggerated, and the wholly ideal. Ker regarded much of the literature of the Middle Ages as "romantic" in this sense—the only types of narrative free from such "romanticizing" tendencies being the historical and family narrative, or Icelanders' sagas developed in classical Icelandic literature at the end of the 12th and in the early 13th century. The *Chanson de Roland* indulges freely in the fantastic and the unreal: hence Charlemagne's patriarchal age and preternatural strength (he is more than 200 years old when he conquers Spain); or the colossal numbers of those slain by the French; or, again, the monstrous races of men following the Saracen banners. Pious legends, saints' lives, and stories of such apocryphal adventures as those of the Irish St. Brendan (*c.* 486–578) who, as hero of a legend first written down in the 9th-century, *Navigatio Brendani,* and later widely translated and adapted, wanders among strange islands on his way to the earthly paradise—these likewise favour the marvellous. The great 12th-century *Roman d'Alexandre,* a *roman d'antiquité* based on and developing the early Greek romance of Alexander the Great (the Alexander romance), was begun in the first years of the century by Alberic de Briançon and later continued by other poets. It introduces fantastic elements, more especially technological wonders and the marvels of India: the springs of rejuvenation, the flower-maidens growing in a forest, the cynocephali (dog-headed men), the bathyscaphe that takes Alexander to the bottom of the ocean, and the car in which he is drawn through the air by griffins on his celestial journey.

The setting. The fact that so many medieval romances are set in distant times and remote places is not an essential feature of romance but rather a reflection of its

origins. As has been seen, the Old French word *romanz* early came to mean "historical work in the vernacular." All the *romans d'antiquité* have a historical or pseudo-historical theme, whether they evoke Greece, Troy or the legendary world of Alexander; but while making some attempt to give antiquity an exotic aspect by means of marvels or technological wonders, medieval writers were quite unable to create a convincing historical setting; and thus in all important matters of social life and organization they projected the western European world of the 12th century back into the past. Similarly, historical and contemporary geography were not kept separate. The result is often a confused jumble, as, for example, in the Anglo-Norman Hue de Rotelande's *Protesilaus,* in which the characters have Greek names; the action takes place in Burgundy, Crete, Calabria, and Apulia; and Theseus is described as "king of Denmark." This lavish use of exotic personal and geographical names and a certain irresponsibility about settings was still to be found in some of Shakespeare's romantic comedies: the "seacoast of Bohemia" in *The Winter's Tale* is thoroughly medieval in its antecedents. For in the medieval period, myth and folktale and straightforward fact were on an equal footing. Not that any marvel or preternatural happening taking place in secular (as opposed to biblical) history was necessarily to be believed: it was simply that the remote times and regions were convenient locations for picturesque and marvellous incidents. It is, indeed, at precisely this point that the transition begins from the concept of romance as "past history in the vernacular" to that of "a wholly fictitious story."

Errors of history and geography in early romance

MEDIEVAL VERSE ROMANCES

Arthurian romance. *The matter of Britain.* In his *Historia regum Britanniae,* Geoffrey of Monmouth "invented history" by drawing on classical authors, the Bible, and Celtic tradition to create the story of a British kingdom, to some extent paralleling that of Israel. He described the rise of the British people to glory in the reigns of Uther Pendragon and Arthur, then the decline and final destruction of the kingdom, with the exile of the British survivors and their last king, Cadwalader. Romances that have Arthur or some of his knights as main characters were classified as *matière de Bretagne* by Jehan Bodel (flourished 1200) in a well-known poem. There is in this "matter of Britain" a certain amount of material ultimately based on the belief, probably Celtic in origin, in an otherworld into which humans can penetrate, where they can challenge those who inhabit it, or enjoy the love of fairy women. Such themes appear in a highly rationalized form in the lays (*lais*) of the late 12th-century Marie de France, although she mentions Arthur and his queen only in one, the lay of *Lanval.*

Chrétien de Troyes. But it was Chrétien de Troyes (flourished 1165–80) who in five romances (*Erec; Cligès; Lancelot, ou Le Chevalier de la charrette; Yvain ou Le Chevalier au lion;* and *Perceval ou li Conte du Graal*) fashioned a new type of narrative based on the matter of Britain. The internal debate and self-analysis of the *roman d'antiquité* is here used with great artistry. At times, what seems to matter most to the poet is not the plot but the thematic pattern he imposes upon it and the significance he succeeds in conveying, either in individual scenes in which the action is interpreted by the characters in long monologues or through the work as a whole. In addition to this, he attempts what he himself calls a *conjointure*—that is, the organization into a coherent whole of a series of episodes. The adventures begin and end at the court of King Arthur; but the marvels that bring together material from a number of sources are not always meant to be believed, especially as they are somehow dovetailed into the normal incidents of life at a feudal court. Whatever Chrétien's intentions may have been, he inaugurated what may be called a Latin tradition of romance—clear, hard, bright, adorned with rhetoric, in which neither the courtly sentiment nor the enchantments are seriously meant. Chrétien had only one faithful follower, the trouvère Raoul de Houdenc (flourished 1200–30), author of *Méraugis de Portlesguez.* He shared Chrétien's taste for love casuistry, rhetorical adornment, and fantastic adventure. For both of these authors elements of rhetoric and self-analysis remain important, although the dose of rhetoric varies from one romance to another. Even in Chrétien's *Perceval ou Li Conte du Graal* ("Perceval, or the Romance of the Grail")—the work in which the Grail appears for the first time in European literature—the stress is on narrative incident interspersed with predictions of future happenings and retrospective explanations. Arthurian romances of the period 1170–1250 are *romans d'aventure,* exploiting the element of the strange, the supernatural, and the magical in the Arthurian tradition. A number (for example, *La Mule sans frein* ["The Mule Without a Bridle"], *c.* 1200; *L'Âtre périlleux* ["The Perilous Churchyard"], *c.* 1250) have as their hero Arthur's nephew Gawain, who, in the earlier Arthurian verse romances, is a type of the ideal knight.

Introduction of the Grail theme into European literature

Love as a major theme. The treatment of love varies greatly from one romance to another. It is helpful to distinguish sharply here between two kinds of theme: the one, whether borrowed from classical antiquity (such as the story of Hero and Leander, or that of Pyramus and Thisbe, taken from Ovid's *Metamorphoses*) or of much more recent origin, ending tragically; the other ending with marriage, reconciliation, or the reunion of separated lovers. It is noteworthy that "romance," as applied to a love affair in real life, has in modern English the connotation of a happy ending. This is also true of most Old French love romances in verse: the tragic ending is rare and is usually linked with the theme of the lover who, finding his or her partner dead, joins the beloved in death, either by suicide or from grief.

The Tristan story. The greatest tragic love story found as a romance theme is that of Tristan and Iseult. It was given the form in which it has become known to succeeding generations in about 1150–60 by an otherwise unknown Old French poet whose work, although lost, can be reconstructed in its essentials from surviving early versions based upon it. Probably closest in spirit to the original is the fragmentary version of *c.* 1170–90 by the Norman poet Béroul. From this it can be inferred that the archetypal poem told the story of an all-absorbing passion caused by a magic potion, a passion stronger than death, yet unable to triumph over the feudal order to which the heroes belong. The story ended with Iseult's death in the embrace of her dying lover and with the symbol of two trees growing from the graves of the lovers and intertwining their branches so closely that they could never be separated. Most later versions, including a courtly version by an Anglo-Norman poet known only as Thomas, attempt to resolve the tragic conflict in favour of the sovereignty of passion and to turn the magic potion into a mere symbol. Gottfried von Strassburg's German version, *Tristan und Isolde* (*c.* 1210), based on Thomas, is one of the great courtly romances of the Middle Ages; but although love is set up as the supreme value and as the object of the lovers' worship, the mellifluous and limpid verse translates the story into the idyllic mode. Another tragic and somewhat unreal story is that told in the anonymous *Chastelaine de Vergi* (*c.* 1250), one of the gems of medieval poetry, in which the heroine dies of grief because, under pressure, her lover has revealed their secret and adulterous love to the duke of Burgundy. The latter tells it to his own wife, who allows the heroine to think that her lover has betrayed her. The theme of the dead lover's heart served up by the jealous husband to the lady—tragic, sophisticated, and far-fetched—appears in the anonymous *Chastelain de Couci* (*c.* 1280), and again in *Daz Herzmaere* by the late 13th-century German poet Konrad von Würzburg. The theme of the outwitting of the jealous husband, common in the fabliaux (short verse tales containing realistic, even coarse detail, and written to amuse), is frequently found in 13th-century romance and in lighter lyric verse. It occurs both in the *Chastelain de Couci* and in the Provençal romance, *Flamenca* (*c.* 1234), in which it is treated comically.

Later versions of the Tristan and Iseult theme

The theme of separation and reunion. But the theme that has left the deepest impress on romance is that of a happy resolution, after many trials and manifold dangers, of lovers' difficulties. As has been seen, this theme was derived from late classical Greek romance by way of *Apollonius of Tyre* and its numerous translations and variants.

A somewhat similar theme, used for pious edification, is that of the legendary St. Eustace, reputedly a high officer under the Roman emperor Trajan, who lost his position, property, and family only to regain them after many tribulations, trials, and dangers. The St. Eustace theme appears in *Guillaume d'Angleterre,* a pious tale rather than a romance proper, which some have attributed to Chrétien de Troyes.

The Floire and Blancheflor romance

A variant on the theme of separation and reunion is found in the romance of *Floire et Blancheflor* (*c.* 1170), in which Floire, son of the Saracen "king" of Spain, is parted by his parents from Blancheflor, daughter of a Christian slave of noble birth, who is sold to foreign slave dealers. He traces her to a tower where maidens destined for the sultan's harem are kept, and the two are reunited when he gains access to her there by hiding in a basket of flowers. This romance was translated into Middle High German, Middle Dutch, Norse, and Middle English (as *Floris and Blancheflur, c.* 1250) and in the early 13th century was imitated in *Aucassin et Nicolette,* which is a *chantefable* (a story told in alternating sections of sung verse and recited prose) thought by some critics to share a common source with *Floire et Blancheflor.* In it, the roles and nationality, or religion, of the main characters are reversed; Nicolette, a Saracen slave converted to Christianity, who proves to be daughter of the king of Carthage, disguises herself as a minstrel in order to return to Aucassin, son of Count Gavin of Beaucaire. Jean Renart's *L'Escoufle* (*c.* 1200–02) uses the theme of lovers who, accidentally separated while fleeing together from the emperor's court, are eventually reunited; and the highly esteemed and influential *Guillaume de Palerne* (*c.* 1200) combines the theme of escaping lovers with that of the "grateful animal" (here a werewolf, which later resumes human shape as a king's son) assisting the lovers in their successful flight. The popular *Partenopeus de Blois* (*c.* 1180), of which 10 French manuscripts and many translated versions are known, resembles the Cupid and Psyche story told in the Roman writer Apuleius' *Golden Ass* (2nd century AD), although there is probably no direct connection. In the early 13th-century *Galeran de Bretagne,* Galeran loves Fresne, a foundling brought up in a convent; the correspondence between the two is discovered, and Fresne is sent away but appears in Galeran's land just in time to prevent him from marrying her twin sister, Fleurie.

The "Imogen theme"

The theme of a knight who undertakes adventures to prove to his lady that he is worthy of her love is represented by a variety of romances including the *Ipomedon* (1174–90) of Hue de Rotelande and the anonymous mid-13th-century Anglo-Norman *Gui de Warewic.* Finally, there are many examples of the "persecuted heroine" theme; in one variety a person having knowledge of some "corporal sign"—a birthmark or mole—on a lady wagers with her husband that he will seduce her and offer proof that he has done so (this is sometimes called the "Imogen theme" from its use in Shakespeare's *Cymbeline*). The deceit is finally exposed and the lady's honour vindicated. In the early 13th-century *Guillaume de Dôle* by Jean Renart, the birthmark is a rose; and in the *Roman de Violette,* written after 1225 by Gerbert de Montreuil, it is a violet. Philippe de Beaumanoir's *La Manekine* (*c.* 1270), Jean Maillart's *La Contesse d'Anjou* (1361), and Chaucer's *Man of Law's Tale* (after 1387) all treat the theme of the tribulations of a wife falsely accused and banished but, after many adventures, reunited with her husband.

MEDIEVAL PROSE ROMANCES

Arthurian themes. The Arthurian prose romances arose out of the attempt, made first by Robert de Boron in the verse romances *Joseph d'Arimathie, ou le Roman de l'estoire dou Graal* and *Merlin* (*c.* 1190–1200), to combine the fictional history of the Holy Grail with the chronicle of the reign of King Arthur. Robert gave his story an allegorical meaning, related to the person and work of Christ. A severe condemnation of secular chivalry and courtly love characterize the Grail branch of the prose Lancelot-Grail, or Vulgate, cycle as well as some parts of the post-Vulgate "romance of the Grail" (after 1225); in the one case, Lancelot (here representing fallen human nature) and, in the other, Balain (who strikes the Dolorous Stroke) are contrasted with Galahad, a type of the Redeemer. The conflict between earthly chivalry and the demands of religion is absent from the *Perlesvaus* (after 1230?), in which the hero Perlesvaus (that is, Perceval) has Christological overtones and in which the task of knighthood is to uphold and advance Christianity. A 13th-century prose *Tristan* (*Tristan de Léonois*), fundamentally an adaptation of the Tristan story to an Arthurian setting, complicates the love theme of the original with the theme of a love rivalry between Tristan and the converted Saracen Palamède and represents the action as a conflict between the treacherous villain King Mark and the "good" knight Tristan.

The ideals of chivalry

In the 14th century, when chivalry enjoyed a new vogue as a social ideal and the great orders of secular chivalry were founded, the romance writers, to judge from what is known of the voluminous *Perceforest* (written *c.* 1330 and still unpublished in its entirety), evolved an acceptable compromise between the knight's duty to his king, to his lady, and to God. Chivalry as an exalted ideal of conduct finds its highest expression in the anonymous Middle English *Sir Gawayne and the Grene Knight* (*c.* 1370), whose fantastic beheading scene (presumably taken from a lost French prose romance source) is made to illustrate the fidelity to the pledged word, the trust in God, and the unshakable courage that should characterize the knight.

The special structure of the Lancelot-Grail romance

Structure. The Vulgate *Lancelot-Grail* cycle displays a peculiar technique of interweaving that enables the author (or authors) to bring together a large number of originally independent themes. The story of Lancelot, of Arthur's kingdom, and the coming of Galahad (Lancelot's son) are all interconnected by the device of episodes that diverge, subdivide, join, and separate again, so that the work is a kind of interlocking whole, devoid of unity in the modern sense but forming as impregnable a structure as any revolving around a single centre. One of its most important features is its capacity for absorbing contrasting themes, such as the story of Lancelot's love for Guinevere, Arthur's queen, and the Quest of the Grail; another feature is its ability to grow through continuations or elaborations of earlier themes insufficiently developed. The great proliferation of prose romances at the end of the Middle Ages would have been impossible without this peculiarity of structure. Unlike any work that is wholly true to the Aristotelian principle of indivisibility and isolation (or organic unity), the prose romances satisfy the first condition, but not the second: internal cohesion goes with a tendency to seek connections with other similar compositions and to absorb an increasingly vast number of new themes. Thus the prose *Tristan* brings together the stories of Tristan and Iseult, the rise and fall of Arthur's kingdom, and the Grail Quest. It early gave rise to an offshoot, the romance of *Palamède* (before 1240), which deals with the older generation of Arthur's knights. A similar example of "extension backward" is the *Perceforest,* which associates the beginnings of knighthood in Britain with both Brutus the Trojan (reputedly Aeneas' grandson and the legendary founder of Britain) and Alexander the Great and makes its hero, Perceforest, live long before the Christian era.

LATER DEVELOPMENTS

The Arthurian prose romances were influential in both Italy and Spain; and this favoured the development in these countries of works best described as *romans d'aventure,* with their constantly growing interest in tournaments, enchantments, single combat between knights, love intrigues, and rambling adventures. In Italy, early prose compilations of Old French epic material from the Charlemagne cycle were subsequently assimilated to the other great bodies of medieval French narrative fiction and infused with the spirit of Arthurian prose romance. The great Italian heroic and romantic epics, Matteo Boiardo's *Orlando innamorato* (1483) and Ludovico Ariosto's *Orlando furioso* (1516), are based on this fusion. The serious themes of the Holy Grail and death of Arthur left no mark in Italy. The romantic idealism of Boiardo and Ariosto exploits instead the worldly adventures and the love sentiment of Arthurian prose romance, recounted lightly and with a sophisticated humour.

The Spanish romance

In Spain the significant development is the appearance, as early as the 14th, or even the 13th, century, of a native prose romance, the *Amadís de Gaula.* Arthurian in spirit but not in setting and with a freely invented episodic content, this work, in the form given to it by Garci Rodríguez de Montalvo in its first known edition of 1508, captured the imagination of the polite society of western Europe by its blend of heroic and incredible feats of arms and tender sentiment and by its exaltation of an idealized and refined concept of chivalry. Quickly translated and adapted into French, Italian, Dutch, and English and followed by numerous sequels and imitations in Spanish and Portuguese, it remained influential for more than four centuries, greatly affecting the outlook and sensibility of western society. Cervantes parodied the fashion inspired by *Amadís* in *Don Quixote* (1605); but his admiration for the work itself caused him to introduce many of its features into his own masterpiece, so that the spirit and the character of chivalric romance may be said to have entered into the first great modern novel.

Parallels with sculpture and painting

More important still for the development of the novel form was the use made by romance writers of the technique of multiple thematic structure and "interweaving" earlier mentioned. Like the great examples of Romanesque ornamental art, both sculptural and pictorial, the cyclic romances of the late Middle Ages, while showing a strong sense of cohesion, bear no trace whatever of the classical concept of subordination to a single theme: an excellent proof, if proof were needed, of the limited relevance of this concept in literary aesthetics. Even those romances which, like the *Amadís* and its ancestor, the French prose *Lancelot,* had one great figure as the centre of action, cannot be said to have progressed in any way toward the notion of the unity of theme.

The spread and popularity of romance literature. This is as true of medieval romances as of their descendants, including the French and the English 18th-century novel and the pastoral romance, which, at the time of the Renaissance, revived the classical traditions of pastoral poetry and led to the appearance, in 1504, of the *Arcadia* by the Italian poet Jacopo Sannazzaro and, in about 1559, of the *Diana* by the Spanish poet and novelist Jorge de Montemayor. Both works were widely influential in translation, and each has claims to be regarded as the first pastoral romance, but in spirit *Diana* is the true inheritor of the romance tradition, giving it, in alliance with the pastoral, a new impetus and direction.

The social milieu

Medieval romance began in the 12th century when clerks, working for aristocratic patrons, often ladies of royal birth such as Eleanor of Aquitaine and her daughters, Marie de Champagne and Matilda, wife of Henry the Lion, duke of Saxony, began to write for a leisured and refined society. Like the courtly lyric, romance was a vehicle of a new aristocratic culture which, based in France, spread to other parts of western Europe. Translations and adaptations of French romances appear early in German: the *Roman d'Enéas,* in a version written by Heinrich von Veldeke before 1186, and the archetypal Tristan romance in Eilhart von Oberge's *Tristant* of *c.* 1170–80. In England many French romances were adapted, sometimes very freely, into English verse and prose from the late 13th to the 15th century; but by far the most important English contribution to the development and popularization of romance was the adaptation of a number of French Arthurian romances completed by Sir Thomas Malory in 1469–70 and published in 1485 by William Caxton under the title of *Le Morte Darthur.* In the Scandinavian countries the connection with the Angevin rulers of England led to importation of French romances in the reign (1217–63) of Haakon of Norway.

The decline of romance. As has been seen, in the later Middle Ages the prose romances were influential in France, Italy, and Spain, as well as in England; and the advent of the printed book made them available to a still wider audience. But although they continued in vogue into the 16th century, with the spread of the ideals of the New Learning, the greater range and depth of vernacular literature, and the rise of the neoclassical critics, the essentially medieval image of the perfect knight was bound to change into that of the scholar-courtier, who, as presented by the Italian Baldassare Castiglione in his *Il Cortegiano* (published 1528), embodies the highest moral ideals of the Renaissance. The new Spanish romances continued to enjoy international popularity until well into the 17th century and in France gave rise to compendious sentimental romances with an adventurous, pastoral, or pseudo-historical colouring popular with Parisian *salon* society until *c.* 1660. But the French intellectual climate, especially after the beginning of the so-called classical period in the 1660s, was unfavourable to the success of romance as a "noble" genre. Before disappearing, however, the romances lent the French form of their name to such *romans* as Antoine Furetière's *Le Roman bourgeois* (1666) and Paul Scarron's *Le Roman comique* (1651–57). These preserved something of the outward form of romance but little of its spirit; and while they transmitted the name to the kind of narrative fiction that succeeded them, they were in no sense intermediaries between its old and its new connotations. The great critical issue dominating the thought of western Europe from about 1660 onward was that of "truth" in literature; and romance, as being "unnatural" and unreasonable, was condemned. Only in England and Germany did it find a home with poets and novelists. Thus, while Robert Boyle, the natural philosopher, in his *Occasional Discourses* (1666) was inveighing against gentlemen whose libraries contained nothing more substantial than "romances," Milton, in *Paradise Lost,* could still invoke "what resounds/In fable or romance of Uther's son . . ."

The search for "truth" in literature

The 18th-century romantic revival. The 18th century in both England and Germany saw a strong reaction against the rationalistic canons of French classicism—a reaction that found its positive counterpart in such romantic material as had survived from medieval times. The Gothic romances, of which Horace Walpole's *Castle of Otranto* (1764; dated 1765) is the most famous, are perhaps of less importance than the ideas underlying the defense of romance by Richard Hurd in his *Letters on Chivalry and Romance* (1762). To Hurd, romance is not truth but a delightful and necessary holiday from common sense. This definition of romance (to which both Ariosto and Chrétien de Troyes would no doubt have subscribed) inspired on the one hand the romantic epic *Oberon* (1780) and on the other the historical romances of Sir Walter Scott. But influential though Scott's romantic novels may have been in every corner of Europe (including the Latin countries), it was the German and English Romantics who, with a richer theory of the imagination than Hurd's, were able to recapture something of the spirit and the structure of romance—the German Romantics by turning to their own medieval past; the English, by turning to the tradition perpetuated by Edmund Spenser and Shakespeare.

(E.Vi./F.Wh.)

Saga

In medieval Iceland the literary term saga denoted any kind of story or history in prose, irrespective of the kind or nature of the narrative or the purposes for which it was written. Used in this general sense, the term applies to a wide range of literary works, including those of hagiography (biography of saints), historiography, and secular fiction in a variety of modes. Lives of the saints and other stories for edification are entitled sagas, as are the Norse versions of French romances and the Icelandic adaptations of various Latin histories. Chronicles and other factual records of the history of Scandinavia and Iceland down to the 14th century are also included under the blanket term saga literature. In a stricter sense, however, the term saga is confined to legendary and historical fictions, in which the author has attempted an imaginative reconstruction of the past and organized the subject matter according to certain aesthetic principles. Using the distinctive features of the hero as principal guideline, medieval Icelandic narrative fiction can be classified as: (1) kings' sagas, (2) legendary sagas, and (3) sagas of Icelanders.

The origin and evolution of saga writing in Iceland are largely matters for speculation. A common pastime on

Icelandic farms, from the 12th century down to modern times, was the reading aloud of stories to entertain the household, known as *sagnaskemmtun* ("saga entertainment"). It seems to have replaced the traditional art of storytelling. All kinds of written narratives were used in *sagnaskemmtun:* secular, sacred, historical, and legendary. The Icelandic church took a sympathetic view of the writing and reading of sagas, and many of the authors whose identity is still known were monks or priests.

NONFICTIONAL SAGA LITERATURE

Translations. European narratives were known in Iceland in the 12th and 13th centuries and undoubtedly served as models for Icelandic writers when they set out to form a coherent picture of early Scandinavian history. Translations of lives of the saints and the apostles and accounts of the Holy Virgin testify to the skill of Icelandic prose writers in handling the vernacular for narrative purposes from the 12th century onward. Histories were also adapted and translated from Latin, based on those of the 7th- and 8th-century Anglo-Saxon writer Bede, the 7th-century Spanish historian Isidore of Seville, and others; on fictitious accounts of the Trojan wars, notably, one of the 5th century attributed to Dares Phrygius and one of the 4th century attributed to Dictys Cretensis; on the 12th-century British chronicler Geoffrey of Monmouth; and on the 1st-century Roman historians Sallust and Lucan. In the 13th century, Abbot Brandr Jónsson wrote a history of the Jews based on the Vulgate, on the 10th-century biblical scholar Peter Comestor, and on other sources.

In the 13th century, saga literature was also enriched by Norwegian prose translations of French romance literature. These soon found their way into Iceland, where they were popular and a strong influence on native storywriting. Probably the earliest, *Tristrams saga* (the story of Tristan and Iseult), was translated in 1226. Most of the themes of French romance appear in Icelandic versions; *e.g., Karlamagnús saga* was based on Charlemagne legends.

Themes of French romance literature

Native historical accounts. Icelandic historians seem to have started writing about their country's past toward the end of the 11th century. Saemundr Sigfússon, trained as a priest in France, wrote a Latin history of the kings of Norway, now lost but referred to by later authors. The first Icelander to use the vernacular for historical accounts was Ari Thorgilsson, whose *Íslendingabók* (or *Libellus Islandorum* [*The Book of the Icelanders*]) survives. It is a concise description of the course of Icelandic history from the beginning of the settlement (*c.* 870) to 1118. Ari seems to have written this book about 1125, but before that date he may already have compiled (in collaboration with Kolskeggr Ásbjarnarson) the so-called *Landnámabók* ("Book of Settlements"), which lists the names and land claims of about 400 settlers. Because this work survives only in 13th- and 14th-century versions, it is impossible to tell how much of it is Ari's. Both books gave the Icelanders a clear picture of the beginning of their society; both works served to stimulate public interest in the period during which events recounted in the sagas of Icelanders (see below) are supposed to have taken place. Other factual accounts of the history of Iceland followed later: *Kristni saga* describes Iceland's conversion to Christianity about the end of the 10th century and the emergence of a national church. *Hungrvaka* ("The Appetizer") contains accounts of the lives of the first five bishops of Skálholt, from the mid-11th to the third quarter of the 12th century; the biographies of other prominent bishops are in the *Biskupa sögur.* Though some of these have a strong hagiographical flavour, others are soberly written and of great historical value. The period *c.* 1100–1264 is also dealt with in several secular histories, known collectively as *Sturlunga saga,* the most important of which is the *Íslendinga saga* ("The Icelanders' Saga") of Sturla Thórdarson, who describes in memorable detail the bitter personal and political feuds that marked the final episode in the history of the Icelandic commonwealth (*c.* 1200–64).

LEGENDARY AND HISTORICAL FICTION

Kings' sagas. After Saemundr Sigfússon, Icelandic and Norwegian authors continued to explore the history of Scandinavia in terms of rulers and royal families, some of them writing in Latin and others in the vernacular. Broadly speaking, the kings' sagas fall into two distinct groups: contemporary (or near contemporary) biographies and histories of remoter periods. To the first group belonged a now-lost work, written in about 1170 by an Icelander called Eiríkr Oddsson, dealing with several 12th-century kings of Norway. *Sverris saga* describes the life of King Sverrir (reigned 1184–1202). The first part was written by Abbot Karl Jónsson under the supervision of the King himself, but it was completed (probably by the Abbot) in Iceland after Sverrir's death. Sturla Thórdarson wrote two royal biographies: *Hákonar saga* on King Haakon Haakonsson (*c.* 1204–63) and *Magnús saga* on his son and successor, Magnus VI Lawmender (Lagabøter; reigned 1263–80); of the latter only fragments survive. In writing these sagas Sturla used written documents as source material and, like Abbot Karl before him, he also relied on the accounts of eyewitnesses. Works on the history of the earlier kings of Norway include two Latin chronicles of Norwegian provenance, one of which was compiled *c.* 1180, and two vernacular histories, also written in Norway, the so-called *Ágrip* (*c.* 1190) and *Fagrskinna* (*c.* 1230). The Icelandic *Morkinskinna* (*c.* 1220) deals with the kings of Norway from 1047–1177; an outstanding feature of it is that it tells some brilliant stories of Icelandic poets and adventurers who visited the royal courts of Scandinavia.

Contemporary writings of remoter periods

The kings' sagas reached their zenith in the *Heimskringla,* or *Noregs konunga sögur* ("History of the Kings of Norway"), of Snorri Sturluson, which describes the history of the royal house of Norway from legendary times down to 1177. Snorri was a leading 13th-century Icelandic poet, who used as sources all the court poetry from the 9th century onward that was available to him. He also used many earlier histories of the kings of Norway and other written sources. *Heimskringla* is a supreme literary achievement that ranks Snorri Sturluson with the great writers of medieval Europe. He interpreted history in terms of personalities rather than politics, and many of his character portrayals are superbly drawn. Two of the early kings of Norway, Olaf Tryggvason (reigned 995–1000) and Olaf Haraldsson (Olaf the Saint; reigned 1015–30), received special attention from Icelandic antiquarians and authors. Only fragments of a 12th-century *Ólafs saga helga* ("St. Olaf's Saga") survive; a 13th-century biography of the same king by Styrmir Kárason is also largely lost. (Snorri Sturluson wrote a brilliant saga of St. Olaf, rejecting some of the grosser hagiographical elements in his sources; this work forms the central part of his *Heimskringla.*) About 1190 a Benedictine monk, Oddr Snorrason, wrote a Latin life of Olaf Tryggvason, of which an Icelandic version still survives. A brother in the same monastery, Gunnlaugr Leifsson, expanded this biography, and his work was incorporated into later versions of *Ólafs saga Tryggvasonar.* Closely related to the lives of the kings of Norway are *Færeyinga saga,* describing the resistance of Faeroese leaders to Norwegian interference during the first part of the 11th century, and *Orkneyinga saga,* dealing with the rulers of the earldom of Orkney from about 900 to the end of the 12th century. These two works were probably written about 1200. The history of the kings of Denmark from *c.* 940 to 1187 is told in *Knýtlinga saga.*

Legendary sagas. The learned men of medieval Iceland took great pride in their pagan past and copied traditional poems on mythological and legendary themes. In due course some of these narrative poems served as the basis for sagas in prose. In his *Edda* (probably written *c.* 1225), Snorri Sturluson tells several memorable stories, based on ancient mythological poems, about the old gods of the North, including such masterpieces as the tragic death of Balder and the comic tale of Thor's journey to giantland. Snorri's book also contains a summary of the legendary Nibelungen cycle. (A much fuller treatment of the same theme is to be found in *Völsunga saga* and *Thidriks saga,* the latter composed in Norway and based on German sources.) Other Icelandic stories based on early poetic tradition include *Heidreks saga; Hrólfs saga kraka,* which has a certain affinity with the Old English poem *Beowulf; Hálfs saga ok Hálfsrekka; Gautreks saga;* and *Ásmundar*

saga kappabana, which tells the same story as the Old High German *Hildebrandslied.* The term legendary sagas also covers a number of stories the antecedents and models of which are not exclusively native. These sagas are set in what might be called the legendary heroic age at one level and also vaguely in the more recent Viking age at the other, the action taking place in Scandinavia and other parts of the Viking world, from Russia to Ireland, but occasionally also in the world of myth and fantasy. It is mostly through valour and heroic exploits that the typical hero's personality is realized. He is, however, often a composite character, for some of his features are borrowed from a later and more refined ethos than that of early Scandinavia. He is in fact the synthesis of Viking ideals on the one hand and of codes of courtly chivalry on the other. Of individual stories the following are notable: *Egils saga ok Ásmundar,* which skillfully employs the flashback device; *Bósa saga ok Herrauds,* exceptional for its erotic elements; *Fridthjófs saga,* a romantic love story; *Hrólfs saga Gautrekssonar; Göngu-Hrólfs saga;* and *Halfadanar saga Eysteinssonar.* There are many more. The legendary sagas are essentially romantic literature, offering an idealized picture of the remote past, and many of them are strongly influenced by French romance literature. In these sagas the main emphasis is on a lively narrative, entertainment being their primary aim and function. Some of the themes in the legendary sagas are also treated in the *Gesta Danorum* of the 12th-century Danish historian Saxo Grammaticus, who states that some of his informants for the legendary history of Denmark were Icelanders.

Sagas of Icelanders. In the late 12th century, Icelandic authors began to fictionalize the early part of their history (*c.* 900–1050), and a new literary genre was born: the sagas of Icelanders. Whereas the ethos of the kings' sagas and of the legendary sagas is aristocratic and their principal heroes warlike leaders, the sagas of Icelanders describe characters who are essentially farmers or farmers' sons or at least people who were socially not far above the author's public, and their conduct and motivation are measurable in terms of the author's own ethos. These authors constantly aimed at geographic, social, and cultural verisimilitude; they made it their business to depict life in Iceland as they had experienced it or as they imagined it had actually been in the past. Though a good deal of the subject matter was evidently derived from oral tradition and thus of historical value for the period described, some of the best sagas are largely fictional; their relevance to the authors' own times mattered perhaps no less than their incidental information about the past. An important aim of this literature was to encourage people to attain a better understanding of their social environment and a truer knowledge of themselves through studying the real and imagined fates of their forbears. A spirit of humanism, sometimes coloured by a fatalistic heroic outlook, pervades the narrative. The edificatory role, however, was never allowed to get out of hand or dominate the literary art; giving aesthetic pleasure remained the saga writer's primary aim and duty.

Aim of the sagas of Icelanders

Nothing is known of the authorship of the sagas of Icelanders, and it has proved impossible to assign a definite date to many of them. It seems improbable that in their present form any of them could have been written before *c.* 1200. The period *c.* 1230–90 has been described as the golden age of saga writing because such masterpieces as *Egils saga, Víga-Glúms saga, Gísla saga, Eyrbyggja saga, Hrafnkels saga Freysgoda, Bandamanna saga, Hænsa-Thóris saga,* and *Njáls saga* appear to have been written during that time. Although a number of sagas date from the 14th century, only one, *Grettis saga,* can be ranked with the classical ones.

The sagas of Icelanders can be subdivided into several categories according to the social and ethical status of the principal heroes. In some, the hero is a poet who sets out from the rural society of his native land in search of fame and adventure to become the retainer of the king of Norway or some other foreign ruler. Another feature of these stories is that the hero is also a lover. To this group belong some of the early-13th-century sagas, including *Kormáks saga, Hallfredar saga,* and *Bjarnar saga Hítdaelakappa.* In *Gunnlaugs saga ormstungu,* which may have been written after the middle of the 13th century, the love theme is treated more romantically than in the others. *Fostbraeda saga* ("The Blood-Brothers' Saga") describes two contrasting heroes: one a poet and lover, the other a ruthless killer. *Egils saga* offers a brilliant study of a complex personality—a ruthless Viking who is also a sensitive poet, a rebel against authority from early childhood who ends his life as a defenseless, blind old man. In several sagas the hero becomes an outlaw fighting a hopeless battle against the social forces that have rejected him. To this group belong *Hardar saga ok Hólmverja* and *Droplaugarsona saga;* but the greatest of the outlaw sagas are *Gísla saga,* describing a man who murders his own brother-in-law and whose sister reveals his dark secret; and *Grettis saga,* which deals with a hero of great talents and courage who is constantly fighting against heavy odds and is treacherously slain by an unscrupulous enemy.

Most of the sagas of Icelanders, however, are concerned with people who are fully integrated members of society, either as ordinary farmers or as farmers who also act as chieftains. *Hrafnkels saga* describes a chieftain who murders his shepherd, is then tortured and humiliated for his crime, and finally takes cruel revenge on one of his tormentors. The hero who gives his name to *Hænsa-Thoris saga* is a man of humble background who makes money as a peddler and becomes a wealthy but unpopular landowner. His egotism creates trouble in the neighbourhood, and after he has set fire to one of the farmsteads, killing the farmer and the entire household, he is prosecuted and later put to death. *Ölkofra tháttr* (the term *tháttr* is often used for a short story) and *Bandamanna saga* ("The Confederates' Saga") satirize chieftains who fail in their duty to guard the integrity of the law and try to turn other people's mistakes into profit for themselves. The central plot in *Laxdæla saga* is a love triangle, in which the jealous heroine forces her husband to kill his best friend. *Eyrbyggja saga* describes a complex series of feuds between several interrelated families; *Hávardar saga* is about an old farmer who takes revenge on his son's killer, the local chieftain; *Víga-Glúms saga* tells of a ruthless chieftain who commits several killings and swears an ambiguous oath in order to cover his guilt; while *Vatnsdæla saga* is the story of a noble chieftain whose last act is to help his killer escape.

In the sagas of Icelanders justice, rather than courage, is often the primary virtue, as might be expected in a literature that places the success of an individual below the welfare of society at large. This theme is an underlying one in *Njáls saga,* the greatest of all the sagas. It is a story of great complexity and richness, with a host of brilliantly executed character portrayals and a profound understanding of human strengths and weaknesses. Its structure is highly complex, but at its core is the tragedy of an influential farmer and sage who devotes his life to a hopeless struggle against the destructive forces of society but ends it inexorably when his enemies set fire to his house, killing his wife and sons with him. (He.P.)

Theme of *Njáls saga*

Novel

The novel is a genre of fiction, and fiction may be defined as the art or craft of contriving, through the written word, representations of human life that instruct or divert or both. The various forms that fiction may take are best seen less as a number of separate categories than as a continuum or, more accurately, a cline, with some such brief form as the anecdote at one end of the scale and the longest conceivable novel at the other. When any piece of fiction is long enough to constitute a whole book, as opposed to a mere part of a book, then it may be said to have achieved novelhood. But this state admits of its own quantitative categories, so that a relatively brief novel may be termed a novella (or, if the insubstantiality of the content matches its brevity, a novelette), and a very long novel may overflow the banks of a single volume and become a *roman-fleuve,* or river novel. Length is very much one of the dimensions of the genre.

The term novel is a truncation of the Italian word *novella*

(from the plural of Latin *novellus,* a late variant of *novus,* meaning "new"), so that what is now, in most languages, a diminutive denotes historically the parent form. The *novella* was a kind of enlarged anecdote like those to be found in the 14th-century Italian classic Boccaccio's *Decameron,* each of which exemplifies the etymology well enough. The stories are little new things, novelties, freshly minted diversions, toys; they are not reworkings of known fables or myths, and they are lacking in weight and moral earnestness. It is to be noted that, despite the high example of novelists of the most profound seriousness, such as Tolstoy, Henry James, and Virginia Woolf, the term novel still, in some quarters, carries overtones of lightness and frivolity. And it is possible to descry a tendency to triviality in the form itself. The ode or symphony seems to possess an inner mechanism that protects it from aesthetic or moral corruption, but the novel can descend to shameful commercial depths of sentimentality or pornography. It is the purpose of this section to consider the novel not solely in terms of great art but also as an all-purpose medium catering for all the strata of literacy.

The novel and the epic compared

Such early ancient Roman fiction as Petronius' *Satyricon* of the 1st century AD and Lucius Apuleius' *Golden Ass* of the 2nd century contain many of the popular elements that distinguish the novel from its nobler born relative the epic poem. In the fictional works, the medium is prose, the events described are unheroic, the settings are streets and taverns, not battlefields and palaces. There is more low fornication than princely combat; the gods do not move the action; the dialogue is homely rather than aristocratic. It was, in fact, out of the need to find—in the period of Roman decline—a literary form that was anti-epic in both substance and language that the first prose fiction of Europe seems to have been conceived. The most memorable character in Petronius is a *nouveau riche* vulgarian; the hero of Lucius Apuleius is turned into a donkey; nothing less epic can well be imagined.

The medieval chivalric romance (from a popular Latin word, probably *Romanice,* meaning written in the vernacular, not in traditional Latin) restored a kind of epic view of man—though now as heroic Christian, not heroic pagan. At the same time, it bequeathed its name to the later genre of continental literature, the novel, which is known in French as *roman,* in Italian as *romanzo,* etc. (The English term romance, however, carries a pejorative connotation.) But that later genre achieved its first great flowering in Spain at the beginning of the 17th century in an antichivalric comic masterpiece—the *Don Quixote* of Cervantes, which, on a larger scale than the *Satyricon* or *The Golden Ass,* contains many of the elements that have been expected from prose fiction ever since. Novels have heroes, but not in any classical or medieval sense. As for the novelist, he must, in the words of the contemporary British-American W.H. Auden,

> Become the whole of boredom, subject to
> Vulgar complaints like love, among the Just
> Be just, among the Filthy filthy too,
> And in his own weak person, if he can,
> Must suffer dully all the wrongs of Man.

The novel attempts to assume those burdens of life that have no place in the epic poem and to see man as unheroic, unredeemed, imperfect, even absurd. This is why there is room among its practitioners for writers of hard-boiled detective thrillers such as the contemporary American Mickey Spillane or of sentimental melodramas such as the prolific 19th-century English novelist Mrs. Henry Wood, but not for one of the unremitting elevation of outlook of a John Milton.

The reader may also be interested in the analogous treatment of a comparable genre in the section *Short story* which follows. Critical approaches are discussed in the article the ARTS, CRITICISM OF THE; and the section *Literary criticism* below.

ELEMENTS

Plot. The novel is propelled through its hundred or thousand pages by a device known as the story or plot. This is frequently conceived by the novelist in very simple terms, a mere nucleus, a jotting on an old envelope: for example, Charles Dickens' *Christmas Carol* (1843) might have been conceived as "a misanthrope is reformed through certain magical visitations on Christmas Eve"; or Jane Austen's *Pride and Prejudice* (1813) as "a young couple destined to be married have first to overcome the barriers of pride and prejudice"; or Fyodor Dostoyevsky's *Crime and Punishment* (1866) as "a young man commits a crime and is slowly pursued in the direction of his punishment." The detailed working out of the nuclear idea requires much ingenuity, since the plot of one novel is expected to be somewhat different from that of another, and there are very few basic human situations for the novelist to draw upon. The dramatist may take his plot ready-made from fiction or biography—a form of theft sanctioned by Shakespeare—but the novelist has to produce what look like novelties.

Original and borrowed plots

The example of Shakespeare is a reminder that the ability to create an interesting plot, or even any plot at all, is not a prerequisite of the imaginative writer's craft. At the lowest level of fiction, plot need be no more than a string of stock devices for arousing stock responses of concern and excitement in the reader. The reader's interest may be captured at the outset by the promise of conflicts or mysteries or frustrations that will eventually be resolved, and he will gladly—so strong is his desire to be moved or entertained—suspend criticism of even the most trite modes of resolution. In the least sophisticated fiction, the knots to be untied are stringently physical, and the denouement often comes in a sort of triumphant violence. Serious fiction prefers its plots to be based on psychological situations, and its climaxes come in new states of awareness—chiefly self-knowledge—on the parts of the major characters.

Melodramatic plots, plots dependent on coincidence or improbability, are sometimes found in even the most elevated fiction; E.M. Forster's *Howards End* (1910) is an example of a classic British novel with such a plot. But the novelist is always faced with the problem of whether it is more important to represent the formlessness of real life (in which there are no beginnings and no ends and very few simple motives for action) or to construct an artifact as well balanced and economical as a table or chair; since he is an artist, the claims of art, or artifice, frequently prevail. But there are ways of constructing novels in which plot may play a desultory part or no part at all. The traditional picaresque novel—a novel with a rogue as its central character—like Alain Lesage's *Gil Blas* (1715) or Henry Fielding's *Tom Jones* (1749), depends for movement on a succession of chance incidents. In the works of Virginia Woolf, the consciousness of the characters, bounded by some poetic or symbolic device, sometimes provides all the fictional material. Marcel Proust's great *roman-fleuve, À la recherche du temps perdu* (1913–27; *Remembrance of Things Past*), has a metaphysical framework derived from the time theories of the philosopher Henri Bergson, and it moves toward a moment of truth that is intended to be literally a revelation of the nature of reality. Strictly, any scheme will do to hold a novel together—raw action, the hidden syllogism of the mystery story, prolonged solipsist contemplation—so long as the actualities or potentialities of human life are credibly expressed, with a consequent sense of illumination, or some lesser mode of artistic satisfaction, on the part of the reader.

Character. The inferior novelist tends to be preoccupied with plot; to the superior novelist the convolutions of the human personality, under the stress of artfully selected experience, are the chief fascination. Without character it was once accepted that there could be no fiction. In the period since World War II, the creators of what has come to be called the French *nouveau roman* (*i.e.,* new novel) have deliberately demoted the human element, claiming the right of objects and processes to the writer's and reader's prior attention. Thus, in books termed *chosiste* (literally "thing-ist"), they make the furniture of a room more important than its human incumbents. This may be seen as a transitory protest against the long predominance of character in the novel, but, even on the popular level, there have been indications that readers can be held by things as much as by characters. Henry James could be

Objects as "characters"

vague in *The Ambassadors* (1903) about the provenance of his chief character's wealth; if he wrote today he would have to give his readers a tour around the factory or estate. The popularity of much undistinguished but popular fiction has nothing to do with its wooden characters; it is machines, procedures, organizations that draw the reader. The success of Ian Fleming's British spy stories in the 1960s had much to do with their hero, James Bond's car, gun, and preferred way of mixing a martini.

But the true novelists remain creators of characters—prehuman, such as those in William Golding's *Inheritors* (1955); animal, as in Henry Williamson's *Tarka the Otter* (1927) or Jack London's *Call of the Wild* (1903); caricatures, as in much of Dickens; or complex and unpredictable entities, as in Tolstoy, Dostoyevsky, or Henry James. The reader may be prepared to tolerate the most wanton-seeming stylistic tricks and formal difficulties because of the intense interest of the central characters in novels as diverse as James Joyce's *Ulysses* (1922) and *Finnegans Wake* (1939) and Laurence Sterne's *Tristram Shandy* (1760–67).

Characters as symbols

It is the task of literary critics to create a value hierarchy of fictional character, placing the complexity of the Shakespearean view of man—as found in the novels of Tolstoy and Joseph Conrad—above creations that may be no more than simple personifications of some single characteristic, like some of those by Dickens. It frequently happens, however, that the common reader prefers surface simplicity—easily memorable cartoon figures like Dickens' never-despairing Mr. Micawber and devious Uriah Heep—to that wider view of personality, in which character seems to engulf the reader, subscribed to by the great novelists of France and Russia. The whole nature of human identity remains in doubt, and writers who voice that doubt—like the French exponents of the *nouveau roman* Alain Robbe-Grillet and Nathalie Sarraute, as well as many others—are in effect rejecting a purely romantic view of character. This view imposed the author's image of himself—the only human image he properly possessed—on the rest of the human world. For the unsophisticated reader of fiction, any created personage with a firm position in time–space and the most superficial parcel of behavioral (or even sartorial) attributes will be taken for a character. Though the critics may regard it as heretical, this tendency to accept a character is in conformity with the usages of real life. The average person has at least a suspicion of his own complexity and inconsistency of makeup, but he sees the rest of the world as composed of much simpler entities. The result is that novels whose characters are created out of the author's own introspection are frequently rejected as not "true to life." But both the higher and the lower orders of novel readers might agree in condemning a lack of memorability in the personages of a work of fiction, a failure on the part of the author to seem to add to the reader's stock of remembered friends and acquaintances. Characters that seem, on recollection, to have a life outside the bounds of the books that contain them are usually the ones that earn their creators the most regard. Depth of psychological penetration, the ability to make a character real as oneself, seems to be no primary criterion of fictional talent.

Scene, or setting. The makeup and behaviour of fictional characters depend on their environment quite as much as on the personal dynamic with which their author endows them: indeed, in Émile Zola, environment is of overriding importance, since he believed it determined character. The entire action of a novel is frequently determined by the locale in which it is set. Thus, Gustave Flaubert's *Madame Bovary* (1857) could hardly have been placed in Paris, because the tragic life and death of the heroine have a great deal to do with the circumscriptions of her provincial milieu. But it sometimes happens that the main locale of a novel assumes an importance in the reader's imagination comparable to that of the characters and yet somehow separable from them. Wessex is a giant brooding presence in Thomas Hardy's novels, whose human characters would probably not behave much differently if they were set in some other rural locality of England. The popularity of Sir Walter Scott's "Waverley" novels is due in part to their evocation of a romantic Scotland. Setting may be the prime consideration of some readers, who can be drawn to Conrad because he depicts life at sea or in the East Indies; they may be less interested in the complexity of human relationships that he presents.

Regional novels

The regional novel is a recognized species. The sequence of four novels that Hugh Walpole began with *Rogue Heries* (1930) was the result of his desire to do homage to the part of Cumberland, in England, where he had elected to live. The great Yoknapatawpha cycle of William Faulkner, a classic of 20th-century American literature set in an imaginary county in Mississippi, belongs to the category as much as the once-popular confections about Sussex that were written about the same time by the English novelist Sheila Kaye-Smith. Many novelists, however, gain a creative impetus from avoiding the same setting in book after book and deliberately seeking new locales. The English novelist Graham Greene apparently needed to visit a fresh scene in order to write a fresh novel. His ability to encapsulate the essence of an exotic setting in a single book is exemplified in *The Heart of the Matter* (1948); his contemporary Evelyn Waugh stated that the West Africa of that book replaced the true remembered West Africa of his own experience. Such power is not uncommon: the Yorkshire moors have been romanticized because Emily Brontë wrote of them in *Wuthering Heights* (1847), and literary tourists have visited Stoke-on-Trent, in northern England, because it comprises the "Five Towns" of Arnold Bennett's novels of the early 20th century. Others go to the Monterey, California, of John Steinbeck's novels in the expectation of experiencing a *frisson* added to the locality by an act of creative imagination. James Joyce, who remained inexhaustibly stimulated by Dublin, has exalted that city in a manner that even the guidebooks recognize.

The setting of a novel is not always drawn from a real-life locale. The literary artist sometimes prides himself on his ability to create the totality of his fiction—the setting as well as the characters and their actions. In the Russian expatriate Vladimir Nabokov's *Ada* (1969) there is an entirely new space–time continuum, and the English scholar J.R.R. Tolkien in his *Lord of the Rings* (1954–55) created an "alternative world" that appeals greatly to many who are dissatisfied with the existing one. The world of interplanetary travel was imaginatively created long before the first moon landing. The properties of the future envisaged by H.G. Wells's novels or by Aldous Huxley in *Brave New World* (1932) are still recognized in an age that those authors did not live to see. The composition of place can be a magical fictional gift.

Whatever the locale of his work, every true novelist is concerned with making a credible environment for his characters, and this really means a close attention to sense data—the immediacies of food and drink and colour—far more than abstractions like "nature" and "city." The London of Charles Dickens is as much incarnated in the smell of wood in lawyers' chambers as in the skyline and vistas of streets.

Narrative method and point of view. Where there is a story, there is a storyteller. Traditionally, the narrator of the epic and mock-epic alike acted as an intermediary between the characters and the reader; the method of Fielding is not very different from the method of Homer. Sometimes the narrator boldly imposed his own attitudes; always he assumed an omniscience that tended to reduce the characters to puppets and the action to a predetermined course with an end implicit in the beginning. Many novelists have been unhappy about a narrative method that seems to limit the free will of the characters, and innovations in fictional technique have mostly sought the objectivity of the drama, in which the characters appear to work out their own destinies without prompting from the author.

The epistolary method, most notably used by Samuel Richardson in *Pamela* (1740) and by Jean-Jacques Rousseau in *La nouvelle Héloïse* (1761), has the advantage of allowing the characters to tell the story in their own words, but it is hard to resist the uneasy feeling that a kind of divine editor is sorting and ordering the letters into his own pattern. The device of making the

narrator also a character in the story has the disadvantage of limiting the material available for the narration, since the narrator-character can know only those events in which he participates. There can, of course, be a number of secondary narratives enclosed in the main narrative, and this device—though it sometimes looks artificial—has been used triumphantly by Conrad and, on a lesser scale, by W. Somerset Maugham. A, the main narrator, tells what he knows directly of the story and introduces what B and C and D have told him about the parts that he does not know.

Seeking the most objective narrative method of all, Ford Madox Ford used, in *The Good Soldier* (1915), the device of the storyteller who does not understand the story he is telling. This is the technique of the "unreliable observer." The reader, understanding better than the narrator, has the illusion of receiving the story directly. Joyce, in both his major novels, uses different narrators for the various chapters. Most of them are unreliable, and some of them approach the impersonality of a sort of disembodied parody. In *Ulysses,* for example, an episode set in a maternity hospital is told through the medium of a parodic history of English prose style. But, more often than not, the sheer ingenuity of Joyce's techniques draws attention to the manipulator in the shadows. The reader is aware of the author's cleverness where he should be aware only of the characters and their actions. The author is least noticeable when he is employing the stream of consciousness device, by which the inchoate thoughts and feelings of a character are presented in interior monologue—apparently unedited and sometimes deliberately near-unintelligible. It is because this technique seems to draw fiction into the psychoanalyst's consulting room (presenting the raw material of either art or science, but certainly not art itself), however, that Joyce felt impelled to impose the shaping devices referred to above. Joyce, more than any novelist, sought total objectivity of narration technique but ended as the most subjective and idiosyncratic of stylists.

The presence of the author

The problem of a satisfactory narrative point of view is, in fact, nearly insoluble. The careful exclusion of comment, the limitation of vocabulary to a sort of reader's lowest common denominator, the paring of style to the absolute minimum—these puritanical devices work well for an Ernest Hemingway (who, like Joyce, remains, nevertheless, a highly idiosyncratic stylist) but not for a novelist who believes that, like poetry, his art should be able to draw on the richness of word play, allusion, and symbol. For even the most experienced novelist, each new work represents a struggle with the unconquerable task of reconciling all-inclusion with self-exclusion. It is noteworthy that Cervantes, in *Don Quixote,* and Nabokov, in *Lolita* (1955), join hands across four centuries in finding most satisfactory the device of the fictitious editor who presents a manuscript story for which he disclaims responsibility. But this highly useful method presupposes in the true author a scholarly, or pedantic, faculty not usually associated with novelists.

Scope, or dimension. No novel can theoretically be too long, but if it is too short it ceases to be a novel. It may or may not be accidental that the novels most highly regarded by the world are of considerable length—Cervantes' *Don Quixote,* Dostoyevsky's *Brothers Karamazov,* Tolstoy's *War and Peace,* Dickens' *David Copperfield,* Proust's *À la recherche du temps perdu,* and so on. On the other hand, since World War II, brevity has been regarded as a virtue in works like the later novels of the Irish absurdist author Samuel Beckett and the *ficciones* of the Argentine Jorge Luis Borges, and it is only an aesthetic based on bulk that would diminish the achievement of Ronald Firbank's short novels of the post-World War I era or the Evelyn Waugh who wrote *The Loved One* (1948). It would seem that there are two ways of presenting human character—one, the brief way, through a significant episode in the life of a personage or group of personages; the other, which admits of limitless length, through the presentation of a large section of a life or lives, sometimes beginning with birth and ending in old age. The plays of Shakespeare show that a full delineation of character can be effected in a very brief compass, so that, for this aspect of the novel, length confers no special advantage. Length, however, is essential when the novelist attempts to present something bigger than character—when, in fact, he aims at the representation of a whole society or period of history.

No other cognate art form—neither the epic poem nor the drama nor the film—can match the resources of the novel when the artistic task is to bring to immediate, sensuous, passionate life the somewhat impersonal materials of the historian. *War and Peace* is the great triumphant example of the panoramic study of a whole society—that of early 19th-century Russia—which enlightens as the historian enlightens and yet also conveys directly the sensations and emotions of living through a period of cataclysmic change. In the 20th century, another Russian, Boris Pasternak, in his *Doctor Zhivago* (1957), expressed—though on a less than Tolstoyan scale—the personal immediacies of life during the Russian Revolution. Though of much less literary distinction than either of these two books, Margaret Mitchell's *Gone with the Wind* (1936) showed how the American Civil War could assume the distanced pathos, horror, and grandeur of any of the classic struggles of the Old World.

Needless to say, length and weighty subject matter are no guarantee in themselves of fictional greatness. Among American writers, for example, James Jones's celebration of the U.S. Army on the eve of World War II in *From Here to Eternity* (1951), though a very ambitious project, repels through indifferent writing and sentimental characterization; Norman Mailer's *Naked and the Dead* (1948), an equally ambitious military novel, succeeds much more because of a tautness, a concern with compression, and an astringent objectivity that Jones was unable to match. Frequently the size of a novel is too great for its subject matter—as with Marguerite Young's *Miss MacIntosh, My Darling* (1965), reputedly the longest single-volume novel of the 20th century, John Barth's *Giles Goat-Boy* (1966), and John Fowles's *Magus* (1965). Diffuseness is the great danger in the long novel, and diffuseness can mean slack writing, emotional self-indulgence, sentimentality.

Even the long picaresque novel—which, in the hands of a Fielding or his contemporary Tobias Smollett, can rarely be accused of sentimentality—easily betrays itself into such acts of self-indulgence as the multiplication of incident for its own sake, the coy digression, the easygoing jogtrot pace that subdues the sense of urgency that should lie in all fiction. If Tolstoy's *War and Peace* is a greater novel than Fielding's *Tom Jones* or Dickens' *David Copperfield,* it is not because its theme is nobler, or more pathetic, or more significant historically; it is because Tolstoy brings to his panoramic drama the compression and urgency usually regarded as the monopolies of briefer fiction.

Works in several books

Sometimes the scope of a fictional concept demands a technical approach analogous to that of the symphony in music—the creation of a work in separate books, like symphonic movements, each of which is intelligible alone but whose greater intelligibility depends on the theme and characters that unify them. The French author Romain Rolland's *Jean-Christophe* (1904–12) sequence is, very appropriately since the hero is a musical composer, a work in four movements. Among works of English literature, Lawrence Durrell's *Alexandria Quartet* (1957–60) insists in its very title that it is a tetralogy rather than a single large entity divided into four volumes; the concept is "relativist" and attempts to look at the same events and characters from four different viewpoints. Anthony Powell's *Dance to the Music of Time,* a multivolume series of novels that began in 1951 (collected 1962), may be seen as a study of a segment of British society in which the chronological approach is eschewed, and events are brought together in one volume or another because of a kind of parachronic homogeneity. C.P. Snow's *Strangers and Brothers,* a comparable series that began in 1940 and continued to appear throughout the '50s and into the '60s, shows how a fictional concept can be realized only in the act of writing, since the publication of the earlier volumes antedates the historical events portrayed in later ones. In other words, the author could not know what the subject matter of the sequence would be until he was in sight of its end. Behind all these works lies the giant example of

Proust's *roman-fleuve,* whose length and scope were properly coterminous with the author's own life and emergent understanding of its pattern.

Myth, symbolism, significance. The novelist's conscious day-to-day preoccupation is the setting down of incident, the delineation of personality, the regulation of exposition, climax, and denouement. The aesthetic value of his work is frequently determined by subliminal forces that seem to operate independently of the writer, investing the properties of the surface story with a deeper significance. A novel will then come close to myth, its characters turning into symbols of permanent human states or impulses, particular incarnations of general truths perhaps only realized for the first time in the act of reading. The ability to perform a quixotic act anteceded *Don Quixote,* just as *bovarysme* existed before Flaubert found a name for it.

But the desire to give a work of fiction a significance beyond that of the mere story is frequently conscious and deliberate, indeed sometimes the primary aim. When a novel—like Joyce's *Ulysses* or John Updike's *Centaur* (1963) or Anthony Burgess' *Vision of Battlements* (1965)—is based on an existing classical myth, there is an intention of either ennobling a lowly subject matter, satirizing a debased set of values by referring them to a heroic age, or merely providing a basic structure to hold down a complex and, as it were, centrifugal picture of real life. Of *Ulysses,* Joyce said that his Homeric parallel (which is worked out in great and subtle detail) was a bridge across which to march his 18 episodes; after the march the bridge could be "blown skyhigh." But there is no doubt that, through the classical parallel, the account of an ordinary summer day in Dublin is given a richness, irony, and universality unattainable by any other means.

The mythic or symbolic intention of a novel may manifest itself less in structure than in details which, though they appear naturalistic, are really something more. The shattering of the eponymous golden bowl in Henry James's 1904 novel makes palpable, and hence truly symbolic, the collapse of a relationship. Even the choice of a character's name may be symbolic. Sammy Mountjoy, in William Golding's *Free Fall* (1959), has fallen from the grace of heaven, the mount of joy, by an act of volition that the title makes clear. The eponym of *Doctor Zhivago* is so called because his name, meaning "the living," carries powerful religious overtones. In the Russian version of the Gospel According to St. Luke, the angels ask the women who come to Christ's tomb: *"Chto vy ischyote zhivago mezhdu myortvykh?"*—"Why do you seek the living among the dead?" And his first name, Yuri, the Russian equivalent of George, has dragon-slaying connotations.

The symbolism of names

The symbol, the special significance at a subnarrative level, works best when it can fit without obtrusion into a context of naturalism. The optician's trade sign of a huge pair of spectacles in F. Scott Fitzgerald's *Great Gatsby* (1925) is acceptable as a piece of scenic detail, but an extra dimension is added to the tragedy of Gatsby, which is the tragedy of a whole epoch in American life, when it is taken also as a symbol of divine myopia. Similarly, a cinema poster in Malcolm Lowry's *Under the Volcano* (1947), advertising a horror film, can be read as naturalistic background, but it is evident that the author expects the illustrated fiend—a concert pianist whose grafted hands are those of a murderer—to be seen also as a symbol of Nazi infamy; the novel is set at the beginning of World War II, and the last desperate day of the hero, Geoffrey Firmin, stands also for the collapse of Western civilization.

There are symbolic novels whose infranarrative meaning cannot easily be stated, since it appears to subsist on an unconscious level. Herman Melville's *Moby Dick* (1851) is such a work, as is D.H. Lawrence's novella *St. Mawr* (1925), in which the significance of the horse is powerful and mysterious.

USES

Interpretation of life. Novels are not expected to be didactic, like tracts or morality plays; nevertheless, in varying degrees of implicitness, even the "purest" works of fictional art convey a philosophy of life. The novels of Jane Austen, designed primarily as superior entertainment, imply a desirable ordered existence, in which the comfortable decorum of an English rural family is disturbed only by a not too serious shortage of money, by love affairs that go temporarily wrong, and by the intrusion of self-centred stupidity. The good, if unrewarded for their goodness, suffer from no permanent injustice. Life is seen, not only in Jane Austen's novels but in the whole current of bourgeois Anglo-American fiction, as fundamentally reasonable and decent. When wrong is committed, it is usually punished, thus fulfilling Miss Prism's summation in Oscar Wilde's play *The Importance of Being Earnest* (1895), to the effect that in a novel the good characters end up happily and the bad characters unhappily: "that is why it is called fiction."

That kind of fiction called realistic, which has its origins in 19th-century France, chose the other side of the coin, showing that there was no justice in life and that the evil and the stupid must prevail. In the novels of Thomas Hardy there is a pessimism that may be taken as a corrective of bourgeois Panglossianism—the philosophy that everything happens for the best, satirized in Voltaire's *Candide* (1759)—since the universe is presented as almost impossibly malevolent. This tradition is regarded as morbid, and it has been deliberately ignored by most popular novelists. The "Catholic" novelists such as François Mauriac in France, Graham Greene in England, and others—see life as mysterious, full of wrong and evil and injustice inexplicable by human canons but necessarily acceptable in terms of the plans of an inscrutable God. Between the period of realistic pessimism, which had much to do with the agnosticism and determinism of 19th-century science, and the introduction of theological evil into the novel, writers such as H.G. Wells attempted to create a fiction based on optimistic liberalism. As a reaction, there was the depiction of "natural man" in the novels of D.H. Lawrence and Ernest Hemingway.

For the most part, the view of life common to American and European fiction since World War II posits the existence of evil—whether theological or of that brand discovered by the French Existentialists, particularly Jean-Paul Sartre—and assumes that man is imperfect and life possibly absurd. The fiction of the Communist states is based on a very different assumption, one that seems naïve and old-fashioned in its collective optimism to readers in the disillusioned democracies. It is to be noted that in the Soviet Union aesthetic evaluation of fiction has been replaced by ideological judgment. The works of the popular British writer A.J. Cronin, since they seem to depict personal tragedy as an emanation of capitalistic infamy, are rated higher than those of Conrad, James, and their peers. The novel as an art form stands or falls by its capacity to express a view of life acceptable to the Soviet authorities.

The Communist interpretation

Entertainment or escape. In a period that takes for granted that the written word should be "committed"—to the exposure of social wrong or the propagation of progressive ideologies—novelists who seek merely to take the reader out of his dull or oppressive daily life are not highly regarded, except by that reading public that has never expected a book to be anything more than a diversion. Nevertheless, the provision of laughter and dreams has been for many centuries a legitimate literary occupation. It can be condemned by serious devotees of literature only if it falsifies life through oversimplification and tends to corrupt its readers into belief that reality is as the author presents it. The novelettes once beloved of mill girls and domestic servants, in which the beggar maid was elevated to queendom by a king of high finance, were a mere narcotic, a sort of enervating opium of the oppressed; the encouragement of such subliterature might well be one of the devices of social oppression. Adventure stories and spy novels may have a healthy enough astringency, and the very preposterousness of some adventures can be a safeguard against any impressionable young reader's neglecting the claims of real life to dream of becoming a secret agent. The subject matter of some humorous novels—such as the effete British aristocracy created by P.G. Wodehouse, which is no longer in existence if it ever was—can never be identified with a real human society; the dream is accepted as a dream. The same may be said of Evelyn Waugh's early novels—such as *Decline and*

Fall (1928) and *Vile Bodies* (1930)—but these are raised above mere entertainment by touching, almost incidentally, on real human issues (the relation of the innocent to a circumambient malevolence is a persistent theme in all Waugh's writing).

Any reader of fiction has a right to an occasional escape from the dullness or misery of his existence, but he has the critical duty of finding the best modes of escape—in the most efficiently engineered detective or adventure stories, in humor that is more than sentimental buffonery, in dreams of love that are not mere pornography. The fiction of entertainment and escape frequently sets itself higher literary standards than novels with a profound social or philosophical purpose. Books like John Buchan's *Thirty-nine Steps* (1915), Graham Greene's *Travels with My Aunt* (1969), Dashiell Hammett's *Maltese Falcon* (1930), and Raymond Chandler's *Big Sleep* (1939) are distinguished pieces of writing that, while diverting and enthralling, keep a hold on the realities of human character. Ultimately, all good fiction is entertainment, and if it instructs or enlightens it does so best through enchanting the reader.

Propaganda. The desire to make the reader initiate certain acts—social, religious, or political—is the essence of all propaganda, and, though it does not always accord well with art, the propagandist purpose has often found its way into novels whose prime value is an aesthetic one. The *Nicholas Nickleby* (1839) of Charles Dickens attacked the abuses of schools to some purpose, as his *Oliver Twist* (1838) drew attention to the horrors of poorhouses and his *Bleak House* (1853) to the abuses of the law of chancery. The weakness of propaganda in fiction is that it loses its value when the wrongs it exposes are righted, so that the more successful a propagandist novel is, the briefer the life it can be expected to enjoy. The genius of Dickens lay in his ability to transcend merely topical issues through the vitality with which he presented them, so that his contemporary disclosures take on a timeless human validity—chiefly through the power of their drama, character, and rhetoric.

The ephemeral quality of propaganda

The purely propagandist novel—which Dickens was incapable of writing—quickly becomes dated. The "social" novels of H.G. Wells, which propounded a rational mode of life and even blueprinted utopias, were very quickly exploded by the conviction of man's irredeemable irrationality that World War I initiated and World War II corroborated, a conviction the author himself came to share toward the end of his life. But the early scientific romances of Wells remain vital and are seen to have been prophetic. Most of the fiction of the Soviet Union, which either glorifies the regime or refrains from criticizing it, is dull and unreal, and the same may be said of Communist fiction elsewhere. Propaganda too frequently ignores man as a totality, concentrating on him aspectively—in terms of politics or sectarian religion. When a didactic attack on a system, as in Harriet Beecher Stowe's attack on slavery in the United States in *Uncle Tom's Cabin* (1852), seems to go beyond mere propaganda, it is because the writer makes the reader aware of wrongs and injustices that are woven into the permanent human fabric. The reader's response may be a modification of his own sensibility, not an immediate desire for action, and this is one of the legitimate effects of serious fiction. The propagandist Dickens calls for the immediate righting of wrongs, but the novelist Dickens says, mainly through implication, that all men—not just schoolmasters and state hirelings—should become more humane. If it is possible to speak of art as possessing a teaching purpose, this is perhaps its only lesson.

Reportage. The division in the novelist's mind is between his view of his art as a contrivance, like a Fabergé watch, and his view of it as a record of real life. The versatile English writer Daniel Defoe, on the evidence of such novels as his *Journal of the Plague Year* (1722), a recreation of the London plague of 1665, believed that art or contrivance had the lesser claim and proceeded to present his account of events of which he had had no direct experience in the form of plain journalistic reportage. This book, like his *Robinson Crusoe* (1719) and *Moll Flanders* (1722), is more contrived and cunning than it appears, and the hurried, unshaped narrative is the product of careful preparation and selective ordering. His example, which could have been a very fruitful one, was not much followed until the 20th century, when the events of the real world became more terrifying and marvellous than anything the novelist could invent, and seemed to ask for that full imaginative treatment that only the novelist's craft can give.

In contemporary American literature, John Hersey's *Hiroshima* (1946), though it recorded the actual results of the nuclear attack on the Japanese city in 1945, did so in terms of human immediacies, not scientific or demographic abstractions, and this approach is essentially novelistic. Truman Capote's *In Cold Blood* (1966) took the facts of a multiple murder in the Midwest of the United States and presented them with the force, reality, tone, and (occasionally) overintense writing that distinguish his genuine fiction. Norman Mailer, in *The Armies of the Night* (1968), recorded, in great personal detail but in a third-person narration, his part in a citizens' protest march on Washington, D.C. It would seem that Mailer's talent lies in his ability to merge the art of fiction and the craft of reportage, and his *Of a Fire on the Moon* (1970), which deals with the American lunar project, reads like an episode in an emergent *roman-fleuve* of which Mailer is the central character.

The presentation of factual material as art is the purpose of such thinly disguised biographies as Somerset Maugham's *Moon and Sixpence* (1919), undisguised biographies fleshed out with supposition and imagination like Helen Waddell's *Peter Abelard* (1933), and many autobiographies served up—out of fear of libel or of dullness—as novels. Conversely, invented material may take on the lineaments of journalistic actuality through the employment of a Defoe technique of flat understatement. This is the way of such science fiction as Michael Crichton's *Andromeda Strain* (1969), which uses sketch maps, computer projections, and simulated typewritten reports.

Agent of change in language and thought. Novelists, being neither poets nor philosophers, rarely originate modes of thinking and expression. Poets such as Chaucer and Shakespeare have had much to do with the making of the English language, and Byron was responsible for the articulation of the new romantic sensibility in it in the early 19th century. Books like the Bible, Karl Marx's *Kapital,* and Adolf Hitler's *Mein Kampf* may underlie permanent or transient cultures, but it is hard to find, except in the early Romantic period, a novelist capable of arousing new attitudes to life (as opposed to aspects of the social order) and forging the vocabulary of such attitudes.

Sentiment

With the 18th-century precursors of Romanticism—notably Richardson, Sterne, and Rousseau—the notion of sentiment entered the European consciousness. Rousseau's *Nouvelle Héloïse* fired a new attitude to ward love—more highly emotional than ever before—as his *Émile* (1762) changed educated views on how to bring up children. The romantic wave in Germany, with Goethe's *Sorrows of Werther* (1774) and the works of Jean-Paul Richter a generation later, similarly aroused modes of feeling that rejected the rational constraints of the 18th century. Nor can the influence of Sir Walter Scott's novels be neglected, both on Europe and on the American South (where Mark Twain thought it had had a deplorable effect). With Scott came new forms of regional sentiment, based on a romantic reading of history.

It is rarely, however, that a novelist makes a profound mark on a national language, as opposed to a regional dialect (to which, by using it for a literary end, he may impart a fresh dignity). It is conceivable that Alessandro Manzoni's *I promessi sposi* (1825–27; *The Betrothed*), often called the greatest modern Italian novel, gave 19th-century Italian intellectuals some notion of a viable modern prose style in an Italian that might be termed "national," but even this is a large claim. Günter Grass, in post-Hitler Germany, sought to revivify a language that had been corrupted by the Nazis; he threw whole dictionaries at his readers, in the hope that new freedom, fantasy, and exactness in the use of words might influence the publicists, politicians, and teachers in the direction of a new liberalism of thought and expression.

It is difficult to say whether the French Existentialists, such as Sartre and Albert Camus, have influenced their age primarily through their fiction or their philosophical writings. Certainly, Sartre's early novel *Nausea* (1938) established unforgettable images of the key terms of his philosophy, which has haunted a whole generation, as Camus's novel *The Stranger* (1942) created for all time the lineaments of "Existential man." In the same way, the English writer George Orwell's *Nineteen Eighty-four* (1949) incarnated brilliantly the nature of the political choices that are open to 20th-century humanity, and, with terms like "Big Brother" (*i.e.,* the leader of an authoritarian state) and "doublethink" (belief in contradictory ideas simultaneously), modified the political vocabulary. But no novelist's influence can compare to that of the poet's, who can give a language a soul and define, as Shakespeare and Dante did, the scope of a culture.

Expression of the spirit of its age. The novelist, like the poet, can make the inchoate thoughts and feelings of a society come to articulation through the exact and imaginative use of language and symbol. In this sense, his work seems to precede the diffusion of new ideas and attitudes and to be the agent of change. But it is hard to draw a line between this function and that of expressing an existing climate of sensibility. Usually the nature of a historical period—that spirit known in German as the *Zeitgeist*—can be understood only in long retrospect, and it is then that the novelist can provide its best summation. The sickness of the Germany that produced Hitler had to wait some time for fictional diagnosis in such works as Thomas Mann's *Doctor Faustus* (1947) and, later, Günter Grass's *Tin Drum* (1959). Evelyn Waugh waited several years before beginning, in the trilogy *Sword of Honour,* to depict that moral decline of English society that started to manifest itself in World War II, the conduct of which was both a cause and a symptom of the decay of traditional notions of honour and justice.

The lag between events and their use in novels

The novel can certainly be used as a tool for the better understanding of a departed age. The period following World War I had been caught forever in Hemingway's *Sun Also Rises* (1926; called *Fiesta* in England), F. Scott Fitzgerald's novels and short stories about the so-called Jazz Age, the *Antic Hay* (1923) and *Point Counter Point* (1928) of Aldous Huxley, and D.H. Lawrence's *Aaron's Rod* (1922) and *Kangaroo* (1923). The spirit of the English 18th century, during which social, political, and religious ideas associated with rising middle classes conflicted with the old Anglican Tory rigidities, is better understood through reading Smollett and Fielding than by taking the cerebral elegance of Pope and his followers as the typical expression of the period.

Similarly, the unrest and bewilderment of the young in the period after World War II still speak in novels like J.D. Salinger's *Catcher in the Rye* (1951) and Kingsley Amis' *Lucky Jim* (1954). It is notable that with novels like these—and the beat-generation books of Jack Kerouac; the American-Jewish novels of Saul Bellow, Bernard Malamud, and Philip Roth; and the black novels of Ralph Ellison and James Baldwin—it is a segmented spirit that is expressed, the spirit of an age group, social group, or racial group, and not the spirit of an entire society in a particular phase of history. But probably a *Zeitgeist* has always been the emanation of a minority, the majority being generally silent. The 20th century seems, from this point of view, to be richer in vocal minorities than any other period in history.

Creator of life-style and arbiter of taste. Novels have been known to influence, though perhaps not very greatly, modes of social behaviour and even, among the very impressionable, conceptions of personal identity. But more young men have seen themselves as Hamlet or Childe Harold than as Julien Sorel, the protagonist of Stendhal's novel *The Red and the Black* (1830), or the sorrowing Werther. Richardson's novel may popularize Pamela, or Galsworthy's *Forsyte Saga* (1906–22) Jon, as a baptismal name, but it rarely makes a deeper impression on the mode of life of literate families. On the other hand, the capacity of Oscar Wilde's *Picture of Dorian Gray* (1891) to influence young men in the direction of sybaritic amorality, or of D.H. Lawrence's *Lady Chatterley's Lover* (1928) to engender a freer attitude to sex, has never been assessed adequately. With the lower middle class reading public, the effect of devouring *The Forsyte Saga* was to engender genteelisms—cucumber sandwiches for tea, supper renamed dinner—rather than to learn that book's sombre lesson about the decline of the old class structure. Similarly, the ladies who read Scott in the early 19th century were led to barbarous ornaments and tastefully arranged folk songs.

Fiction has to be translated into one of the dramatic media—stage, film, or television—before it can begin to exert a large influence. *Tom Jones* as a film in 1963 modified table manners and coiffures and gave American visitors to Great Britain a new (and probably false) set of expectations. The stoic heroes of Hemingway, given to drink, fights, boats, and monosyllables, became influential only when they were transferred to the screen. They engendered other, lesser heroes—incorruptible private detectives, partisans brave under interrogation—who in their turn have influenced the impressionable young when seeking an identity. Ian Fleming's James Bond led to a small revolution in martini ordering. But all these influences are a matter of minor poses, and such poses are most readily available in fiction easily adapted to the mass media—which means lesser fiction. Proust, though he recorded French patrician society with painful fidelity, had little influence on it, and it is hard to think of Henry James disturbing the universe even fractionally. Films and television programs dictate taste and behaviour more than the novel ever could.

Fiction in other media

STYLE

Romanticism. The Romantic movement in European literature is usually associated with those social and philosophical trends that prepared the way for the French Revolution, which began in 1789. The somewhat subjective, anti-rational, emotional currents of romanticism transformed intellectual life in the revolutionary and Napoleonic periods and remained potent for a great part of the 19th century. In the novel, the romantic approach to life was prepared in the "sentimental" works of Richardson and Sterne and attained its first major fulfillment in the novels of Rousseau. Sir Walter Scott, in his historical novels, turned the past into a great stage for the enactment of events motivated by idealism, chivalry, and strong emotional impulse, using an artificially archaic language full of remote and magical charm. The exceptional soul—poet, patriot, idealist, madman—took the place of dully reasonable fictional heroes, such as Tom Jones, and sumptuous and mysterious settings ousted the plain town and countryside of 18th-century novels.

The romantic novel must be seen primarily as a historical phenomenon, but the romantic style and spirit, once they had been brought into being, remained powerful and attractive enough to sustain a whole subspecies of fiction. The cheapest love story can be traced back to the example of Charlotte Brontë's *Jane Eyre* (1847), or even Rousseau's earlier *Nouvelle Héloïse.* Similarly, best-selling historical novels, even those devoid of literary merit, can find their progenitor in Scott, and science fiction in Mary Shelley's *Frankenstein* (1818), a romantic novel subtitled *The Modern Prometheus,* as well as in Jules Verne and H.G. Wells. The aim of romantic fiction is less to present a true picture of life than to arouse the emotions through a depiction of strong passions, or to fire the imagination with exotic, terrifying, or wonderful scenes and events. When it is condemned by critics, it is because it seems to falsify both life and language; the pseudopoetical enters the dialogue and *récit* alike, and humanity is seen in only one of its aspects—that of feeling untempered with reason.

If such early romantic works as those of Scott and of the Goethe of *The Sorrows of Werther* have long lost their original impact, the romantic spirit still registers power and truth in the works of the Brontës—particularly in Emily Brontë's *Wuthering Heights,* in which the poetry is genuine and the strange instinctual world totally convincing. Twentieth-century romantic fiction records few masterpieces. Writers like Daphne du Maurier, the author of *Jamaica Inn* (1936), *Rebecca* (1938), and many others,

Decline of the romantic novel

are dismissed as mere purveyors of easy dreams. It is no more possible in the 20th century to revive the original romantic élan in literature than it is to compose music in the style of Beethoven. Despite the attempts of Lawrence Durrell to achieve a kind of decadent romantic spirit in his *Alexandria Quartet,* the strong erotic feeling, the exotic setting, the atmosphere of poetic hallucination, the pain, perversion, and elemental force seem to be contrivances, however well they fulfill the original romantic prescription.

Realism. Certain major novelists of the 19th century, particularly in France, reacted against romanticism by eliminating from their work those "softer" qualities—tenderness, idealism, chivalric passion, and the like—which seemed to them to hide the stark realities of life in a dreamlike haze. In Gustave Flaubert's works there are such romantic properties—his novel *Salammbô* (1862), for instance, is a sumptuous representation of a remote pagan past—but they are there only to be punctured with realistic irony. On one level, his *Madame Bovary* may be taken as a kind of parable of the punishment that fate metes out to the romantic dreamer, and it is the more telling because Flaubert recognized a strong romantic vein in himself: "Madame Bovary, c'est moi" ("Madame Bovary is myself"). Stendhal and Balzac, on the other hand, admit no dreams and present life in a grim nakedness without poetic drapery.

Balzac's mammoth fictional work—the 20-year succession of novels and stories he published under the collective title *La Çomédie humaine* (*The Human Comedy*)—and Stendhal's novels of the same period, *The Red and the Black* (1830) and *The Charterhouse of Parma* (1839), spare the reader nothing of those baser instincts in man and society that militate against, and eventually conquer, many human aspirations. Rejecting romanticism so energetically, however, they swing to an extreme that makes "realism" a synonym for unrelenting pessimism. Little comes right for the just or the weak, and base human nature is unqualified by even a modicum of good. But there is a kind of affirmative richness and energy about both writers that seems to belie their pessimistic thesis.

The pessimistic view

In England, George Eliot in her novel *Middlemarch* (1871–72) viewed human life grimly, with close attention to the squalor and penury of rural life. If "nature" in works by romantic poets like Wordsworth connoted a kind of divine benevolence, only the "red in tooth and claw" aspect was permitted to be seen in the novels of the realists. George Eliot does not accept any notion of Divine Providence, whether Christian or pantheistic, but her work is instinct with a powerful moral concern: her characters never sink into a deterministic morass of hopelessness, since they have free will, or the illusion of it. With Thomas Hardy, who may be termed the last of the great 19th-century novelists, the determinism is all-pervasive, and his final novel, *Jude the Obscure* (1896), represents the limit of pessimism. Behind him one is aware of the new science, initiated by the biologists Charles Darwin and T.H. Huxley, which displaces man as a free being, capable of choice, by a view of him as the product of blind mechanistic forces over which he has little control.

Realism in this sense has been a continuing impulse in the 20th-century novel, but few writers would go so far as Hardy in positing man's near-total impotence in a hostile universe, with the gods killing human creatures for their sport. Realism in the Existentialist fiction of 20th-century France, for instance, makes man not merely wretched but absurd, yet it does not diminish his power of self-realization through choice and action. Realism has frequently been put in the service of a reforming design, which implies a qualified optimism. War novels, novels about the sufferings of the oppressed (in prison, ghetto, totalitarian state), studies of human degradation that are bitter cries against man-made systems—in all of these the realistic approach is unavoidable, and realistic detail goes much further than anything in the first realists. But there is a difference in the quality of the anger the reader feels when reading the end of Hardy's *Tess of the D'Urbervilles* (1891) and that generated by Upton Sinclair's *Jungle* (1906) or Erich Maria Remarque's *All Quiet on the Western Front* (1929). In Hardy's novel, pessimistic determinism, reducing human character to pain, frustration, and impotent anger, was—paradoxically—appropriate to an age that knew no major cataclysms or oppressions. The novels of Sinclair and Remarque reflect the 20th century, which saw the origin of all wrong in the human will, and set on a program of diagnosis and reform.

Realism as reform

Naturalism. The naturalistic novel is a development out of realism, and it is, again, in France that its first practitioners are to be found, with Émile Zola leading. It is difficult to separate the two categories, but naturalism seems characterized not only by a pessimistic determinism but also by a more thoroughgoing attention to the physical and biological aspects of human existence. Man is less a soul aspiring upward to its divine source than a product of natural forces, as well as genetic and social influences, and the novelist's task is to present the physical essence of man and his environment. The taste of Balzac's and Stendhal's audiences was not easily able to accommodate itself to utter frankness about the basic processes of life, and the naturalists had to struggle against prejudice, and often censorship, before their literary candour was able to prevail. The 20th century takes the naturalistic approach for granted, but it is more concerned with a technique of presentation than with the somewhat mechanistic philosophy of Zola and his followers.

Naturalism received an impetus after World War I, when novelists felt they had a duty to depict the filth, suffering, and degradation of the soldier's life, without euphemism or circumlocution. Joyce's *Ulysses,* when it appeared in 1922, was the first novel to seek to justify total physical candour in terms of its artistic, as opposed to moral, aim—which was to depict with almost scientific objectivity every aspect of an ordinary urban day. Though Joyce had read Zola, he seems to invoke the spirit of a very much earlier naturalistic writer—the ribald French author of the 16th century, François Rabelais—and this is in keeping with the Catholic tradition that Joyce represents. Zola, of course, was an atheist.

It would have been a sin against his aesthetic canons for Joyce to have shown Leopold Bloom—the protagonist of *Ulysses*—eating breakfast or taking a bath and yet not defecating or masturbating. The technique of the interior monologue, which presented the unedited flow of a character's unspoken thought and emotion, also called for the utmost frankness in dealing with natural functions and urges. Joyce, it is now recognized, had no prurient or scatological intention; his concern was with showing life as it is (without any of the didactic purpose of Zola), and this entailed the presentation of lust, perversion, and blasphemy as much as any of the traditionally acceptable human functions.

The naturalistic novelists have had their social and legal problems—obscenity indictments, confiscation, emasculation by timid publishers—but the cause was ultimately won, at least in Great Britain and the United States, where there are few limits placed on the contemporary novelist's proclaimed right to be true to nature. In comparison with much contemporary fiction the pioneer work of Zola seems positively reticent.

Impressionism. The desire to present life with frank objectivity led certain early 20th-century novelists to question the validity of long-accepted narrative conventions. If truth was the novelist's aim, then the tradition of the omniscient narrator would have to go, to be replaced by one in which a fallible, partially ignorant character—one involved in the story and hence himself subject to the objective or naturalistic approach—recounted what he saw and heard. But the Impressionist painters of late 19th-century France had proclaimed a revision of the whole seeing process: they distinguished between what the observer assumed he was observing and what he actually observed. That cerebral editing which turned visual data into objects of geometric solidity had no place in Impressionist painting; the visible world became less definite, more fluid, resolving into light and colour.

The influence of the painters

The German novelists Thomas Mann and Hermann Hesse, moving from the realist tradition, which concentrated on closely notated detail in the exterior world, sought the lightness and clarity of a more elliptical style, and were

proclaimed Impressionists. But in England Ford Madox Ford went much further in breaking down the imagined rigidities of the space–time continuum, liquidating step-by-step temporal progression and making the visual world shimmer, dissolve, reconstitute itself. In Ford's tetralogy *Parade's End* (1924–28), the reader moves freely within the time continuum, as if it were spatial, and the total picture is perceived through an accumulation of fragmentary impressions. Ford's masterpiece, *The Good Soldier*, pushes the technique to its limit: the narrator tells his story with no special dispensation to see or understand more than a fallible being can, and, in his reminiscences, he fragments whole sequences of events as he ranges freely through time (such freedom had traditionally been regarded as a weakness, a symptom of the disease of inattention).

In the approach to dialogue manifested in a book that Ford wrote jointly with Conrad—*The Inheritors* (1901)—a particular aspect of literary impressionism may be seen whose suggestiveness has been ignored by other modern novelists. As the brain imposes its own logical patterns on the phenomena of the visual world, so it is given to editing into clarity and conciseness the halting utterances of real-life speech; the characters of most novels are impossibly articulate. Ford and Conrad attempted to present speech as it is actually spoken, with many of the meaningful solidities implied rather than stated. The result is sometimes exasperating, but only as real-life conversation frequently is.

The interior monologue, which similarly resists editing, may be regarded as a development of this technique. To show pre-articulatory thought, feeling, and sensuous perception unordered into a rational or "literary" sequence is an impressionistic device that, beginning in Édouard Dujardin's minor novel *Les Lauriers sont coupés* (1888; *We'll to the Woods No More*), served fiction of high importance, from Dorothy Richardson, Joyce, and Virginia Woolf to William Faulkner and Samuel Beckett.

Novelists like Ronald Firbank and Evelyn Waugh (who studied painting and was a competent draftsman) learned, in a more general sense, how to follow the examples of the Impressionist and Postimpressionist painters in their fiction. A spare brilliance of observation, like those paintings in which a whole scene is suggested through carefully selected points of colour, replaced that careful delineation of a whole face, or inventorying of a whole room, that had been the way of Balzac and other realists. In four or five brief lines of dialogue Waugh can convey as much as the 19th-century novelists did in as many pages.

Expressionism. Expressionism was a German movement that found its most congenial media in painting and drama. The artist's aim was to express, or convey the essence of, a particular theme, to the exclusion of such secondary considerations as fidelity to real life. The typical Expressionist play, by Bertolt Brecht, for example, concerns itself with a social or political idea that is hurled at the audience through every possible stage device—symbols, music, cinematic insertions, choral speech, dance. Human character is less important than the idea of humanity, and probability of action in the old realist sense is the least of the dramatist's concerns. The emotional atmosphere is high-pitched, even ecstatic, and the tone is more appropriate to propaganda than to art. Expressionistic technique, as the plays of Brecht prove, was an admirable means of conveying a Communist program, and it was in the service of such a program that John Dos Passos, in the trilogy of novels *U.S.A.* (1937), used literary devices analogous to the dramatic ones of Brecht—headlines, tabloid biographies, popular songs, lyric soliloquies, and the like.

Expressionism as alienation

But the Austro–Czech Franz Kafka, the greatest of the Expressionist novelists, sought to convey what may crudely be termed man's alienation from his world in terms that admit of no political interpretation. Joseph K., the hero of Kafka's novel *The Trial* (1925), is accused of a nameless crime, he seeks to arm himself with the apparatus of a defense, and he is finally executed—stabbed with the utmost courtesy by two men in a lonely place. The hallucinatory atmosphere of that novel, as also of his novel *The Castle* (1926), is appropriate to nightmare, and indeed Kafka's work has been taken by many as an imaginative forecast of the nightmare through which Europe was compelled to live during the Hitler regime. But its significance is more subtle and universal; one of the elements is original sin and another filial guilt. In the story *The Metamorphosis* (1915) a young man changes into an enormous insect, and the nightmare of alienation can go no further.

Kafka's influence has been considerable. Perhaps his most distinguished follower is the English writer Rex Warner, whose *Wild Goose Chase* (1937) and *Aerodrome* (1941) use fantasy, symbol, and improbable action for an end that is both Marxist and Freudian; the filial guilt, however, seems to be taken directly from Kafka, with an innocent hero caught in a monstrously oppressive web that is both the totalitarian state and paternal tyranny. More recently, the American writer William Burroughs has developed his own Expressionistic techniques in *The Naked Lunch* (1959), which is concerned with the alienation from society of the drug addict. His later novels *Nova Express* (1964) and *The Ticket That Exploded* (1962) use obscene fantasy to present a kind of metaphysical struggle between free spirit and enslaved flesh, evidently an extrapolation of the earlier drug theme. Burroughs is a didactic novelist, and didacticism functions best in a fictional ambience that rejects the complexities of character and real-life action.

Avant-gardism. Many innovations in fiction can be classified under headings already considered. Even so revolutionary a work as Joyce's *Finnegans Wake* represents an attempt to show the true nature of a dream; this can be regarded as a kind of Impressionism pushed so far that it looks like Surrealism. The brief novels of Samuel Beckett (which, as they aim to demonstrate the inadequacy of language to express the human condition, become progressively more brief) seem to have a kind of Expressionist derivation, since everything in them is subordinated to a central image of man as a totally deprived creature, resentful of a God he does not believe in. The French anti-novel, dethroning man as a primary concern of fiction, perhaps represents the only true break with traditional technique that the 20th-century novel has seen.

Technical innovation

Dissatisfaction not only with the content of the traditional novel but with the manner in which readers have been schooled to approach it has led the contemporary French novelist Michel Butor, in *Mobile,* to present his material in the form of a small encyclopaedia, so that the reader finds his directions obliquely, through an alphabetic taxonomy and not through the logic of sequential events. Nabokov, in *Pale Fire* (1962), gives the reader a poem of 999 lines and critical apparatus assembled by a madman; again the old sense of direction (beginning at the beginning and going on to the end) has been liquidated, yet *Pale Fire* is a true and highly intelligible novel. In England, B.S. Johnson published similar "false-directional" novels, though the influence of Sterne makes them seem accessible, even cozily traditional. One of Johnson's books is marketed as a bundle of disjunct chapters—which may thus be dealt aleatorially and read in any order.

Available avant-garde techniques are innumerable, though not all of them are salable. There is the device of counterpointing a main narrative with a story in footnotes, which eventually rises like water and floods the other. A novel has been written, though not published, in which the words are set (rather like the mouse's tail or tale in *Alice in Wonderland*) to represent graphically the physical objects in the narrative. Burroughs has experimented with a tricolumnar technique, in which three parallel narratives demand the reader's attention. But the writers like Borges and Nabokov go beyond mere technical innovation: they ask for a reconsideration of the very essence of fiction. In one of his *ficciones,* Borges strips from the reader even the final illusion that he is reading a story, for the story is made to dissolve, the artist evidently losing faith in his own artifact. Novels, as both Borges and Nabokov show, can turn into poems or philosophical essays, but they cannot, while remaining literature, turn into compositions disclaiming all interest in the world of feeling, thought, and sense. The novelist can do anything he pleases with his art so long as he interprets, or even just presents, a world that the reader recognizes as existing, or capable of existing, or capable of being dreamed of as existing.

TYPES OF NOVEL

Historical. For the hack novelist, to whom speedy output is more important than art, thought, and originality, history provides ready-made plots and characters. A novel on Alexander the Great or Joan of Arc can be as flimsy and superficial as any schoolgirl romance. But historical themes, to which may be added prehistoric or mythical ones, have inspired the greatest novelists, as Tolstoy's *War and Peace* and Stendhal's *Charterhouse of Parma* reveal. In the 20th century, distinguished historical novels such as Arthur Koestler's *The Gladiators* (1939), Robert Graves's *I, Claudius* (1934), Zoé Oldenbourg's *Destiny of Fire* (1960), and Mary Renault's *The King Must Die* (1958) exemplify an important function of the fictional imagination—to interpret remote events in human and particular terms, to transform documentary fact, with the assistance of imaginative conjecture, into immediate sensuous and emotional experience.

There is a kind of historical novel, little more than a charade, which frequently has a popular appeal because of a common belief that the past is richer, bloodier, and more erotic than the present. Such novels, which include such immensely popular works as those of Georgette Heyer, or Baroness Orczy's Scarlet Pimpernel stories in England in the early 20th century, and *Forever Amber* (1944) by Kathleen Winsor in the United States, may use the trappings of history but, because there is no real assimilation of the past into the imagination, the result must be a mere costume ball. On the other hand, the American novelist John Barth showed in *The Sot-Weed Factor* (1960) that mock historical scholarship—preposterous events served up with parodic pomposity—could constitute a viable, and not necessarily farcical, approach to the past. Barth's history is cheerfully suspect, but his sense of historical perspective is genuine.

It is in the technical conservatism of most European historical novels that the serious student of fiction finds cause to relegate the category to a secondary place. Few practitioners of the form seem prepared to learn from any writer later than Scott, though Virginia Woolf—in *Orlando* (1928) and *Between the Acts* (1941)—made bold attempts to squeeze vast tracts of historical time into a small space and thus make them as fictionally manageable as the events of a single day. And John Dos Passos' *U.S.A.*, which can be taken as a historical study of a phase in America's development, is a reminder that experiment is not incompatible with the sweep and amplitude that great historical themes can bring to the novel.

Picaresque. In Spain, the novel about the rogue or *pícaro* was a recognized form, and such English novels as Defoe's *The Fortunate Mistress* (1724) can be regarded as picaresque in the etymological sense. But the term has come to connote as much the episodic nature of the original species as the dynamic of roguery. Fielding's *Tom Jones,* whose hero is a bastard, amoral, and very nearly gallows-meat, has been called picaresque, and the *Pickwick Papers* of Dickens—whose eponym is a respectable and even childishly ingenuous scholar—can be accommodated in the category.

The requirements for a picaresque novel are apparently length, loosely linked episodes almost complete in themselves, intrigue, fights, amorous adventure, and such optional items as stories within the main narrative, songs, poems, or moral homilies. Perhaps inevitably, with such a structure or lack of it, the driving force must come from a wild or roguish rejection of the settled bourgeois life, a desire for the open road, with adventures in inn bedrooms and meetings with questionable wanderers. In the modern period, Saul Bellow's *Adventures of Augie March* (1953) and Jack Kerouac's *Dharma Bums* (1959) have something of the right episodic, wandering, free, questing character. But in an age that lacks the unquestioning acceptance of traditional morality against which the old picaresque heroes played out their villainous lives, it is not easy to revive the *novela picaresca* as the anonymous author of *Lazarillo de Tormes* (1554) conceived it, or as such lesser Spanish writers of the beginning of the 17th century as Mateo Alemán, Vicente Espinel, and Luis Vélez de Guevara developed it. The modern criminal wars with the police rather than with society, and his career is one of closed and narrow techniques, not compatible with the gay abandon of the true *pícaro.*

Sentimental. The term sentimental, in its mid-18th-century usage, signified refined or elevated feeling, and it is in this sense that it must be understood in Laurence Sterne's *Sentimental Journey* (1768). Richardson's *Pamela* (1740) and Rousseau's *Nouvelle Héloïse* (1761) are sentimental in that they exhibit a passionate attachment between the sexes that rises above the merely physical. The vogue of the sentimental love novel was one of the features of the Romantic movement, and the form maintained a certain moving dignity despite a tendency to excessive emotional posturing. The germs of mawkishness are clearly present in Sterne's *Tristram Shandy* (1760–67), though offset by a diluted Rabelaisianism and a certain cerebral quality. The debasement by which the term sentimental came to denote a self-indulgence in superficial emotions occurred in the Victorian era, under the influence of sanctimony, religiosity, and a large commercial demand for bourgeois fiction. Sentimental novels of the 19th and 20th centuries are characterized by an invertebrate emotionalism and a deliberately lachrymal appeal. Neither Dickens nor Thackeray was immune to the temptations of sentimentality—as is instanced by their treatment of deathbed scenes. The reported death of Tiny Tim in *A Christmas Carol* (1843) is an example of Dickens' ability to provoke two tearful responses from the one situation—one of sorrow at a young death, the other of relief at the discovery that the death never occurred. Despite such patches of emotional excess, Dickens cannot really be termed a sentimental novelist. Such a designation must be reserved for writers like Mrs. Henry Wood, the author of *East Lynne* (1861). That the sentimental novel is capable of appeal even in the Atomic Age is shown by the success of *Love Story* (1970), by Erich Segal. That this is the work of a Yale professor of classics seems to indicate either that not even intellectuals disdain sentimental appeal or that tearjerking is a process to be indulged in coldly and even cynically. Stock emotions are always easily aroused through stock devices, but both the aim and the technique are generally eschewed by serious writers.

Gothic. The first Gothic fiction appeared with works like Horace Walpole's *Castle of Otranto* (1765) and Matthew Gregory Lewis' *Monk* (1796), which countered 18th-century "rationalism" with scenes of mystery, horror, and wonder. Gothic (the spelling "Gothick" better conveys the contemporary flavour) was a designation derived from architecture, and it carried—in opposition to the Italianate style of neoclassical building more appropriate to the Augustan Age—connotations of rough and primitive grandeur. The atmosphere of a Gothic novel was expected to be dark, tempestuous, ghostly, full of madness, outrage, superstition, and the spirit of revenge. Mary Shelley's *Frankenstein,* which maintains its original popularity and even notoriety, has in overplus the traditional Gothic ingredients, with its weird God-defying experiments, its eldritch shrieks, and, above all, its monster. Edgar Allan Poe developed the Gothic style brilliantly in the United States, and he has been a considerable influence. A good deal of early science fiction, like H.G. Wells's *Island of Doctor Moreau* (1896), seems to spring out of the Gothic movement, and the Gothic atmosphere has been seriously cultivated in England in the later novels of Iris Murdoch and in the Gormenghast sequence beginning in 1946 of Mervyn Peake. It is noteworthy that Gothic fiction has always been approached in a spirit of deliberate suspension of the normal canons of taste. Like a circus trick, a piece of Gothic fiction asks to be considered as ingenious entertainment; the pity and terror are not aspects of a cathartic process but transient emotions to be, somewhat perversely, enjoyed for their own sake.

Ingredients of the typical Gothic novel

Psychological. The psychological novel first appeared in 17th-century France, with Madame de La Fayette's *Princesse de Clèves* (1678), and the category was consolidated by works like the Abbé Prévost's *Manon Lescaut* (1731) in the century following. More primitive fiction had been characterized by a proliferation of action and incidental characters; the psychological novel limited it-

self to a few characters whose motives for action could be examined and analyzed. In England, the psychological novel did not appear until the Victorian era, when George Eliot became its first great exponent. It has been assumed since then that the serious novelist's prime concern is the workings of the human mind, and hence much of the greatest fiction must be termed psychological. Dostoyevsky's *Crime and Punishment* deals less with the ethical significance of a murder than with the soul of the murderer; Flaubert's interest in Emma Bovary has less to do with the consequences of her mode of life in terms of nemesic logic than with the patterns of her mind; in *Anna Karenina,* Tolstoy presents a large-scale obsessive study of feminine psychology that is almost excruciating in its relentless probing. The novels of Henry James are psychological in that the crucial events occur in the souls of the protagonists, and it was perhaps James more than any serious novelist before or since who convinced frivolous novel-readers that the "psychological approach" guarantees a lack of action and excitement.

The movement away from Freud

The theories of Sigmund Freud are credited as the source of the psychoanalytical novel. Freud was anticipated, however, by Shakespeare (in, for example, his treatment of Lady Macbeth's somnambulistic guilt). Two 20th-century novelists of great psychological insight—Joyce and Nabokov—professed a disdain for Freud. To write a novel with close attention to the Freudian or Jungian techniques of analysis does not necessarily produce new prodigies of psychological revelation; Oedipus and Electra complexes have become commonplaces of superficial novels and films. The great disclosures about human motivation have been achieved more by the intuition and introspection of novelists and dramatists than by the more systematic work of the clinicians.

The novel of manners. To make fiction out of the observation of social behaviour is sometimes regarded as less worthy than to produce novels that excavate the human mind. And yet the social gestures known as manners, however superficial they appear to be, are indices of a collective soul and merit the close attention of the novelist and reader alike. The works of Jane Austen concern themselves almost exclusively with the social surface of a fairly narrow world, and yet she has never been accused of a lack of profundity. A society in which behaviour is codified, language restricted to impersonal formulas, and the expression of feeling muted, is the province of the novel of manners, and such fiction may be produced as readily in the 20th century as in the era of Fanny Burney or Jane Austen. Such novels as Evelyn Waugh's *Handful of Dust* (1934) depend on the exact notation of the manners of a closed society, and personal tragedies are a mere temporary disturbance of collective order. Even Waugh's trilogy *Sword of Honour* is as much concerned with the minutiae of surface behaviour in an army, a very closed society, as with the causes for which that army fights. H.H. Munro ("Saki"), in *The Unbearable Bassington* (1912), an exquisite novel of manners, says more of the nature of Edwardian society than many a more earnest work. It is conceivable that one of the novelist's duties to posterity is to inform it of the surface quality of the society that produced him; the great psychological profundities are eternal, manners are ephemeral and have to be caught. Finally, the novel of manners may be taken as an artistic symbol of a social order that feels itself to be secure.

Epistolary. The novels of Samuel Richardson arose out of his pedagogic vocation, which arose out of his trade of printer—the compilation of manuals of letter-writing technique for young ladies. His age regarded letter writing as an art on which could be expended the literary care appropriate to the essay or to fiction, and, for Richardson, the creation of epistolary novels entailed a mere step from the actual world into that of the imagination. His *Pamela* (1740) and *Clarissa* (1748) won phenomenal success and were imitated all over Europe, and the epistolary novel—with its free outpouring of the heart—was an aspect of early romanticism. In the 19th century, when the letter-writing art had not yet fallen into desuetude, it was possible for Wilkie Collins to tell the mystery story of *The Moonstone* (1868) in the form of an exchange of letters, but it would be hard to conceive of a detective novel using such a device in the 20th century, when the well-wrought letter is considered artificial. Attempts to revive the form have not been successful, and Christopher Isherwood's *Meeting by the River* (1967), which has a profoundly serious theme of religious conversion, seems to fail because of the excessive informality and chattiness of the letters in which the story is told. The 20th century's substitute for the long letter is the transcribed tape recording—more, as Beckett's play *Krapp's Last Tape* indicates, a device for expressing alienation than a tool of dialectic. But it shares with the Richardsonian epistle the power of seeming to grant direct communication with a fictional character, with no apparent intervention on the part of the true author.

Pastoral. Fiction that presents rural life as an idyllic condition, with exquisitely clean shepherdesses and sheep immune to foot-rot, is of very ancient descent. Longus' *Daphnis and Chloe,* written in Greek in the 2nd or 3rd century AD, was the remote progenitor of such Elizabethan pastoral romances as Sir Philip Sidney's *Arcadia* (1590) and Thomas Lodge's *Rosalynde* (1590), the source book for Shakespeare's *As You Like It.* The *Paul et Virginie* of Bernardin de St. Pierre (1787), which was immensely popular in its day, seems to spring less from the pastoral utopian convention than from the dawning Romanticism that saw in a state of nature only goodness and innocence. Still, the image of a rural Eden is a persistent one in Western culture, whatever the philosophy behind it, and there are elements of this vision even in D.H. Lawrence's *Rainbow* (1915) and, however improbable this may seem, in his *Lady Chatterley's Lover* (1928). The more realistic and ironic pictures of the pastoral life, with poverty and pig dung, beginning with George Crabbe's late-18th-century narrative poems, continuing in George Eliot, reaching sour fruition in Thomas Hardy, are usually the work of people who know the country well, while the rural idyll is properly a townsman's dream. The increasing stresses of urban life make the country vision a theme still available to serious fiction, as even a work as sophisticated as Saul Bellow's *Herzog* (1964) seems to show. But, since Stella Gibbons' satire *Cold Comfort Farm* (1932), it has been difficult for any British novelist to take seriously pastoral lyricism.

Realism in the pastoral novel

Apprenticeship. The *Bildungsroman,* or novel about upbringing and education, seems to have its beginnings in Goethe's work, *Wilhelm Meisters Lehrjahre* (1796), which is about the processes by which a sensitive soul discovers its identity and its role in the big world. A story of the emergence of a personality and a talent, with its implicit motifs of struggle, conflict, suffering, and success, has an inevitable appeal for the novelist; many first novels are autobiographical and attempt to generalize the author's own adolescent experiences into a kind of universal symbol of the growing and learning processes. Charles Dickens embodies a whole *Bildungsroman* in works like *David Copperfield* (1850) and *Great Expectations* (1861), but allows the emerged ego of the hero to be absorbed into the adult world, so that he is the character that is least remembered. H.G. Wells, influenced by Dickens but vitally concerned with education because of his commitment to socialist or utopian programs, looks at the agonies of the growing process from the viewpoint of an achieved utopia in *The Dream* (1924) and, in *Joan and Peter* (1918), concentrates on the search for the right modes of apprenticeship to the complexities of modern life.

The school story established itself in England as a form capable of popularization in children's magazines, chiefly because of the glamour of elite systems of education as first shown in Thomas Hughes's *Tom Brown's School Days* (1857), which is set at Rugby. In France, *Le Grand Meaulnes* (1913) of Alain-Fournier is the great exemplar of the school novel. The studies of struggling youth presented by Hermann Hesse became, after his death in 1962, part of an American campus cult indicating the desire of the serious young to find literary symbols for their own growing problems.

Samuel Butler's *Way of All Flesh,* which was written by 1885 but not published until 1903, remains one of the

greatest examples of the modern *Bildungsroman;* philosophical and polemic as well as moving and comic, it presents the struggle of a growing soul to further, all unconsciously, the aims of evolution, and is a devastating indictment of Victorian paternal tyranny. But probably James Joyce's *Portrait of the Artist as a Young Man* (1916), which portrays the struggle of the nascent artistic temperament to overcome the represssions of family, state, and church, is the unsurpassable model of the form in the 20th century. That the learning novel may go beyond what is narrowly regarded as education is shown in two remarkable works of the 1950s—William Golding's *Lord of the Flies* (1955), which deals with the discovery of evil by a group of shipwrecked middle class boys brought up in the liberal tradition, and J.D. Salinger's *Catcher in the Rye* (1951), which concerns the attempts of an adolescent American to come to terms with the adult world in a series of brief encounters, ending with his failure and his ensuing mental illness.

Roman à clef. Real, as opposed to imaginary, human life provides so much ready-made material for the novelist that it is not surprising to find in many novels a mere thinly disguised and minimally reorganized representation of actuality. When, for the fullest appreciation of a work of fiction, it is necessary for the reader to consult the real-life personages and events that inspired it, then the work is a *roman à clef,* or novel that needs a key. In a general sense, every work of literary art requires a key or clue to the artist's preoccupations (the jail in Dickens; the mysterious tyrants in Kafka, both leading back to the author's own father), but the true *roman à clef* is more particular in its disguised references. Chaucer's "Nun's Priest's Tale" has puzzling naturalistic details that can be cleared up only by referring the poem to an assassination plot in which the Earl of Bolingbroke was involved. Swift's *Tale of a Tub* (1704), Dryden's *Absalom and Achitophel* (1681), and Orwell's *Animal Farm* (1945) make total sense only when their hidden historical content is disclosed. These, of course, are not true novels, but they serve to indicate a literary purpose that is not primarily aesthetic. Lawrence's *Aaron's Rod* requires a knowledge of the author's personal enmities, and to understand Aldous Huxley's *Point Counter Point* fully one must know, for instance, that the character of Mark Rampion is D.H. Lawrence himself and that of Denis Burlap is the critic John Middleton Murry. Proust's *À la recherche du temps perdu* becomes a richer literary experience when the author's social milieu is explored, and Joyce's *Finnegans Wake* has so many personal references that it may be called the most massive *roman à clef* ever written. The more important the *clef* becomes to full understanding, the closer the work has come to a special kind of didacticism. When it is dangerous to expose the truth directly, then the novel or narrative poem may present it obliquely. But the ultimate vitality of the work will depend on those elements in it that require no key.

Anti-novel. The movement away from the traditional novel form in France in the form of the *nouveau roman* tends to an ideal that may be called the anti-novel—a work of the fictional imagination that ignores such properties as plot, dialogue, human interest. It is impossible, however, for a human creator to create a work of art that is completely inhuman. Contemporary French writers like Alain Robbe-Grillet in *Jealousy* (1957), Nathalie Sarraute in *Tropisms* (1939) and *The Planetarium* (1959), and Michel Butor in *Passing Time* (1957) and *Degrees* (1960) wish mainly to remove the pathetic fallacy from fiction, in which the universe, which is indifferent to man, is made to throw back radar reflections of man's own emotions. Individual character is not important, and consciousness dissolves into sheer "perception." Even time is reversible, since perceptions have nothing to do with chronology, and, as Butor's *Passing Time* shows, memories can be lived backward in this sort of novel. Ultimately, the very appearance of the novel—traditionally a model of the temporal treadmill—must change; it will not be obligatory to start at page 1 and work through to the end; a novel can be entered at any point, like an encyclopaedia.

The two terms most heard in connection with the French anti-novel are *chosisme* and *tropisme.* The first, with which Robbe-Grillet is chiefly associated, relates to the novelist's concern with things in themselves, not things as human symbols or metaphors. The second, which provided a title for Nathalie Sarraute's early novel, denotes the response of the human mind to external stimuli—a response that is general and unmodified by the apparatus of "character." It is things, the furniture of the universe, that are particular and variable; the multiplicity of human observers melts into an undifferentiable mode of response. Needless to say, there is nothing new in this epistemology as applied to the novel. It is present in Laurence Sterne (in whom French novelists have always been interested), as also in Virginia Woolf.

Chosisme and tropisme

Such British practitioners of the anti-novel as Christine Brooke-Rose and Rayner Heppenstall (both French scholars, incidentally) are more empirical than their French counterparts. They object mainly to the falsification of the external world that was imposed on the traditional novel by the exigencies of plot and character, and they insist on notating the minutiae of the surface of life, concentrating in an unhurried fashion on every detail of its texture. A work like Heppenstall's *Connecting Door* (1962), in which the narrator-hero does not even possess a name, is totally unconcerned with action but very interested in buildings, streets, and the sound of music. This is properly a fresh approach to the materials of the traditional novel rather than a total liberation from it. Such innovations as are found in the *nouveau roman* can best show their value in their influence on traditional novelists, who may be persuaded to observe more closely and be wary of the seductions of swift action, contrived relationships, and neat resolutions.

Cult, or coterie, novels. The novel, unlike the poem, is a commercial commodity, and it lends itself less than the materials of literary magazines to that specialized appeal called coterie, intellectual or elitist. It sometimes happens that books directed at highly cultivated audiences—like *Ulysses, Finnegans Wake,* and Djuna Barnes's *Nightwood* (1936)—achieve a wider response, sometimes because of their daring in the exploitation of sex or obscenity, more often because of a vitality shared with more demotic fiction. The duplicated typescript or the subsidized periodical, rather than the commercially produced book, is the communication medium for the truly hermetic novel.

The novel that achieves commercial publication but whose limited appeal precludes large financial success can frequently become the object of cult adulation. In the period since World War II, especially in the United States, such cults can have large memberships. The cultists are usually students (who, in an era of mass education, form a sizable percentage of the total population of the United States), or fringes of youth sharing the student ethos, and the novels chosen for cult devotion relate to the social or philosophical needs of the readers. The fairy stories of Tolkien, *The Lord of the Flies* of Golding, the science fiction of Kurt Vonnegut, Jr., have, for a greater or lesser time, satisfied a hunger for myth, symbols, and heterodox ideas, to be replaced with surprising speed by other books. The George Orwell cult among the young was followed by a bitter reaction against Orwell's own alleged reactionary tendencies, and such a violent cycle of adoration and detestation is typical of literary cults. Adult cultists tend, like young ones, to be centred in universities, from which they circulate newsletters on *Finnegans Wake,* Anthony Powell's *Music of Time* sequence, and the works of Evelyn Waugh. Occasionally new public attention becomes focussed on a neglected author through his being chosen as a cult object. This happened when the novellas of Ronald Firbank, the anonymous comic novel *Augustus Carp, Esq.,* and G.V. Desani's *All About Mr. Hatterr* got back into print because of the urging of minority devotees. Despite attempts to woo a larger public to read it, Malcolm Lowry's *Under the Volcano* obstinately remained a cult book, while the cultists performed their office of keeping the work alive until such time as popular taste should become sufficiently enlightened to appreciate it.

Student cultists

Detective, mystery, thriller. The terms detective story, mystery, and thriller tend to be employed interchangeably. The detective story thrills the reader with mysterious crimes, usually of a violent nature, and puzzles his reason

until their motivation and their perpetrator are, through some triumph of logic, uncovered. The detective story and mystery are in fact synonymous, but the thriller frequently purveys adventurous *frissons* without mysteries, like the spy stories of Ian Fleming, for example, but not like the spy stories of Len Deighton, which have a bracing element of mystery and detection. The detective novel began as a respectable branch of literature with works like Poe's "Murders in the Rue Morgue" (1841), Dickens' unfinished *Edwin Drood* (1870), and Wilkie Collins' *Moonstone* (1868) and *Woman in White* (1860). With the coming of the Sherlock Holmes stories of Sir Arthur Conan Doyle, at the beginning of the 20th century, the form became a kind of infraliterary subspecies, despite the intellectual brilliance of Holmes's detective work and the high literacy of Doyle's writing. Literary men like G.K. Chesterton practiced the form on the margin, and dons read thrillers furtively or composed them pseudonymously (*e.g.*, J.I.M. Stewart, reader in English literature at Oxford, wrote as "Michael Innes"). Even the British poet laureate, C. Day Lewis, subsidized his verse through writing detective novels as "Nicholas Blake." Dorothy L. Sayers, another Oxford scholar, appeared to atone for a highly successful career as a mystery writer by turning to religious drama and the translating of Dante, as well as by making her last mystery novel—*Gaudy Night* (1935)—a highly literary, even pedantic, confection.

Such practitioners as Agatha Christie, Ellery Queen, Erle Stanley Gardner, Raymond Chandler, to say nothing of the highly commercial Edgar Wallace and Mickey Spillane, have given much pleasure and offended only the most exalted literary canons. The fearless and intelligent amateur detective, or private investigator, or police officer has become a typical hero of the modern age. And those qualities that good mystery or thriller writing calls for are not to be despised, since they include economy, skillful sustention of suspense, and very artful plotting.

The mystery novel was superseded in popularity by the novel of espionage, which achieved a large vogue with the James Bond series of Ian Fleming. Something of its spirit, if not its sadism and eroticism, had already appeared in books like John Buchan's *Thirty-nine Steps* and the "entertainments" of Graham Greene, as well as in the admirable novels of intrigue written by Eric Ambler. Fleming had numerous imitators, as well as a more than worthy successor in Len Deighton. The novels of John Le Carré found a wide audience despite their emphasis on the less glamorous, often even squalid aspects of international espionage; his works include *The Spy Who Came in from the Cold* (1963) and *Smiley's People* (1980).

Western. Man's concern with taming wild land, or advancing frontiers, or finding therapy in reversion from the civilized life to the atavistic is well reflected in adventure novels, beginning with James Fenimore Cooper's novels of the American frontier *The Pioneers* (1823) and *The Last of the Mohicans* (1826). As the 19th century advanced, and new tracts of America were opened up, a large body of fiction came out of the men who were involved in pioneering adventure. Mark Twain's *Roughing It* (1872) may be called a frontier classic. Bret Harte wrote shorter fiction, like "The Luck of Roaring Camp" (1868), but helped to spread an interest in frontier writing to Europe, where the cult of what may be termed the western novel is as powerful as in America. Owen Wisters' *Virginian* (1902), Andy Adams' near-documentary *Log of a Cowboy* (1903), Emerson Hough's *Covered Wagon* (1922), from which the first important western film was made in 1923, Hamlin Garland's *Son of the Middle Border* (1917), and O.E. Rölvaag's *Giants in the Earth* (1927) all helped to make the form popular, but it is to Zane Grey—who wrote more than 50 western novels —that lovers of frontier myth have accorded the greatest devotion. The western is now thought of predominantly as a cinematic form, but it arose out of literature. Other frontier fiction has come from another New World, the antipodes—South Africa as well as the Austrialian outback—but the American West has provided the best mythology, and it is still capable of literary treatment. Sophisticated literary devices may be grafted onto the western—surrealistic fantasy or parallels to Shakespeare or to the ancient classics—but the peculiar and perennial appeal of the western lies in its ethical simplicity, the frequent violence, the desperate attempt to maintain minimal civilized order, as well as the stark, near-epic figures from true western history, such as Billy the Kid, Calamity Jane, Wyatt Earp, Annie Oakley, and Jesse James.

The enduring and the ephemeral

The best-seller. A distinction should be made between novels whose high sales are an accolade bestowed on literary merit and novels that aim less at aesthetic worth than at profits. The works of Charles Dickens were best-sellers in their day, but good sales continue, testifying to a vitality that was not purely ephemeral. On the other hand, many best-selling novels have a vogue that is destined not to outlast the time when they were produced. It is a characteristic of this kind of best-seller that the writing is less interesting than the content, and that the content itself has a kind of journalistic oversimplification that appeals to unsophisticated minds. The United States is the primary home of the commercial novel whose high sales accrue from careful, and sometimes cold-blooded, planning. A novel in which a topical subject—such as the Mafia, or corruption in government, or the election of a new pope, or a spate of aircraft accidents, or the censorship of an erotic book—is treated with factual thoroughness, garnished with sex, enlivened by quarrels, fights, and marital infidelities, presented in nonliterary prose, and given lavish promotion by its publisher may well become a best-seller. It is also likely to be almost entirely forgotten a year or so after its publication. The factual element in the novel seems to be necessary to make the reader feel that he is being educated as well as diverted. Indeed, the conditions for the highest sales seem to include the reconciliation of the pornographic and the didactic.

A novel with genuine aesthetic vitality often sells more than the most vaunted best-seller, but the sales are more likely to be spread over decades and even centuries rather than mere weeks and months. The author of such a book may, in time, enrich others, but he is unlikely himself to attain the opulence of writers of best-sellers such as Harold Robbins or Irving Wallace.

Fantasy and prophecy. The term science fiction is a loose one, and it is often made to include fantastic and prophetic books that make no reference to the potentialities of science and technology for changing human life. Nevertheless, a novel like Keith Roberts' *Pavane* (1969), which has as a premise the conquest of England by Spain in 1588, and the consequent suppression rather than development of free Protestant intellectual inquiry, is called science fiction, though such terms as "fiction of hypothesis" and "time fantasy" would be more fitting. The imaginative novelist is entitled to remake the existing world or present possible future worlds, and a large corpus of fiction devoted to such speculative visions has been produced in the last hundred years, more of it based on metaphysical hypotheses than on scientific marvels. Jules Verne and H.G. Wells pioneered what may be properly termed science fiction, mainly to an end of diversion. Since the days of Wells's *Time Machine* (1895) and *Invisible Man* (1897), the fiction of hypothesis has frequently had a strong didactic aim, often concerned with opposing the very utopianism that Wells—mainly in his nonfictional works—built on the potentialities of socialism and technology. Aldous Huxley's *Brave New World* (1932) showed how dangerous utopianism could be, since the desire for social stability might condone conditioning techniques that would destroy the fundamental human right to make free choices. Toward the end of his life Huxley produced a cautious utopian vision in *Island* (1962), but the dystopian horrors of his earlier novel and of his *Ape and Essence* (1948) remain more convincing. Orwell's *Nineteen Eighty-four* (1949) showed a world in which a tyrannic unity is imposed by a collective solipsism, and contradictions are liquidated through the constant revision of history that the controlling party decrees. Anthony Burgess' *Clockwork Orange* (1962) and *Wanting Seed* (1962) portray ghastly futures that extrapolate, respectively, philosophies of crime control and population control out of present-day tendencies that are only potentially dangerous.

A large number of writers practice prophetic fantasy with considerable literary skill and careful factual preparation—Kurt Vonnegut, Jr., Ray Bradbury, Italo Calvino, Isaac Asimov, J.G. Ballard, to name a few—and novelists whose distinction lies mainly in more traditional fields have attempted the occasional piece of future-fiction, like L.P. Hartley with his *Facial Justice* (1961) and Evelyn Waugh in *Love Among the Ruins* (1953). The fantasist who fantasizes without prophetic or warning intent is rarer, but works like Nabokov's *Ada,* Tolkien's *Lord of the Rings* cycle, and Christine Brooke-Rose's *Out* (1964) represent legitimate and heartening stretching of the imagination, assurances that the novelist has the right to create worlds, as well as characters, of his own. But the dystopian novel can have a salutary influence on society, actively correcting regressive or illiberal tendencies, and *Brave New World* and *Nineteen Eighty-four* can be cherished as great didactic landmarks, not just as works of literary art.

Proletarian. The novel that, like Dickens' *Hard Times* (1854), presents the lives of workingmen or other members of the lower orders is not necessarily an example of proletarian fiction. The category properly springs out of direct experience of proletarian life and is not available to writers whose background is bourgeois or aristocratic. Thus, William Godwin's *Caleb Williams* (1794), or Robert Bage's *Hermsprong* (1796), though, like *Hard Times,* sympathetic to the lot of the oppressed worker, is more concerned with the imposition of reform from above than with revolution from within, and the proletarian novel is essentially an intended device of revolution. The Russian Maksim Gorky, with works like *Foma Gordeyev* (1900) and *Mother* (1907), as well as many short stories portraying the bitterness of poverty and unemployment (the pseudonym *Gorky* means "bitter"), may be taken as an exemplary proletarian writer. The United States has produced a rich crop of working-class fiction. Such Socialist writers as Jack London, Upton Sinclair, John Dos Passos, and Edward Dahlberg, however, did not witness the triumph of the workers' revolution in their own country, as Gorky did in his, and it is the fate of the American proletarian novelist, through literary success, either to join the class he once dreamed of overthrowing or to become anarchic and frustrated. In the Soviet Union the proletarian novel was doomed to disappear in the form that Gorky knew; a present-day Gorky would presumably write, in exile, novels about Soviet oppression or else join Yuly Daniel (who writes under the pseudonym Nikolay Arzhak) and Aleksandr Solzhenitsyn in enforced silence and disgrace. For it is the essence of the revolutionary novel only to possess vitality and validity when written under capitalist "tyranny." England has produced its share of working-class novelists exuding bitterness, such as Alan Sillitoe, with his *Saturday Night and Sunday Morning* (1958), but conditions apt for revolution have not existed in Britain for over a century. British novelists who emerged after World War II, such as John Braine (*Room at the Top*), Keith Waterhouse (*There Is a Happy Land*), Kingsley Amis (*Lucky Jim*), and Stan Barstow (*A Kind of Loving*), showed a solution to working-class frustration in a fluid system of class promotion: revolution is an inadmissible dream. Generally speaking, in the novel, which is preoccupied with individuals rather than with groups, it is difficult to make the generalized political statements that are meat and drink to the revolutionary propagandist.

Reform versus revolution

Other types. The categories briefly discussed above are among the most common fictional forms. Theoretically there is no limit to the number available, since changing social patterns provide fresh subjects and fresh taxonomies, and new metaphysical and psychological doctrines may beget new fictional approaches to both content and technique. Other categories of fictional art include the erotic novel (which may or may not be pornographic), the satirical novel, the farcical novel, the novel for or about children, the theological novel, the allegorical novel, and so on. Types of fiction no longer practiced, since their real-life referents no longer exist, include the colonial novel—such as E.M. Forster's *Passage to India* (1924), Henri Fauconnier's *Malaisie* (1930), and the African sequence of Joyce Cary—and space fantasy like H.G. Wells's *First Men in the Moon* (1901). One may read examples of a departed category with pleasure and profit, but the category can no longer yield more than parody or pastiche. New kinds of fiction fill in the gaps, like the novel of negritude, the structuralist novel (following the linguistic sociologists and anthropologists), the homosexual novel, the novel of drug hallucination, and so on. So long as human society continues to exist, the novel will exist as its mirror, an infinitude of artistic images reflecting an infinitude of life patterns.

Outmoded types

THE NOVEL IN ENGLISH

The United Kingdom. England's chief literary achievements lie in the fields of drama and poetry, and the attitude of English novelists to their form was, for a long time, cheerfully empirical and even amateurish. Elizabethan novels, *novelle* rather, imitated the Spanish picaresque story, and Thomas Nashe's *Unfortunate Traveller* (1594) is a good, bustling, vital example of a rapidly composed commercial work more concerned with sensational incident and language than with shape or character. Daniel Defoe (1660?–1731) is often considered to be the true progenitor of the long English novel, but his *Robinson Crusoe* and *Moll Flanders* are loosely constructed, highly episodic, and presented as mock-biography rather than real fiction. It is with *Pamela,* by Samuel Richardson (1689–1761), that the tradition of serious, moral fiction in English may be said to begin, but later 18th-century novelists reverted to the picaresque and comic. Henry Fielding (1707–54) wrote his first novel, *Joseph Andrews,* as a parody on *Pamela,* but his masterpiece, *Tom Jones,* is original if shapeless, an example of English literary genius sinking thankfully back into the casual and improvisatory. Laurence Sterne (1713–68) produced a great mad work, *Tristram Shandy,* that, in its refusal to truckle to any rules of structure, remains still a quarry for avant-garde novelists; and Tobias Smollett (1721–71) wrote picaresque satire in *Roderick Random* and *Peregrine Pickele,* full-blooded portraits of the age that impress more with their vigour than with their art.

The Romantic Age brought, paradoxically, the cool and classically shaped novels of Jane Austen (1775–1817), a major practitioner and still a model for apprentices in the craft. Sir Walter Scott (1771–1832), a Scotsman who wrote about romantic historical Scotland, must be regarded as an international figure whose influence was greater even than Richardson's, since he more than anyone established the historical novel as the primary fictional form in Europe. In his work, nevertheless, the traditional faults of the British novel may be descried—episodic formlessness, an ebullience of texture rather than a clean narrative line. The Victorian Age moved out of the romantic past, or, as with William Makepeace Thackeray (1811–63), stayed with it only to deromanticize it. Charles Dickens (1812–70) was indebted to the picaresque tradition but turned reformist eyes on his own age. Thackeray and Dickens are complementary in that the first attacks the upper classes while the second showers sympathy, sometimes of a mawkish kind, on the lower classes. With George Eliot (1819–80), the first true English psychological novels appear, strong in their moral content, and George Meredith (1828–1909) may be said to have anticipated—in *The Ordeal of Richard Feverel* and *The Egoist*—the approach in depth that characterized the psychological novel of the 20th century.

The Romantic and Victorian ages

Both Charlotte Brontë (1816–55) and Emily Brontë (1818–48) exemplify the capacity of the English novel to achieve solitary "sports" unrelated to any current or tradition. Both *Wuthering Heights*—a superb evocation of the soul of a locality, with a love story that is fierce and primitive but recounted with poetic sophistication—and *Jane Eyre,* an exceedingly frank and still shocking study of a love that rides over Victorian conventions, are unlike any other books of their time, or of any other time, though their qualities have been diluted into hundreds of popular 20th-century romances. Later Victorians, particularly Samuel Butler (1835–1902) and Thomas Hardy (1840–1928), reflected those changes in the educated English sen-

sibility that had been brought about by the new science. Hardy's world is one in which the Christian God has been replaced by a malevolent Providence—the poet-novelist's theologization of scientific determinism. Butler's *Way of All Flesh,* a work that contrives to be both bitterly realistic and highly-comic, demonstrates the working of Darwinian evolution in social institutions such as the family and even the church. In many ways, Butler led English fiction into the modern age.

John Galsworthy (1867–1933) showed himself interested in the processes by which old institutions, such as a great Whig family, decay as history advances, but in both style and characterization he is firmly set in the Victorian Age. Of Arnold Bennett (1867–1931) it may be said that he brought to a kind of Thackerayan social realism something of the spirit of the French novel, particularly the anglicized tones of Balzac and Zola. W. Somerset Maugham (1874–1965), though his most considerable achievements lie in the short story form, wrote novels, like *Of Human Bondage,* that are infused with the delicious acerbities of French naturalism, and all his work, long or short, is exalted by the influence of Guy de Maupassant. Early 20th-century British fiction needed the impact of an alien tradition to jolt it out of bourgeois empiricism, and perhaps the two major influences were both foreigners who had elected to write fiction in England—Joseph Conrad (1857–1924), Polish-born, to whom English was a second language, and Henry James (1843–1916), an American who had drunk deeply at French fountains and brought to the exercise of his craft a scrupulousness and a concern with aesthetic values that was almost obsessive and, it may be said, very un-English. James's influence on such major English-born novelists as Virginia Woolf (1882–1941), E.M. Forster (1879–1970), and Ford Madox Ford (1873–1939) was in the direction of that concern with style which the English novel, unlike the French, has always tried to resist, and this influence remains a potent one on succeeding generations of novelists, not only in England.

Other important Englishmen have remained more interested in content than in style, though in Graham Greene, a Catholic convert, who is primarily known for his "theological" content, there is a very interesting attempt to bring to this a Conradian concern for the solitary-man theme and a Jamesian preoccupation with style. D.H. Lawrence (1885–1930) remains the great modern English novelist who reconciles a highly traditional style (in *The Rainbow,* for example, he often resembles George Eliot) with a subject matter that is revolutionary in the profundity of its human relationships. And Aldous Huxley (1894–1963), though he indulged in formal experiment in *Point Counter Point* and *Eyeless in Gaza,* was always essentially a didactic writer who used the novel form somewhat casually. The same may be said of George Orwell (1903–50), who eschewed stylistic complexity in the interests of a clear message. On the other hand, Evelyn Waugh (1902–66) and Anthony Powell, in a cautious and conservative way, consulted the claims of allusive and evocative prose, as did Wyndham Lewis (1884–1957), who also brought to his novels the aesthetic of an alien art—that of painting.

Novelists like Kingsley Amis and Angus Wilson are seen to belong to traditions already old in the time of Samuel Butler. Amis derives from Fielding (Amis' *Take a Girl Like You* has a moral quality reminiscent of Fielding's *Amelia*), and Wilson never pretends to be other than a disciple of novelists like George Eliot and Dickens. The Victorian kind of novel, as practiced by writers like J.B. Priestley, satisfies large numbers of British readers, and Jamesian scrupulosity and Joycean experiment alike are regarded with amiable suspicion. Such Englishmen as have experimented in fictional technique have usually found a larger audience abroad than at home. Thus Lawrence Durrell found his *Alexandria Quartet* (1962) hailed as a masterpiece in France, while English readers merely liked or disliked it. *A Clockwork Orange* (1962) by Anthony Burgess achieved a large readership in the United States, but it recorded little positive response in the country of its origin.

The fiction of the British Isles is not clearly differentiated into national or regional groups, and to speak of Irish novelists is often to do no more than refer to an irrelevant place of birth. Oscar Wilde (1854–1900) was more French than Irish; his one novel, *The Picture of Dorian Gray,* has no progenitors in English fiction, but it owes much to the late 19th-century French novelist Joris-Karl Huysmans. George Moore (1852–1933) was born in County Mayo, Ireland, but studied art in Paris, and works like *A Modern Lover* and *A Mummer's Wife* were intended to teach the British novel how to absorb the spirit of Flaubert and Zola. Moore's *Lake,* written under the influence of the Irish literary revival, is dutifully set in Ireland, but this essentially international writer was quick to move back to France for *Héloïse and Abélard* and on to Palestine for *The Brook Kerith.* James Joyce (1882–1941), conceivably the greatest novelist in English in the 20th century, never moves from his native Dublin in any of his fiction, but no less parochial writer can well be imagined. *Ulysses,* in the depth of its historical coverage, may be regarded as one of the last great artistic monuments of the Austro-Hungarian Empire, and *Finnegans Wake* was stimulated by the avant-garde climate of Paris, the resting place of so many Irish "wild geese." It is only perhaps negatively that Ireland shaped Joyce's literary personality; the oppressiveness of the new nationalism made him react in the direction of internationalism, and it eventually forced him to a lifelong exile. Flann O'Brien (1911–66), possibly Joyce's true successor, stayed in Dublin and drew, in works like *At Swim-Two-Birds* (1939) and *The Hard Life* (1961), on the native demotic experience, but his techniques came from Europe, not from the Anglo-Irish bourgeois stockpot.

Irish and Scottish novels

The Scottish novel hardly exists as a national entity; there is nothing in fiction that matches the exploitation of Lallans in the poetry of Burns of Hugh MacDiarmid (C.M. Grieve). Scott, as has been indicated, belongs to European literature, and later novelists like Robert Louis Stevenson (1850–94) and Sir James Barrie (1860–1937), though Scottish themes and settings are featured triumphantly in their fiction, are essentially men whose literary metropolis was London rather than Edinburgh.

Wales sedulously cultivates Welsh as a living artistic medium, but few novels by Welshmen reach the English or international market, unless—like Emyr Humphreys—they write their works in English as well as Welsh. The English-language novel from Wales hardly exists, except in the form of best-sellers like Richard Llewellyn's *How Green Was My Valley* (1940)—a deliberate capitalization on Welsh picturesqueness—and the very Joycean prose of the poet Dylan Thomas (1914–53), whose one novel, *Adventures in the Skin Trade,* has very little of Wales in it.

The United States. If the American novel appears to begin with the *Wieland* and *Edgar Huntly* of Charles Brockden Brown (1771–1810), then the fiction of the mother country may be said to have no start on that of the newly independent daughter. Admittedly, Brown owes something to the English Gothic novel, but already a typically American note is struck in the choice of a violent and bizarre form that, in various mutations, was to prove fruitful in the development of American fiction. James Fenimore Cooper (1789–1851) struck out, in the pentateuch called the "Leatherstocking" tales, in another direction suitable to the American genius and experience: *The Last of the Mohicans* and *The Prairie,* though their prose style is perhaps overelaborate, are full of the spirit of a young nation confronting the wilderness and advancing its frontiers. Harriet Beecher Stowe (1811–96) produced, in *Uncle Tom's Cabin,* an antislavery novel whose sensational devices owe something to the Gothic movement but whose style is rooted in that European romantic tradition fathered by Scott. Even Edgar Allan Poe (1809–49) and Nathaniel Hawthorne (1804–64) are incompletely emancipated from Europe, though Hawthorne's *Scarlet Letter* shows a preoccupation with sin that finds no fictional counterpart in the mainstream of the European novel. America, aware of the darkness and mystery of a land still mainly undiscovered, was correspondingly aware, in both Gothic and eschatological fiction, of the dark places of the mind.

The true emancipation from Europe comes in the works of Hawthorne's friend Herman Melville (1819–91), whose

Moby Dick creats a totally American fusion by combining plain adventure with a kind of Manichaean symbolism—implying a belief that the universe is under the dominion of two opposing principles, one good and the other evil. Mark Twain (1835–1910) brought the Mississippi frontier region into the literary geography of the world with *Tom Sawyer* and *Huckleberry Finn,* combining very American humour with harsh social criticism, as also in the unique *Pudd'nhead Wilson.* From Twain on, the new regions of the United States become the materials of a wealth of fiction. Thus, while Bret Harte (1836–1902) wrote about California and George W. Cable (1844–1925) on Louisiana, Indiana was celebrated in *The Hoosier Schoolmaster* of Edward Eggleston (1837–1902) and the *Penrod* stories of Booth Tarkington (1869–1947).

American innocence and European sophistication

But America could not wholly turn its back on Europe, and the early 20th century was characterized by the discovery of the phenomenon called American "innocence" in confrontation with the decadent wisdom and sophistication of the Old World. If, in *Innocents Abroad,* Mark Twain and his fellow travellers remained unimpressed by Europe, Henry James (1843–1916) devoted millions of words to the response of impressionable Americans to the subtle deciduous culture their ancestors had abandoned in the search for a new life. Zolaesque realism entered the American novel with William Dean Howells (1837–1920). Theodore Dreiser (1871–1945) shocked his audiences with the candour and pessimism of *Sister Carrie* and *An American Tragedy,* while Stephen Crane (1871–1900) and Frank Norris (1870–1902) made realism enter, respectively, the traditionally romantic field of war (*The Red Badge of Courage*) and the pastoral life of the promised land of California (*The Octopus*). American innocence ceased to be a fictional property. The brutality in Jack London (1876–1916) joins hands with the depictions of "natural man" in the works of Ernest Hemingway (1899–1961), while the urban fiction of James T. Farrell (1904–79) and John O'Hara (1905–70) has a wholly American realism that leaves the naturalism of Zola far behind.

The 20th-century American novel is "compartmentalized" in a manner that seems to give the lie to any allegation of cultural unity. The cult of the regional novel has continued. Ohio came under the microscopic scrutiny of Sherwood Anderson (1876–1941) in *Winesburg, Ohio,* just as the small-minded hypocrisy and materialism of the Midwest fired the brilliant sequence of Sinclair Lewis (1885–1951) from *Main Street* on. But the best regional fiction has come from the South, with William Faulkner (1897–1962), whose technical master is Joyce, Ellen Glasgow (1874–1945), Flannery O'Connor (1922–64), Eudora Welty, Erskine Caldwell, and Robert Penn Warren. The immigrant communities of Nebraska were the subject of *My Antonia,* by Willa Cather (1873–1947), while O.E. Rölvaag (1876–1931) dealt with the South Dakota Norwegians in *Giants in the Earth.* The urban Jew has become the very spokesman of the contemporary American experience with writers like Saul Bellow, Herbert Gold and Bernard Malamud; while the black American has been made highly articulate in works like *Go Tell It on the Mountain* by James Baldwin, *Invisible Man* by Ralph Ellison, and *Home to Harlem* by Jamaican-born Claude McKay (1890–1948).

Meanwhile, the fiction of social criticism that derives from Howells and was made popular by men like Upton Sinclair (1878–1968) goes on, either in "muckraking" best-sellers or in such specialized attacks on the American ethos as are exemplified by war books like Joseph Heller's *Catch-22* and Kurt Vonnegut's *Slaughterhouse-Five.* American realism has gained a reputation for candour unequalled in any other literature of the world, so that the *Tropic of Cancer* and *Tropic of Capricorn* of Henry Miller (1891–1980) remain models for sexual frankness that encouraged writers like Hubert Selby, Jr. (*Last Exit to Brooklyn*), and Philip Roth (*Portnoy's Complaint*) to uncover areas of eroticism still closed to the European novel.

To balance the concern with naked subject matter, American fiction has also fulfilled the promise of its earlier expatriates beginning with Gertrude Stein (1874–1946) to match the French in a preoccupation with style. The distinction of F. Scott Fitzgerald (1896–1940) lay in the lyrical intensity of his prose as much as in his "jazz age" subject matter, and writers like Truman Capote and John Updike have dedicated themselves similarly to perfecting a prose instrument whose effects (like those of Joyce) touch the borders of poetry. Perhaps the glories and potentialities of American fiction are best summed up in the novels of Vladimir Nabokov (1899–1977). His early works belonged to Europe, but when he took to writing in English he created a sophisticated, cosmopolitan, and highly poetic style. His work, as in *Lolita* and *Pale Fire,* concentrates with unparalleled intensity on the immediacies of American life in the 20th century.

The British Commonwealth. Canada has two literatures—one in French as well as one in English—and, in the taxonomy of this article, it will be as well to forget the course of history and consider French-Canadian novelists along with those of the separated, and officially abandoned, mother country. The Canadian novel in English begins, appropriately enough, with a Richardson—John Richardson (1796–1852), author of *Wacousta,* a story of the Indian uprising led by Pontiac. But, with the exception of James DeMille (1836–80), author of a remarkable novel, *A Strange Manuscript Found in a Copper Cylinder,* and William Kirby (1817–1906), whose *Golden Dog* is an interesting long romance of 18th-century Quebec, no novelist of stature emerged in what was a great period of literary expansion in the United States. The historical novels of Sir Gilbert Parker (1862–1932), the western romances of Ralph Connor (1860–1937), and the world-famous *Anne of Green Gables* by Lucy Maud Montgomery (1874–1942) exhibit a stylistic tameness, along with that jejune oversimplification of psychology that is a mark of the deliberately "popular" novel.

More distinctive Canadian work came in the early 20th century, as in the prairie novels of Robert Stead (1880–1959) and Frederick Philip Grove (1871–1948). Morley Callaghan, a chronicler of urban life, may be seen as a writer approaching international stature, and that tough "European" realism that braces his work is also to be found, along with a rather exotic elegance, in the novels of Robertson Davies. Although some younger novelists, like Margaret Atwood and Alice Munro, deal with growing up and rites of passage in Canada, others, like Brian Moore and Mordecai Richler, have sought artistic stimulation outside Canada, and they show general uneasiness about the lack of a cultivated audience in their great but underpopulated land.

Australian and New Zealand novels

A similar complaint may be heard in Australia and New Zealand, where the fictional tradition has long accommodated itself to the simple literary needs of unsophisticated settlers. The first true Australian novel was probably *For the Term of His Natural Life,* by Marcus Clarke (1846–81), which dealt, appropriately, considering the penal origins of the Australian settlements, with life in prison; and Rolf Boldrewood, the pseudonym of Thomas A. Browne (1826–1915), had a cognate approach of excessive simplicity and melodrama to the days of the gold rush in his *Robbery Under Arms.* Australian farm life was the theme of the novels that Arthur Hoey Davis (1868–1935) wrote under the pen name of Steele Rudd, beginning with *On Our Selection.* But something that can only be termed native Australianism—which has less to do with trades and scenic backgrounds than with a whole new language and collective philosophy of life—appears for the first time in *My Brilliant Career* (1901), by Miles Franklin (1879–1954). The expression of a national personality is not, however, enough.

Progressive 20th-century Australian novelists look hungrily toward Europe, aware that native conservatism will not support the kind of experimentalism in literature that it is prepared to take for granted in painting and architecture. The most considerable modern Australian novelist is Patrick White, but even such massive achievements as *Voss* and *Riders in the Chariot* are set firmly in the Dostoyevskian tradition. Morris West is Australian only by birth; he seeks international themes and the lure of the American best-seller market. It seems that an Australian writer can succeed only if he renounces his Australianism

and goes into exile. It is still a source of humiliation to Australian litterateurs that the finest novel about Australia was written by a mere visitor—*Kangaroo* (1923), by D.H. Lawrence.

New Zealand, with its much smaller population, presents even greater problems for the native novelist seeking an audience, and this—along with New Zealand's closer tie with the mother country—explains why such fiction writers as Katherine Mansfield (1888–1923) and Dame Ngaio Marsh (1899–1982) made London their centre. Young novelists like Janet Frame, Ian Cross, and Maurice Shadbolt show, however, a heartening willingness to render the New Zealand scene in idioms and techniques more progressive than anything to be found in Australia, and they are prepared to resist the temptations of the expatriate life.

Inevitably, territories like India and Africa entered the federation of English literature very much later than those dominions founded on the English language. With few exceptions, African and Indian novelists employ English as a second language, and one of the charms of their novels lies in a creative tension between the adopted language and the native vernacular (needless to say, this is usually self-consciously exploited—often for poetic, but more frequently for comic, effect).

English as a second language of the novelist

Of Indian novelists, R.K. Narayan, in works like *The English Teacher* (U.S. title *Grateful to Life and Death*) and *The Man-Eater of Malgudi,* exhibits an individual combination of tenderness and humour, as well as a sharp eye for Indian foibles. Raja Rao, whose best known novel is *The Serpent and the Rope,* achieves remarkable prosodic effects through allowing Sanskrit rhythm and idiom to fertilize English. Of younger Indian novelists, Balachandra Rajan is notable, in *The Dark Dancer* and *Too Long in the West,* for an ability to satirize the Anglo-American way of life with the same suave elegance that informs his tragicomic view of the East. Khushwant Singh presents, in *I Shall Not Hear the Nightingale,* a powerful chronicle of Sikh life during that period of imperial dissolution that began with World War II. English seems established as the medium for the Indian novel, and it is interesting to note an ability on the part of nonnative Indian residents who have practiced the form to be absorbed into a new "Indo-Anglian" tradition. Rudyard Kipling's *Kim* is respected by Indian writers, and E.M. Forster's *Passage to India* is a progenitor of one kind of Indian novel. The novels of Paul Scott (1920–78)—such as the tetralogy *The Raj Quartet* (1976)—spring out of a love of the country and an understanding of its complexities not uncommon among former British soldiers and administrators. The Indian novel is perhaps a product of territory rather than of blood.

African novelists

The most important of the new African writers come from the West Coast, traditional fount of artists, and they are mostly characterized by immense stylistic vigour, a powerful realism, and, often, a satirical candour unseduced by the claims of the new nationalism. Chinua Achebe, in *Things Fall Apart* and *No Longer at Ease,* renders remarkably the tones of Umuaro speech and thought, and exhibits, as also in *Arrow of God,* a concern for that rich native culture whose extirpation is threatened by imported Western patterns of life and government. His *Man of the People* deals sharply with corruption and personality cult in a newly independent African state modelled on his own Nigeria. A fellow Nigerian, Amos Tutuola, has gained an international reputation with *The Palm-Wine Drinkard*—a richly humorous novel permeated with the spirit of folklore. Cyprian Ekwensi is best known for his *Jagua Nana,* a wry study of the impact of the new materialism (symbolized by the "Jagua" car of the title) on the tribal mind. Onuora Nzekwu, in *Blade Among the Boys,* has a graver theme: the conflict between Ibo religion and imported Christianity in the upbringing of a sensitive and confused young man.

South African novels have, traditionally, dealt with those pioneering themes still exemplified in the work of Laurens van der Post and Stuart Cloete, but the territory has made its entry into world literature comparatively recently chiefly because the official racist policies have inspired a powerful fiction of protest. William Plomer (1903–73) may be regarded as the father of anti-apartheid literature, although his *Turbott Wolfe* appeared in 1925, long before the state doctrine was articulated. The theme of this novel was the necessity for white and black blood to mix and ensure a liberal South African future not given over purely to white domination. Alan Paton, in *Cry, the Beloved Country,* has produced the most popular novel of protest, but Dan Jacobson, Nadine Gordimer, and especially Doris Lessing have amplified mere protest into what may be termed a kind of philosophical fiction, often distinguished enough to rank with the best work of Europe and America. Lessing's sequence *Children of Violence* finds in the South African system of government a starting point for denunciation of wrongs that turn out to be social and sexual as well as racial.

Caribbean works

The varied fictional achievements of the Caribbean are large. Trinidad has produced the two best known West Indian writers—Samuel Selvon and V.S. Naipaul, both of East Indian descent. Selvon's work (*A Brighter Sun, An Island Is a World, The Lonely Londoners*) is grim, bitter, capable of vivid evocation of the Trinidadian scene, but Naipaul, after *A House for Mr. Biswas,* has shown signs of habituation to his English exile, so that his *Mr. Stone and the Knights Companion* seems to be a novel totally nourished by the London world in which it is set. Most Caribbean novelists, finding their publishers and their audiences in England, transfer themselves thither and cut themselves off from all but remembered roots. This is true of Edgar Mittelhölzer (1909–65), an ebullient and prolific writer whose later books all had English settings. Wilson Harris has, in English exile, created an astonishing Guianan tetralogy in which poetry and myth and symbolic difficulty have a place. George Lamming and John Hearne are both notable for firm and economical prose and masterly scene painting. The aesthetic prospects for the West Indian novel seem excellent, but the absorption of its practitioners into the larger English-speaking world represents a symptom that plagues the practice of literature in so many Commonwealth countries—the lack of adequate publishing facilities and, more than that, the failure of a cultivated readership to emerge. It is undoubtedly unhealthy for an author to have to seek primary communication with foreigners.

EUROPE

Russian. The Russian novel properly begins with Nikolay Karamzin (1766–1826), who introduced into Russian literature not only the exotic sentimental romanticism best seen in his novels *Poor Lisa* and *Natalya, the Boyar's Daughter* but that large Gallic vocabulary that was to remain a feature of the literary language. But those qualities that best distinguish Russian fiction—critical realism and spirituality—first appeared in the work of Mikhail Lermontov (1814–41), whose *Hero of Our Time* is the pioneering Russian psychological novel. Nikolay Gogol (1809–52), satirizing provincial mores in *Dead Souls,* ushered in, perhaps unintentionally, a whole fictional movement—the "literature of accusation." Vissarion Belinsky (1811–48), the father of Russian literary criticism, formulated the theory of literature in the service of society, and Ivan Turgenev (1818–83), produced a classic "accusatory" novel in *A Sportsman's Sketches.* But he, and the other major novelists who emerged in the mid-19th century, can hardly be considered in terms of literary movements. Fyodor Dostoyevsky (1821–81), with *The Possessed, The Idiot, Crime and Punishment,* and *The Brothers Karamazov,* affirmed with idiosyncratic power the great spiritual realities, and Leo Tolstoy (1828–1910) produced two of the greatest novels of all time—*War and Peace* and *Anna Karenina,* revelatory of the Russian soul but also of the very nature of universal man and human society.

The influence of the Soviet government

The later days of the 19th century saw a shift in fictional radicalism. The theories of Karl Marx influenced the accusatory writers in the direction of the plight of the proletariat, not, as had been the old way, that of the peasantry. Maksim Gorky (1868–1936), Leonid Andreyev (1871–1919), Aleksandr Kuprin (1870–1938), and the Nobel Prize winner (1933) Ivan Bunin (1870–1953) wrote of the Russian urban experience and helped to create the literary climate of the Soviet regime. Generally, since

the beginning of the first five-year plan in 1928, there has been a division between what the regime regards as valuable in the practice of the novel and what the rest of the world thinks. Mikhail Sholokhov (1905–84) depicted, with no evident propagandist slanting, the Revolution and civil war in *Quiet Flows the Don,* and Fyodor Gladkov (1883–1958) was one of the few novelists readable outside the ranks of the Soviet devout in the new category of economic or industrial fiction. The Russian novel of the age following World War II, however, has undergone a dramatic schism, in which writers like Aleksandr Solzhenitsyn and Boris Pasternak (1890–1960) could receive the Nobel Prize but be officially condemned in the Soviet Union, while the fictional darlings of the regime are recognized, in the non-Communist world, as possessing little or no aesthetic merit.

German. Goethe (1749–1832), who practiced so many arts with such notable brilliance, may be regarded as the first major novelist of Germany. His life covers the whole period of the Enlightenment, with its insistence on a national literary spirit, the *Sturm und Drang* movement, and that phase of *Weltschmerz* which *The Sorrows of Young Werther* fostered. It was *Werther* more than any other work that carried German literature into the world arena; it remained influential in Europe when German Romanticism had already burned itself out. The German reaction against Romanticism was expressed in the regionalism of Theodor Storm (1817–88) and Fritz Reuter (1810–74), who sought to render with objective fidelity the life of their native provinces (northwestern and northeastern Germany, respectively). Swiss writers like Gottfried Keller (1819–90) belonged to the movement of mainstream German regional realism, featuring the picturesque solidities of Switzerland. But the post-Goethean major achievements in the novel had to wait for the Impressionist movement, which produced the works of Thomas Mann (1875–1955) and Hermann Hesse (1877–1962)—fiction concerned less with a roughly hacked slice of life than with form and aesthetic delicacy. World War I brought Expressionism—the nightmares of the German-Czech Franz Kafka (1883–1924) and the psychoanalytical novels of Jakob Wassermann (1873–1934), with their pleas for humanity and justice.

Suppression of the novel by the Nazis

When the true historical nightmare of the Nazi regime followed—predicted, in a sense, by Kafka and Mann—liberal German fiction was suppressed, and liberal German novelists like Mann and Erich Maria Remarque (1898–1970), author of *All Quiet on the Western Front,* went into exile. The brutal, philistine, and nationalistic novels of the true Nazi novelists exist only as curious and frightening relics of an era of infamy. The task of postwar novelists like Günter Grass—author of *The Tin Drum* and *Dog Years*—and Uwe Johnson has been to diagnose the long sickness and force German fiction into new directions—often with the help of surrealism, irony, and verbal experiment.

French. Although the mid-16th-century satirical piece *Gargantua and Pantagruel* of François Rabelais has had, and is still having, a profound influence on the world novel, it would be wrong to place it in the line of true French fiction, which has always manifested a preoccupation with form, order, and economy—qualities totally un-Rabelaisian. The first notable French novel—*The Princess of Clèves* by Madame de La Fayette (1634–93)—shows none of the vices of the English novels of a century later: it is firmly constructed and takes character seriously, as does *Manon Lescaut* by l'Abbé Prévost (1697–1763). With such works the psychological novel was established in Europe. A keen hold on reality and a concern with the problems of man as a social being animates even the fiction of the Romantic school—works like *Indiana* and *Lélia* by George Sand (1804–76) and the heavyweight romances of Victor Hugo (1802–85). But the true glories of French fiction come with the reaction to romanticism and sentimentality, as exemplified in the novels of Jean-Jacques Rousseau (1712–78). The great fathers of realism are Stendhal (1783–1842), Balzac (1799–1850), and Flaubert (1821–80), and their influence is still active. Émile Zola (1840–1902) moved away from the artistic detachment that Flaubert preached and practiced, but, in his *Chronicles of the Rougon Macquart Family,* he tried to emulate the encyclopaedic approach to the novel of Balzac, whose *Human Comedy* is meant to be a history of society in a hundred episodes. The leader of the Naturalist school, Zola saw human character as a product of heredity and environment.

The realist masters

It was left to the 20th-century French novel to cast doubt on a mechanistic or deterministic view of man, to affirm the irrational element in his makeup, and to emphasize the primacy of the will. The temporal treadmill of Balzac and Zola has no place in the masterpiece of Marcel Proust (1871–1922); *Remembrance of Things Past,* if it has a philosophy, says more about the creative élan vital of Bergson, the human essences that underlie the shifting phenomena of time and space, than the social jungle the realists had taken for reality. André Gide (1869–1951) seems to make a plea for human aloofness from environment so that the essentially human capacity for change and growth may operate. André Malraux (1901–76), the forerunner of the Existentialist novelists, demonstrated, in *Man's Fate,* the necessity for human involvement in action as the only answer to the absurdity of his position in a huge and indifferent or malevolent universe. Jean-Paul Sartre (1905–80) and Albert Camus (1913–60) similarly emphasized man's freedom to choose, to say no to evil, to define himself through action.

Other French novelists have been more concerned with recording the minutiae of human life as a predominantly sensuous and emotional experience, like Colette (1873–1954), or with taking the religious sensibility as a fictional theme, like François Mauriac (1855–1970). The practitioners of the *anti-roman*—Butor, Robbe-Grillet, Nathalie Sarraute—pursue their attempts at liquidating human character in the traditional novelistic sense. Samuel Beckett, an Irishman who has turned himself into a major French stylist, goes his own way, presenting—in works like *Molloy, Malone Dies, Watt,* and *The Unnamable*—mankind reduced to degradation and absurdity but somehow admirable because it survives.

French-Canadian novels

French-Canadian fiction inevitably suffers from comparison with the glories of the mother country. Before 1900 there is little of value to record—except perhaps the historical romance of Philippe de Gaspé (1786–1871), *Les Anciens Canadiens*—but Louis Hémon (1880–1913) produced a genuine classic in *Maria Chapdelaine,* a story of Canadian pioneer life, original, moving, and sensitive. The somewhat provincial character of French-Canadian life, dominated by the Church and by outmoded notions of morality, has not been conducive either to fictional candour or to formal experiment, and metropolitan France is unimpressed, for the most part, by the literature of the separated brethren. But there is time for an avant-garde to develop and great masters to appear.

The Golden Age in Spain

Spanish. The great age known as "El Siglo de Oro," or the Golden Age, produced what is conceivably (it must contest this claim with *War and Peace*) the most magnificent of all novels—the *Don Quixote* of Miguel de Cervantes Saavedra (1547–1616). A satire on chivalry that ends as a humane affirmation of the chivalric principle, it encloses—in tender or comic distortion—other fictional forms that flourished in the Spanish Golden Age. The *novela picaresca* fathered a whole European movement, and its best monuments are perhaps the anonymous *Lazarillo de Tormes* (1554), *Guzmán de Alfarache* by Mateo Alemán (1547–?1614), and *El diablo cojuelo* (*The Limping Devil*) by Luis Vélez de Guevara (1579–1644). The pastoral novel, another popular but highly stylized form, was less true fiction than a sort of prose poem, in which lovers in shepherd's disguise bewailed an unattainable or treacherous mistress. But here, since the fictional lovers were often real-life personages in the cloak of a rustic name, the germs of the *roman à clef* are seen stirring. The *novela Morisca* was an inimitable Spanish form, a kind of fictional documentary about the wars between Christians and Muslims, as in *Guerras civiles de Granada* (*Civil Wars of Granada*) by Ginés Pérez de Hita (*c.* 1544–*c.* 1619).

The decline of Spain as a European power is associated, in literature, with feeble nostalgia for the Golden Age or feebler imitations of French classicism. The Spanish novel

began to reemerge only in the early 19th century, when a journalism celebrating regional customs encouraged the development of the realistic regional novel, with Fernán Caballero (1796–1877), Armando Palacio Valdés (1853–1938), and the important Vicente Blasco Ibáñez (1867–1928), whose *Sangre y arena* (*Blood and Sand*), *Mare nostrum,* and *Los cuatro jinetes del Apocalipsis* (*The Four Horsemen of the Apocalypse*) achieved universal fame and were adapted for the screen in the 1920s.

The fiction of the Generation of '98—which took its name from the year of the Spanish-American War, a cataclysmic event for Spain that bereaved it of the last parts of a once great empire—wasted no time on national nostalgia or self-pity, but concentrated on winning a new empire of style. Ramón María del Valle-Inclan (1866–1936) and Ramón Pérez de Ayala (1880–1962) brought a highly original lyricism to the novel, while Pío Baroja (1872–1956) concentrated on representing a world of discrete events, unbound by a unifying philosophy. The fiction that came out of the Civil War of 1936–39 returned to a kind of didactic realism, as with the novels of José Mariá Gironella, who depicted a ravaged and suffering Spain. Camilo José Cela, perhaps the most important modern Spanish novelist, combines realism with a highly original style. His *Familia de Pascual Duarte*—harrowing, compassionate, brilliantly economical—is a novel of towering merit.

Italian. Though Italy originated the *novella,* it was slow in coming to the full-length novel. There is little to record before *I promessi sposi* (*The Betrothed*) by Alessandro Manzoni (1785–1873), a romantic and patriotic novel that describes life in Italy under Spanish domination in the 17th century. That combination of regionalism and realism already noted in the fiction of Germany and Spain did not appear in Italy until after the unification in 1870, when Giovanni Verga (1840–1922) celebrated his native Sicily in *I malavoglia* (*The House by the Medlar Tree*) and Antonio Fogazzaro (1842–1911) showed life in northern Italy during the struggle for unification in *Piccolo mondo antico* (*The Little World of the Past*). Gabriele D'Annunzio (1863–1938) and Luigi Pirandello (1867–1936) were too original for easy classification, and both worked in all the literary media. Pirandello, especially, helped to bring Italian literature into the modern world through such philosophical novels as *Il fu Mattia Pascal* (*The Late Matthew Pascal*), which raises very profound questions about the nature of human identity and yet contrives to be witty, sunny, and essentially Italianate. The importance of Italo Svevo (1861–1928) was obscured for some time because of the difficulty of Italian literati in accepting his relatively unadorned style, but works like *Zeno* and *Senilitá* (a title that James Joyce, Svevo's friend and English teacher, translated *As a Man Grows Older*) are generally recognized as being major contributions to the international novel.

The significant fiction of the Mussolini regime was produced by anti-Fascist exiles like Giuseppe Borgese (1882–1952) and Ignazio Silone (1900–78), whose *Pane e vino* (*Bread and Wine*) is accepted as a 20th-century classic. Alberto Moravia, with *La romana* (*The Woman of Rome*) and *La noia* (*The Empty Canvas*), is perhaps the most popular Italian novelist outside Italy, but he is probably less important than Giuseppe Berto, Cesare Pavese (1908–50), and Elio Vittorini (1908–66). Giuseppe di Lampedusa (1896–1957) created a solitary masterpiece in *Il gattopardo* (*The Leopard*). The experimental writing of Italo Calvino and Carlo Emilio Gadda has become better known outside Italy, but, generally speaking, the Italian novel remains linguistically conservative and needs the impact of some new, powerfully iconoclastic literary figure like James Joyce.

Scandinavian languages. Norway is better known for Ibsen's contribution to the drama and Grieg's to music than for fiction of the first quality. Nevertheless, the novels of Knut Hamsun (1859–1952) earned him the Nobel Prize in 1920, and Sigrid Undset (1882–1949) received the same honour in 1928, though the work of both has failed to engage the lasting attention of world readers of fiction. In Denmark, Johannes Vilhelm Jensen (1873–1950), another Nobel Prize winner, Isak Dinesen (the pen name of Baroness Karen Blixen, 1885–1962), and Martin Nexö (1869–1954) have contributed to their country's fictional literature but made little mark beyond. Except for the Nobel Prize winners Eyvind Johnson (1900–76) and Harry Martinson (1904–78), the achievement of Sweden's novelists is inconsiderable. The novels of Halldór Laxness have restored to Iceland some of the literary fame it once earned for its sagas.

Slavic and East European languages. Czechoslovakia's greatest novelist, Franz Kafka, wrote in German, but its writers in the vernacular include world-famous names, such as Jaroslav Hašek (1883–1923), whose *The Good Soldier Schweik* is acknowledged to be a comic masterpiece, and Karel Čapek (1890–1938), best known for the plays *R.U.R.* (which gave "robot" to the world's vocabulary) and *The Life of the Insects,* but also notable for the novels *The Absolute at Large, Krakatit,* and *The War with the Newts.* The Czech fictional genius tends to the satirical and the fantastic. The fiction of Yugoslavia has made little impact on the progress of the novel, but Radomir Konstantinović should be noted: his *Exitus* is a remarkable study of the death of Christ. Poland can claim two Nobel prizewinning novelists in Henryk Sienkiewicz (1846–1916), the author of *Quo Vadis?,* and Władysław S. Reymont (1868–1925), whose novel *The Peasants* is an aromatic piece of bucolic realism. Witold Gombrowicz (1904–69) wrote a novel, *Ferdydurke,* that has been subjected to two separate modes of suppression—first Fascist, later Communist. A remarkable surrealist essay on "anal tyranny" and depersonalization, it still awaits the acclaim that is due. Romania, which produced outstanding novelists in Eugen Lovinescu and Titu Maiorescu, has suffered, like other Balkan countries, from the totalitarian suppression of the free creative spirit; but Dumitru Radu Popescu, author of *The Blue Lion,* was bold enough in the 1960s to question Communist orthodoxy through the medium of fiction. The work of the Hungarian Gusztáv Rab (1901–66) awaits recognition. His brilliant *Sabaria* develops very courageously the theme of the conflict between Communism and Christianity and reaches conclusions favourable to neither. It is a disturbing and beautifully composed book.

Modern Greece, perhaps more famous for its poets George Seferis and Constantine Cavafy, has produced at least one major novelist in Níkos Kazantzakís (1885–1957), whose *Zorba the Greek* became famous through its film adaptation. *The Last Temptation of Christ,* which presents the life of Jesus as a struggle to overcome "the dark immemorial forces of the Evil One, human and prehuman," glows with the writer's personality and bristles with the dialectic that was one of the first Greek gifts to Western civilization.

The Jewish novel. The literature of the Diaspora—the dispersion of the Jews after their exile from ancient Babylonia—records many large achievements in the languages of exile and that dialect of Low German—Yiddish—which the Ashkenazi Jews have taken around the world. Perhaps the most interesting of modern Jewish novelists in Yiddish is Isaac Bashevis Singer, a naturalized American who refuses to be absorbed linguistically into America, unlike Bellow and Malamud, who have brought to the Anglo-American language typical tones and rhythms of the ghetto.

Israel, which is producing its own rich crop of national writers, began with an existing corpus of literature in modern Hebrew, a language promoted and nurtured by such scholars as Eliezer ben Yehuda (1858–1923) and the members of the neologizing Hebrew Academy of Israel. Among the early Hebrew novelists is Abraham Mapu (1808–67); among the later ones is the brilliant Moshe Shamir. The first Nobel Prize for Literature ever awarded to a Hebrew novelist went, justly, to Shmuel Yosef Agnon (1888–1970). The contemporary Hebrew novel is notable for a fusion of international sophistication and earthy homegrown realism.

ASIA, AFRICA, LATIN AMERICA

China. The tradition of storytelling is an ancient one in China, and the full-length novel can be found as far back as the late 16th century. The fiction of the 18th century

shows a variety of themes and techniques not dissimilar to those of Europe, with social satire, chivalric romance, and adventure stories. Ts'ao Chan (1715?–63) wrote a novel called *Hung lou meng* (translated as *Dream of the Red Chamber* in 1892), which has features not unlike those of Galsworthy's *The Forsyte Saga* and Mann's *Buddenbrooks*—a story of a great aristocratic family in decline, garnished with love interest and shot through with pathos. The early days of the 20th century saw the foundation of the Chinese republic and the development of popular fiction written in the vernacular style of Chinese called *pai-hua.*

From 1917 until the Sino-Japanese War (1937–45), a period of social and intellectual ferment, there was a great influx of Western novels, and, under their influence, movements like those devoted to realism and naturalism in Europe produced a great number of didactic novels. Novels like Lao She's *Rickshaw Boy* and Pa Chin's *Chinese Earth* appeared in the postwar period, but the Communist regime has quelled all but propagandist fiction. Such novels as Liu Ching's *Wall of Bronze,* Chao Shu-li's *Changes in Li Village,* and Ting Ling's *Sun Shines over the Sangkan River* are typical glorifications of the Maoist philosophy and the socialist achievements of the people.

Japan. A literature that admires economy, like the Japanese, is bound to favour, in fictional art, the short story above the full-length novel. Nevertheless, some of the ancient pillow-books, with their diary jottings and anecdotes, have the ring of autobiographical novels; while Murasaki Shikibu's *Tale of Genji,* produced nearly a thousand years ago, is a great and sophisticated work of fictional history. The 20th-century Japanese novel has been developed chiefly under Western influence, like the work of Akutagawa Ryūnosuke (1892–1927), whose short stories "Rashōmon" and "Yabu no naka" became the highly applauded film *Rashomon.* Tanizaki Jun-ichirō (1886–1965) is well known in America and Europe for *The Key, The Makioka Sisters,* and *Diary of a Mad Old Man.* His novels, all set in a modern, Westernized Japan, are assured of universal popularity because of their frank and lavish sexual content. Mishima Yukio (1925–70), who committed ceremonial suicide at the height of his reputation, was perhaps the most successful, and certainly the most prolific, of all modern Japanese novelists. Ten of his fictions had been filmed; he had won all the major Japanese literary awards and appeared to be destined for the Nobel Prize. His works are characterized by ruthless violence and a perversity that, however much it seems to derive from the pornographic excesses of Western fiction, is certainly in the Japanese tradition. His work is hard to judge, but he was the most considerable literary figure of the East.

Western influences

India and East Asia. The Indian novel is a branch of British Commonwealth literature. That is to say, it is practiced by writers who have received a British literary education and, for the most part, publish their books in London. There is no evidence of any great development of fiction in any of the native Indian tongues: a taste for reading novels, as opposed to seeing films or reading short stories in the Hindi or Punjabi press, is acquired in India along with an education in English, which is still the unifying tongue of the subcontinent. In Malaysia, where the same tradition holds, short stories are being written in Malay, Chinese, and Tamil, but the full-length novel is almost exclusively written in English. It is as much a matter of literary markets as of literary education. No novel in any of the tongues of the peninsula is likely to be a remunerative publishing proposition, as Han Suyin—a best-selling Chinese woman doctor from Johore, who found large fame with *A Many-Splendoured Thing*—would be the first to admit, for all her dislike of the British colonial tradition. Indonesia's abandonment of all vestiges of its Dutch colonial past is associated with the encouragement of a literature in Bahasa Indonesia, a variety of Malay, and there are a number of Indonesian novelists still looking for a large educated audience—inevitably in translation, in the West. Among these Mochtar Lubis is notable; his *Twilight in Djakarta* is a bitter indictment of the Sukarno regime, which promptly sent him to prison. This was not the kind of fiction that the new Indonesia had in mind.

Africa. What applies to India and the East Indies applies also to Africa: there is still an insufficiently large audience for fiction written in any of the major African languages, and African novelists brought up on English are only too happy to continue working in it. Other European languages—chiefly Afrikaans (a South African variety of Dutch) and French—are employed as a fictional medium. Arthur Fula, who wrote *Janie Giet die Beeld* (*Janie Casts the Image*), is an Afrikaans novelist whose reputation has not, as yet, stretched to Europe or America, but the novelists of the former French possessions are gaining a name among serious students of the African novel. Mongo Beti, from Cameroon, is known for his *Pauvre Christ de Bomba* and *Le Roi miraculé;* the Ivory Coast has Aké Loba, and Hamidou Kane represents Senegal. This new African French deserves to be regarded as a distinct literary language, unrelated to that of Paris or Quebec, but the critical and linguistic tools for appraising the fiction of French-speaking Africans are not yet available.

Latin America. One of the greatest contemporary writers of fiction, Jorge Luis Borges, is an Argentinian, but his *ficciones* are very short short stories and must, with regrets, be excluded from any survey of the novel. It is significant, however, that the circumstances for the creation of a great fictional literature are in existence in Latin America, signifying (unlike the case in many former British dependencies) a shedding of the old colonial provincialism, which relied on the opinion of Madrid or, in Brazil, of Lisbon. The first Latin American novel was probably *El periquillo sarniento* (1816; *The Itching Parrot*), by the Mexican José Joaquín Fernández de Lizardi (1776–1827), a picaresque work satirizing colonial conditions. In Argentina, José Mármol (1817–1871) published the first major novel of the continent—*Amalia* (1851–55), a powerful study of the fear and degradation that were rife in Buenos Aires during the dictatorship of the corrupt and tyrannical Juan Manuel de Rosas. The Romantic period in Europe had its counterpart in the sentimental wave that overtook such novelists as the Colombian Jorge Isaacs (1837–95), while the new humanitarianism found a voice in *Birds Without a Nest,* a protest-novel on the conditions forced on the Indians of Peru, written by Clorinda Matto de Turner (1854–1901). Juan León Mera followed the same trend in *Cumandá,* a novel about the oppression of the Ecuadorian Indians. As Latin America moved toward the modern age, the inevitable novels of urban social protest made their appearance. Alberto Blest Gana (1830–1920) of Chile, Carlos Reyles (1868–1938) of Uruguay, and Gustavo Martínez Zuviría of Argentina are names of some historical significance; and Reyles's naturalistic novel *La Raza de Caín* (*Cain's Race*) is original in that it finds a parallel between the breeding of stock and the building of a human society.

A 20th-century reaction against the bourgeois novel led to the movement known as nativism, with its concentration on the land itself, the lot of the indigenous peoples, the need for an anti-racist revolution with the aim of true egalitarianism. *Los de Abajo* (1915; *The Underdogs*), by the Mexican Mariano Azuela (1873–1952), and *El Señor Presidente* (1946), by Miguel Angel Asturias of Guatemala, typify the new revolutionary novel. Much of the didactic energy that went to the making of such work resulted in a lack of balance between subject matter and style: the literature of revolt tends to be shrill and crude. The inevitable reaction to a more sophisticated kind of literature, in which the individual became more important than society and the unconscious more interesting than the operation of reason, led to the highly refined experimentation that is the mark of Borges, the Brazilian Érico Veríssimo, and Eduardo Barrios of Chile.

Rise of nativism

The fiction of Brazil is in many ways more interesting than the fiction of the mother country, which has produced only one major novelist in the realist José Maria de Eça de Queirós (1845–1900). Gregório de Matos Guerra, as early as the 17th century, wrote bitterly of colonial administration and painted a realistic picture of Brazilian life. Irony and keen observation have been characteristic of the Brazilian novel ever since—as in the *Memórias Póstumus de Brás Cubas* (1881) and *Dom Casmurro* of

Joaquim Maria Machado de Assis (1839–1908); the *Canaã* (1902; *Canaan*) of José Pereira da Graça Aranha (1868–1931), a remarkable study of the disillusion of German settlers in a new land; and the *Os Sertões* (*Rebellion in the Backlands*) of Euclides da Cunha (1866–1909), an account of a revolt against the newly formed Brazilian republic. The social consciousness of Brazilian novelists is remarkable, especially when it is associated with a strong concern for stylistic economy and grace. Monteiro Lobato (1882–1948), with his rustic hero Jeca Tatú; José Lins do Rêgo (1901–1957), chronicler of the decay of the old plantation life; Jorge Amado, a socialist novelist much concerned with slum life in Bahia; Érico Veríssimo, an experimental writer grounded in the traditional virtues of credible plot and strong characterization—these attest a vigour hardly to be found in the fiction published in Lisbon.

SOCIAL AND ECONOMIC ASPECTS

Though publishers of fiction recognize certain obligations to art, even when these are unprofitable (as they usually are), they are impelled for the most part to regard the novel as a commercial property and to be better pleased with large sales of indifferent work than with the mere unremunerative acclaim of the intelligentsia for books of rare merit. For this reason, any novelist who seeks to practice his craft professionally must consult the claims of the market and effect a compromise between what he wishes to write and what the public will buy. Many worthy experimental novels, or novels more earnest than entertaining, gather dust in manuscript or are circulated privately in photocopies. Indeed, the difficulty that some unestablished novelists find in gaining a readership (which means the attention of a commercial publisher) has led them to take the copying machine as seriously as the printing press and to make the composition, mimeographing, binding, and distribution of a novel into a single cottage industry. For the majority of novelists the financial rewards of their art are nugatory, and only a strong devotion to the form for its own sake can drive them to the building of an *oeuvre*. The subsidies provided by university sinecures sustain a fair number of major American novelists; others, in most countries, support their art by practicing various kinds of subliterature—journalism, film scripts, textbooks, even pseudonymous pornography. Few novelists write novels and novels only.

Awards, patronage, and ancillary activities

There are certain marginal windfalls, and the hope of gaining one of these tempers the average novelist's chronic desperation. America has its National Book Award as well as its book club choices; France has a great variety of prizes; there are also international bestowals; above all, there glows the rarest and richest of all accolades—the Nobel Prize for Literature. Quite often the Nobel Prize winner needs the money as much as the fame, and his election to the honour is not necessarily a reflection of a universal esteem which, even for geniuses like Samuel Beckett, means large sales and rich royalties. When Sinclair Lewis received the award in 1930, wealth and fame were added to wealth and fame already sufficiently large; when William Faulkner was chosen in 1949, most of his novels had been long out of print in America (if not in Sweden).

Prizes come so rarely, and often seem to be bestowed so capriciously, that few novelists build major hopes on them. They build even fewer hopes on patronage: Harriet Shaw Weaver, James Joyce's patroness, was probably the last of a breed that, from Maecenas on, once intermittently flourished; state patronage—as represented, for instance, by the annual awards of the Arts Council of Great Britain—can provide little more than a temporary palliative for the novelist's indigence. Novelists have more reasonable hopes from the world of the film or the stage, where adaptations can be profitable and even salvatory. The long struggles of the British novelist T.H. White came to an end when his Arthurian sequence *The Once and Future King* (1958) was translated into a stage musical called *Camelot,* though by treating the lump sum paid to him as a single year's income instead of a reward for decades of struggle, nearly all the windfall would have gone for taxes if White had not taken his money into low-tax exile. Such writers as Graham Greene, nearly all of whose novels have been filmed, must be tempted to regard mere book sales as an inconsiderable aspect of the rewards of creative writing. There are few novelists who have not received welcome and unexpected advances on film options, and sometimes the hope of film adaptation has influenced the novelist's style. In certain countries, such as Great Britain but not the United States, television adaptation of published fiction is common, though it pays the author less well than commercial cinema.

When a novelist becomes involved in film-script writing—either in the adaptation of his own work or that of others—the tendency is for him to become subtly corrupted by what seems to him an easier as well as more lucrative technique than that of the novel. Most novelists write dialogue with ease, and their contribution to a film is mostly dialogue: the real problem in novel writing lies in the management of the *récit*. A number of potentially fine novelists, like Terry Southern and Frederic Raphael, have virtually abandoned the literary craft because of their continued success with script writing. In 70-odd years the British novelist Richard Hughes produced only three novels, the excellence of which has been universally recognized; fiction lovers have been deprived of more because of the claims of the film world on Hughes's talent. This kind of situation finds no counterpart in any other period of literary history, except perhaps in the Elizabethan, when the commercial lure of the drama made some good poets write poor plays.

The importance of the paperback

The majority of professional novelists must look primarily to book sales for their income, and they must look decreasingly to hard-cover sales. The novel in its traditional format, firmly stitched and sturdily clothbound, is bought either by libraries or by readers who take fiction seriously enough to wish to acquire a novel as soon as it appears: if they wait 12 months or so they can buy the novel in paper covers for at most a fifth of its original price. The paperback edition of a novel has become, for the vast majority of fiction readers, the form in which they first meet it, and the novelist who does not achieve paperback publication is missing a vast potential audience. He may not repine at this, since the quantitative approach to literary communication may safely be disregarded: the legend on a paperback cover—FIVE MILLION COPIES SOLD—says nothing about the worth of the book within. Nevertheless, the advance he will receive from his hard-cover publisher is geared to eventual paperback expectations, and the "package deal" has become the rule in negotiations between publisher and author's agent. The agent, incidentally, has become important to both publisher and author to an extent that writers like Daniel Defoe and Samuel Richardson would, if resurrected, find hard to understand.

The novelist may reasonably expect to augment his income through the sale of foreign rights in his work, though the rewards accruing from translation are always uncertain. The translator himself is usually a professional man and demands a reasonable reward for his labours, more indeed than the original author may expect: the reputations of some translators are higher than those of some authors, and even their names may be better known. Moreover, the author who earns most from publication in his own language will usually earn most in translation, since it is the high initial home sales that attract foreign publishers to a book. The more "literary" a novel is, the more it exploits the resources of the author's own language, the less likely is it to achieve either popularity at home or publication abroad. Best-selling novels like Mario Puzo's *Godfather* (1969) or Arthur Hailey's *Airport* (1968) are easy to read and easy to translate, so they win all round. It occasionally happens that an author is more popular abroad than he is at home: the best-selling novels of the Scottish physician-novelist A.J. Cronin are perhaps no longer highly regarded in England and America, as they were in the 1930s and 1940s, but they still sell by the million in the U.S.S.R. But a novelist is wisest to expect most from his own country and to regard foreign popularity as an inexplicable bonus.

Problems of obscenity and libel

As though his financial problems were not enough, the novelist frequently has to encounter those dragons unleashed by public morality or by the law. The struggles

of Flaubert, Zola, and Joyce, denounced for attempting to advance the frontiers of literary candour, are well known and still vicariously painful, but lesser novelists, working in a more permissive age, can record cognate agonies. Generally speaking, any novelist writing after the publication in the 1960s of Hubert Selby's *Last Exit to Brooklyn* or Gore Vidal's *Myra Breckenridge* can expect little objection, on the part of either publisher or police, to language or subject matter totally unacceptable, under the obscenity laws then operating, in 1922, when *Ulysses* was first published. This is certainly true of America, if not of Ireland or Malta. But many serious novelists fear an eventual reaction against literary permissiveness as a result of the exploitation by cynical obscenity mongers or hard-core pornographers of the existing liberal situation.

In some countries, particularly Great Britain, the law of libel presents insuperable problems to novelists who, innocent of libellous intent, are nevertheless sometimes charged with defamation by persons who claim to be the models for characters in works of fiction. Disclaimers to the effect that "resemblances to real-life people are wholly coincidental" have no validity in law, which upholds the right of a plaintiff to base his charge on the corroboration of "reasonable people." Many such libel cases are settled before they come to trial, and publishers will, for the sake of peace and in the interests of economy, make a cash payment to the plaintiff without considering the author's side. They will also, and herein lies the serious blow to the author, withdraw copies of the allegedly offensive book and pulp the balance of a whole edition. Novelists are seriously hampered in their endeavours to show, in a traditional spirit of artistic honesty, corruption in public life; they have to tread carefully even in depicting purely imaginary characters and situations, since the chance collocation of a name, a profession, and a locality may produce a libellous situation.

EVALUATION AND STUDY

It has been only in comparatively recent times that the novel has been taken sufficiently seriously by critics for the generation of aesthetic appraisal and the formulation of fictional theories. The first critics of the novel developed their craft not in full-length books but in reviews published in periodicals: much of this writing—in the late 18th and early 19th centuries—was of an occasional nature, and not a little of it casual and desultory; nor, at first, did critics of fiction find it easy to separate a kind of moral judgment of the subject matter from an aesthetic judgement of the style. Such fragmentary observations on the novel as those made by Dr. Johnson in conversation or by Jane Austen in her letters, or, in France, by Gustave Flaubert during the actual process of artistic gestation, have the charm and freshness of insight rather than the weight of true aesthetic judgment. It is perhaps not until the beginning of the 20th century, when Henry James wrote his authoritative prefaces to his own collected novels, that a true criteriology of fiction can be said to have come into existence. The academic study of the novel presupposes some general body of theory, like that provided by Percy Lubbock's *Craft of Fiction* (1921) or E.M. Forster's *Aspects of the Novel* (1927) or the subsequent writings of the critics Edmund Wilson and F.R. Leavis. Since World War II it may be said that university courses in the evaluation of fiction have attained the dignity traditionally monopolized by poetry and the drama.

The distinction between reviewing and criticism

A clear line should be drawn between the craft of fiction criticism and the journeyman work of fiction reviewing. Reviews are mainly intended to provide immediate information about new novels: they are done quickly and are subject to the limitations of space; they not infrequently make hasty judgments that are later regretted. The qualifications sought in a reviewer are not formidable: smartness, panache, waspishness—qualities that often draw the attention of the reader to the personality of the reviewer rather than the work under review—will always be more attractive to circulation-hunting editors than a less spectacular concern with balanced judgment. A thoughtful editor will sometimes put the reviewing of novels into the hands of a practicing novelist, who—knowing the labour that goes into even the meanest book—will be inclined to sympathy more than to flamboyant condemnation. The best critics of fiction are probably novelists *manqués,* men who have attempted the art and, if not exactly failed, not succeeded as well as they could have wished. Novelists who achieve very large success are possibly not to be trusted as critics: obsessed by their own individual aims and attainments, shorn of self-doubt by the literary world's acclaim or their royalty statements, they bring to other men's novels a kind of magisterial blindness.

Novelists can be elated by good reviews and depressed by bad ones, but it is rare that a novelist's practice is much affected by what he reads about himself in the literary columns. Genuine criticism is a very different matter, and a writer's approach to his art can be radically modified by the arguments and summations of a critic he respects or fears. As the hen is unable to judge of the quality of the egg it lays, so the novelist is rarely able to explain or evaluate his work. He relies on the professional critic for the elucidation of the patterns in his novels, for an account of their subliminal symbolism, for a reasoned exposition of their stylistic faults. As for the novel reader, he will often learn enthusiasm for particular novelists through the writings of critics rather than from direct confrontation with the novels themselves. The essays in Edmund Wilson's *Axel's Castle* (1931) aroused an interest in the Symbolist movement which the movement was not easily able to arouse by itself; the essay on *Finnegans Wake,* collected in Wilson's *Wound and the Bow* (1941), eased the way into a very difficult book in a manner that no grim work of solid exegesis could have achieved. The essence of the finest criticism derives from wisdom and humanity more than from mere expert knowledge. Great literature and great criticism possess in common a sort of penumbra of wide but unsystematic learning, a devotion to civilized values, an awareness of tradition, and a willingness to rely occasionally on the irrational and intuitive.

All this probably means that the criticism of fiction can never, despite the efforts of aestheticians schooled in modern linguistics, become an exact science. A novel must be evaluated in terms of a firmly held literary philosophy, but such a philosophy is, in the final analysis, based on the irrational and subjective. If the major premises on which F.R. Leavis bases his judgments of George Eliot, Mark Twain, and D.H. Lawrence are accepted, then an acceptance of the judgments themselves is inescapable. But many students of fiction who are skeptical of Leavis will read him in order that judgments of their own may emerge out of a purely negative rejection of his. In reading criticism a kind of dialectic is involved, but no synthesis is ever final. The process of revaluation goes on for ever. One of the sure tests of a novel's worth is its capacity for engendering critical dialectic: no novel is beyond criticism, but many are beneath it.

THE FUTURE OF THE NOVEL

It is apparent that neither law nor public morality nor the public's neglect nor the critic's scorn has ever seriously deflected the dedicated novelist from his self-imposed task of interpreting the real world or inventing alternative worlds. Statistics since World War II have shown a steady increase in the number of novels published annually, and beneath the iceberg tip of published fiction lies a submarine Everest of unpublished work. It has been said that every person has at least one novel in him, and the near-universal literacy of the West has produced dreams of authorship in social ranks traditionally deprived of literature. Some of these dreams come true, and taxi drivers, pugilists, criminals, and film stars have competed, often successfully, in a field that once belonged to professional writers alone. It is significant that the amateur who dreams of literary success almost invariably chooses the novel, not the poem, essay, or autobiography. Fiction requires no special training and can be readable, even absorbing, when it breaks the most elementary rules of style. It tolerates a literary incompetence unthinkable in the poem. If all professional novelists withdrew, the form would not languish: amateurs would fill the market with first and only novels, all of which would find readership.

Amateur novelists

But the future of any art lies with its professionals. Here a distinction has to be made between the Joyces, Henry Jameses, and Conrads on the one hand, and the more ephemeral Mickey Spillanes, Harold Robbinses, and Irving Wallaces on the other. Of the skill of the latter class of novelists there can be no doubt, but it is a skill employed for limited ends, chiefly the making of money, and through it the novel can never advance as art. The literary professionals, however, are dedicated to the discovery of new means of expressing, through the experiential immediacies that are the very stuff of fiction, the nature of man and society. In the symbiosis of publishing, the best-seller will probably continue to finance genuine fictional art. Despite the competition from other art media, and the agonies and the indigence, there are indications that the serious novel will flourish in the future.

It will flourish because it is the one literary form capable of absorbing all the others. The technique of the stage drama or the film can be employed in the novel (as in *Ulysses* and *Giles Goât-Boy*), as can the devices of poetry (as in Philip Toynbee's *Pantaloon* and the novels of Wilson Harris and Janet Frame). In France, as Michel Butor has pointed out, the new novel is increasingly performing some of the tasks of the old essay; in America, as Capote's *In Cold Blood* and Mailer's *Armies of the Night* have shown, the documentary report can gain strength from its presentation as fictional narrative. There are few limits on what the novel can do, there are many experimental paths still to be trod, and there is never any shortage of subject matter.

For all this, periods of decline and inanition may be expected, though not everywhere at once. The strength of the American novel in the period after World War II had something to do with the national atmosphere of breakdown and change: political and social urgencies promoted a quality of urgency in the works of such writers as Mailer, Bellow, Ellison, Heller, and Philip Roth. In the same period, Britain, having shed its empire and erected a welfare state, robbed its novelists of anything larger to write about than temporary indentations in the class system, suburban adultery, and manners. An achieved or static society does not easily produce great art. France, which has known much social and ideological turmoil, has generated a new aesthetic of the novel as well as a philosophy that, as Sartre and Camus have shown, is very suitable for fictional expression. A state on which intellectual quietism or a political philosophy of art is imposed by the ruling party can, as the Soviet Union and China show, succeed only in thwarting literary greatness, but the examples of Pasternak and Solzhenitsyn are reminders that repression can, with rare artistic spirits, act as an agonizing stimulus.

Every art in every country is subject to a cyclical process; during a period of decline it is necessary to keep the communication lines open, producing minor art so that it may some day, unexpectedly, turn into major art. Wherever the novel seems to be dying it is probably settling into sleep; elsewhere it will be alive and vigorous enough. It is important to believe that the novel has a future, though not everywhere at once. (An.B.)

Short story

The short story is a kind of prose fiction, usually more compact and intense than the novel and the short novel (novelette). Prior to the 19th century it was not generally regarded as a distinct literary form. But although in this sense it may seem to be a uniquely modern genre, the fact is that short prose fiction is nearly as old as language itself. Throughout history man has enjoyed various types of brief narratives: jests, anecdotes, studied digressions, short allegorical romances, moralizing fairy tales, short myths, and abbreviated historical legends. None of these constitutes a short story as the 19th and 20th centuries have defined the term, but they do make up a large part of the milieu from which the modern short story emerged.

Many of the elements of storytelling common to the short story and the novel are discussed at greater length in the preceding section on the novel. The short stories of particular literary cultures, along with other genres, are discussed in articles such as LITERATURE, THE HISTORY OF WESTERN; and in articles on the arts of various peoples—*e.g.*, SOUTH ASIAN ARTS.

ANALYSIS OF THE GENRE

As a genre, the short story has received relatively little critical attention, and the most valuable studies of the form that exist are often limited by region or era (*e.g.*, Ray B. West's *The Short Story in America,* 1900–50). One recent attempt to account for the genre has been offered by the Irish short story writer Frank O'Connor, who suggests that stories are a means for "submerged population groups" to address a dominating community. Most other theoretical discussions, however, are predicated in one way or another on Edgar Allan Poe's thesis that stories must have a compact, unified effect.

By far the majority of criticism on the short story focusses on techniques of writing. Many, and often the best of the technical works, advise the young reader—alerting him to the variety of devices and tactics employed by the skilled writer. On the other hand, many of these works are no more than treatises on "how to write stories" for the young writer, and not serious critical material.

The prevalence in the 19th century of two words, "sketch" and "tale," affords one way of looking at the genre. In the United States alone there were virtually hundreds of books claiming to be collections of sketches (Washington Irving's *Sketch Book,* William Dean Howells' *Suburban Sketches*) or collections of tales (Poe's *Tales of the Grotesque and Arabesque,* Herman Melville's *Piazza Tales*). These two terms establish the polarities of the milieu out of which the modern short story grew.

The sketch and the tale

The tale is much older than the sketch. Basically, the tale is a manifestation of a culture's unaging desire to name and conceptualize its place in the cosmos. It provides a culture's narrative framework for such things as its vision of itself and its homeland or for expressing its conception of its ancestors and its gods. Usually filled with cryptic and uniquely deployed motifs, personages, and symbols, tales are frequently fully understood only by members of the particular culture to which they belong. Simply, tales are intracultural. Seldom created to address an outside culture, a tale is a medium through which a culture speaks to itself and thus perpetuates its own values and stabilizes its own identity. The old speak to the young through tales.

The sketch, by contrast, is intercultural, depicting some phenomenon of one culture for the benefit or pleasure of a second culture. Factual and journalistic, in essence the sketch is generally more analytic or descriptive and less narrative or dramatic than the tale. Moreover, the sketch by nature is *suggestive,* incomplete; the tale is often *hyperbolic,* overstated.

The primary mode of the sketch is written; that of the tale, spoken. This difference alone accounts for their strikingly different effects. The sketch writer can have, or pretend to have, his eye on his subject. The tale, recounted at court or campfire—or at some place similarly removed in time from the event—is nearly always a recreation of the past. The tale-teller is an agent of *time,* bringing together a culture's past and its present. The sketch writer is more an agent of *space,* bringing an aspect of one culture to the attention of a second.

It is only a slight oversimplification to suggest that the tale was the only kind of short fiction until the 16th century, when a rising middle class interest in social realism on the one hand and in exotic lands on the other put a premium on sketches of subcultures and foreign regions. In the 19th century certain writers—those one might call the "fathers" of the modern story: Nikolay Gogol, Hawthorne, E.T.A. Hoffmann, Heinrich von Kleist, Prosper Mérimée, Poe—combined elements of the tale with elements of the sketch. Each writer worked in his own way, but the general effect was to mitigate some of the fantasy and stultifying conventionality of the tale and, at the same time, to liberate the sketch from its bondage to strict factuality. The modern short story, then, ranges between the highly imaginative tale and the photographic sketch and in some ways draws on both.

Modern fusions

The short stories of Ernest Hemingway, for example,

may often gain their force from an exploitation of traditional mythic symbols (water, fish, groin wounds), but they are more closely related to the sketch than to the tale. Indeed, Hemingway was able at times to submit his apparently factual stories as newspaper copy. In contrast, the stories of Hemingway's contemporary William Faulkner more closely resemble the tale. Faulkner seldom seems to understate, and his stories carry a heavy flavour of the past. Both his language and his subject matter are rich in traditional material. A Southerner might well suspect that only a reader steeped in sympathetic knowledge of the traditional South could fully understand Faulkner. Faulkner may seem, at times, to be a Southerner speaking to and for Southerners. But, as, by virtue of their imaginative and symbolic qualities, Hemingway's narratives are more than journalistic sketches, so, by virtue of their explorative and analytic qualities, Faulkner's narratives are more than Southern tales.

Whether or not one sees the modern short story as a fusion of sketch and tale, it is hardly disputable that today the short story is a distinct and autonomous, though still developing, genre.

HISTORY

Origins. The evolution of the short story first began before man could write. To aid himself in constructing and memorizing tales, the early storyteller often relied on stock phrases, fixed rhythms, and rhyme. Consequently, many of the oldest narratives in the world, such as the famous Babylonian tale the *Epic of Gilgamesh* (*c.* 2000 BC), are in verse. Indeed, most major stories from the ancient Middle East were in verse: "The War of the Gods," "The Story of Adapa" (both Babylonian), "The Heavenly Bow," and "The King Who Forgot" (both Canaanite). These tales were inscribed in cuneiform on clay during the 2nd millennium BC.

The earliest tales extant

The earliest tales extant from Egypt were composed on papyrus at a comparable date. The ancient Egyptians seem to have written their narratives largely in prose, apparently reserving verse for their religious hymns and working songs. One of the earliest surviving Egyptian tales, "The Shipwrecked Sailor" (*c.* 2000 BC), is clearly intended to be a consoling and inspiring story to reassure its aristocratic audience that apparent misfortune can in the end become good fortune. Also recorded during the 12th dynasty were the success story of the exile Sinuhe and the moralizing tale called "King Cheops [Khufu] and the Magicians." The provocative and profusely detailed story "The Tale of Two Brothers" (or "Anpu and Bata") was written down during the New Kingdom, probably around 1250 BC. Of all the early Egyptian tales, most of which are baldly didactic, this story is perhaps the richest in folk motifs and the most intricate in plot.

The earliest tales from India are not as old as those from Egypt and the Middle East. The *Brāhmaṇas* (*c.* 700 BC) function mostly as theological appendixes to the Four Vedas, but a few are composed as short, instructional parables. Perhaps more interesting as stories are the later tales in the Pāli language, *The Jātaka.* Although these tales have a religious frame that attempts to recast them as Buddhist ethical teachings, their actual concern is generally with secular behaviour and practical wisdom. Another, nearly contemporaneous collection of Indian tales, *The Pañca-tantra* (*c.* AD 500), has been one of the world's most popular books. This anthology of amusing and moralistic animal tales, akin to those of "Aesop" in Greece, was translated into Middle Persian in the 6th century; into Arabic in the 8th century; and into Hebrew, Greek, and Latin soon thereafter. Sir Thomas North's English translation appeared in 1570. Another noteworthy collection is *Kathā-saritsāqara* ("Ocean of Rivers of Stories), a series of tales assembled and recounted in narrative verse in the 11th century by the Sanskrit writer Samadeva. Most of these tales come from much older material, and they vary from the fantastic story of a transformed swan to a more probable tale of a loyal but misunderstood servant.

During the 2nd, 3rd, and 4th centuries BC, the Hebrews first wrote down some of their rather sophisticated narratives, which are now a part of the Old Testament and the Apocrypha. The book of Tobit displays an unprecedented sense of ironic humour; Judith creates an unrelenting and suspenseful tension as it builds to its bloody climax; the story of Susanna, the most compact and least fantastic in the Apocrypha, develops a three-sided conflict involving the innocent beauty of Susanna, the lechery of the elders, and the triumphant wisdom of Daniel. The Old Testament books of Ruth, Esther, and Jonah hardly need mentioning: they may well be the most famous stories in the world.

The didacticism of early tales

Nearly all of the ancient tales, whether from Israel, India, Egypt, or the Middle East, were fundamentally didactic. Some of these ancient stories preached by presenting an ideal for readers to imitate. Others tagged with a "moral" were more direct. Most stories, however, preached by illustrating the success and joy that was available to the "good" man and by conveying a sense of the terror and misery that was in store for the wayward.

The early Greeks contributed greatly to the scope and art of short fiction. As in India, the moralizing animal fable was a common form; many of these tales were collected as "Aesop's fables" in the 6th century BC. Brief mythological stories of the gods' adventures in love and war were also popular in the pre-Attic age. Apollodorus of Athens compiled a handbook of epitomes, or abstracts, of these tales around the 2nd century BC, but the tales themselves are no longer extant in their original form. They appear, though somewhat transformed, in the longer poetical works of Hesiod, Homer, and the tragedians. Short tales found their way into long prose forms as well, as in Hellanicus' *Persika* (5th century BC, extant only in fragments).

Herodotus, the "father of history," saw himself as a maker and reciter of *logoi* (things for telling, tales). His long *History* is interspersed with such fictionalized digressions as the stories of Polycrates and his emerald ring, of Candaules' attractive wife, and of Rhampsinitus' stolen treasure. Xenophon's philosophical history, the *Cyropaedia* (4th century BC), contains the famous story of the soldier Abradates and his lovely and loyal wife Panthea, perhaps the first Western love story. The *Cyropaedia* also contains other narrative interpolations: the story of Pheraules, who freely gave away his wealth; the tale of Gobryas' murdered son; and various anecdotes describing the life of the Persian soldier.

Beginnings of the romance

Moreover, the Greeks are usually credited with originating the romance, a long form of prose fiction with stylized plots of love, catastrophe, and reunion. The early Greek romances frequently took shape as a series of short tales. The *Love Romances* of Parthenius of Nicaea, who wrote during the reign of Augustus Caesar, is a collection of 36 prose stories of unhappy lovers. *The Milesian Tales* (no longer extant) was an extremely popular collection of erotic and ribald stories composed by Aristides of Miletus in the 2nd century BC and translated almost immediately into Latin. As the variety of these short narratives suggests, the Greeks were less insistent than earlier cultures that short fiction be predominantly didactic.

By comparison the contribution of the Romans to short narrative was small. Ovid's long poem, *Metamorphoses,* is basically a reshaping of over 100 short, popular tales into a thematic pattern. The other major fictional narratives to come out of Rome are novel-length works by Petronius (*Satyricon,* 1st century AD) and Apuleius (*The Golden Ass,* 2nd century AD). Like Ovid these men used potential short-story material as episodes within a larger whole. The Roman love of rhetoric, it seems, encouraged the development of longer and more comprehensive forms of expression. Regardless, the trend away from didacticism inaugurated by the Greeks was not reversed.

Middle Ages, Renaissance, and after. *Proliferation of forms.* The Middle Ages was a time of the proliferation, though not necessarily the refinement, of short narratives. The short tale became an important means of diversion and amusement. From the Dark Ages to the Renaissance, various cultures adopted short fiction for their own purposes. Even the aggressive, grim spirit of the invading Germanic barbarians was amenable to expression in short prose. The myths and sagas extant in Scandinavia and Iceland indicate the kinds of bleak and violent tales the invaders took with them into southern Europe.

In contrast, the romantic imagination and high spirits of the Celts remained manifest in their tales. Wherever they appeared—in Ireland, Wales, or Brittany—stories steeped in magic and splendour also appeared. This spirit, easily recognized in such Irish mythological tales as *Longes mac n-Uislenn* (probably 9th-century), infused the chivalric romances that developed somewhat later on the Continent. The romances usually addressed one of three "Matters": the "Matter of Britain" (stories of King Arthur and his knights), the "Matter of France" (the Charlemagne cycle), or the "Matter of Rome" (stories out of antiquity, such as "Pyramus and Thisbe," "Paris and Helen"). Many, but not all, of the romances are too long to be considered short stories. Two of the most influential contributors of short material to the "Matter of Britain" in the 12th century were Chrétien de Troyes and Marie de France. The latter was gifted as a creator of the short narrative poems known as the Breton lays. Only occasionally did a popular short romance like *Aucassin and Nicolette* (13th century) fail to address any of the three Matters.

Exempla and fabliaux

Also widely respected was the exemplum, a short didactic tale usually intended to dramatize or otherwise inspire model behaviour. Of all the exempla, the best known in the 11th and 12th centuries were the lives of the saints, some 200 of which are extant. The *Gesta Romanorum* ("Deeds of the Romans") offered skeletal plots of exempla that preachers could expand into moralistic stories for use in their sermons.

Among the common people of the late Middle Ages there appeared a literary movement counter to that of the romance and exemplum. Displaying a preference for common sense, secular humour, and sensuality, this movement accounted in a large way for the practical-minded animals in beast fables, the coarse and "merry" jestbooks, and the ribald fabliaux. All were important as short narratives, but perhaps the most intriguing of the three are the fabliaux. First appearing around the middle of the 12th century, fabliaux remained popular for 200 years, attracting the attention of Boccaccio and Chaucer. Some 160 fabliaux are extant, all in verse.

The framing circumstance

Often, the medieval storyteller—regardless of the kind of tale he preferred—relied on a framing circumstance that made possible the juxtaposition of several stories, each of them relatively autonomous. Since there was little emphasis on organic unity, most storytellers preferred a flexible format, one that allowed tales to be added or removed at random with little change in effect. Such a format is found in *The Seven Sages of Rome,* a collection of stories so popular that nearly every European country had its own translation. The framing circumstance in *The Seven Sages* involves a prince condemned to death; his advocates (the seven sages) relate a new story each day, thereby delaying the execution until his innocence is made known. This technique is clearly similar to that of *The Arabian Nights,* another collection to come out of the Middle Ages. The majority of the stories in *The Arabian Nights* are framed by the story of Scheherazade in "A Thousand and One Nights." Records indicate that the basis of this framing story was a medieval Persian collection, *Hezar Efsan* ("Thousand Romances," no longer extant). In both the Persian and Arabian versions of the frame, the clever Scheherazade avoids death by telling her king-husband a thousand stories. Though the framing device is identical in both versions, the original Persian stories within the frame were replaced or drastically altered as the collection was adapted by the Arabs during the Muslim Manlūk period (AD 1250–1517).

Boccaccio and Chaucer

Refinement. Short narrative received its most refined treatment in the Middle Ages from Chaucer and Boccaccio. Chaucer's versatility reflects the versatility of the age. In "The Miller's Tale" he artistically combines two fabliaux; in "The Nun's Priest's Tale" he draws upon material common to beast fables; in "The Pardoner's Tale" he creates a brilliantly revealing sermon, complete with a narrative exemplum. This short list hardly exhausts the catalogue of forms Chaucer experimented with. By relating tale to teller and by exploiting relationships among the various tellers, Chaucer endowed *The Canterbury Tales* with a unique, dramatic vitality.

Boccaccio's genius, geared more toward *narrative* than *drama,* is of a different sort. Where Chaucer reveals a character through actions and assertions, Boccaccio seems more interested in stories as pieces of action. With Boccaccio, the characters telling the stories, and usually the characters within, are of subordinate interest. Like Chaucer, Boccaccio frames his well-wrought tales in a metaphoric context. The trip to the shrine at Canterbury provides a meaningful backdrop against which Chaucer juxtaposes his earthy and pious characters. The frame of the *Decameron* (from the Greek *deka,* 10, and *hēmera,* day) has relevance as well: during the height of the Black Plague in Florence, Italy, 10 people meet and agree to amuse and divert each other by telling 10 stories each. Behind every story, in effect, is the inescapable presence of the Black Death. The *Decameron* is fashioned out of a variety of sources, including fabliaux, exempla, and short romances.

Spreading popularity. Immediately popular, the *Decameron* produced imitations nearly everywhere. In Italy alone, there appeared at least 50 writers of *novelle* (as short narratives were called) after Boccaccio.

Italian short narratives

Learning from the success and artistry of Boccaccio and, to a lesser degree, his contemporary Franco Sacchetti, Italian writers for three centuries kept the Western world supplied with short narratives. Sacchetti was no mere imitator of Boccaccio. More of a frank and unadorned realist, he wrote—or planned to write—300 stories (200 of the *Trecentonovelle* ["300 Short Stories"] are extant) dealing in a rather anecdotal way with ordinary Florentine life. Two other well-known narrative writers of the 14th century, Giovanni Fiorentino and Giovanni Sercambi, freely acknowledged their imitation of Boccaccio. In the 15th century Masuccio Salernitano's collection of 50 stories, *Il novellino* (1475), attracted much attention. Though verbosity often substitutes for eloquence in Masuccio's stories, they are witty and lively tales of lovers and clerics.

With Masuccio the popularity of short stories was just beginning to spread. Almost every Italian in the 16th century, it has been suggested, tried his hand at *novelle.* Matteo Bandello, the most influential and prolific writer, attempted nearly everything from brief histories and anecdotes to short romances, but he was most interested in tales of deception. Various other kinds of stories appeared. Agnolo Firenzuolo's popular *Ragionamenti diamore* ("The Reasoning of Love") is characterized by a graceful style unique in tales of ribaldry; Anton Francesco Doni included several tales of surprise and irony in his miscellany, *I marmi* ("The Marbles"); and Gianfrancesco Straparola experimented with common folktales and with dialects in his collection, *Le piacevoli notti* ("The Pleasant Nights"). In the early 17th century, Giambattista Basile attempted to infuse stock situations (often of the fairy-tale type, such as "Puss and Boots") with realistic details. The result was often remarkable—a tale of hags or princes with very real motives and feelings. Perhaps it is the amusing and diverting nature of Basile's collection of 50 stories that has reminded readers of Boccaccio. Or, it may be his use of a frame similar to that in the *Decameron.* Whatever the reason, Basile's *Cunto de li cunti* (1634; *The Story of Stories*) is traditionally linked with Boccaccio and referred to as *The Pentamerone* ("The Five Days"). Basile's similarities to Boccaccio suggest that in the 300 years between them the short story may have gained repute and circulation, but its basic shape and effect hardly changed.

This pattern was repeated in France, though the impetus provided by Boccaccio was not felt until the 15th century. A collection of 100 racy anecdotes, *Les Cent Nouvelles Nouvelles,* "The Hundred New Short Stories" (*c.* 1460), outwardly resembles the *Decameron.* Margaret of Angoulême's *Heptaméron* (1558–59; "The Seven Days"), an unfinished collection of 72 amorous tales, admits a similar indebtedness.

In the early 17th century Béroalde de Verville placed his own Rabelaisian tales within a banquet frame in a collection called *Le Moyen de parvenir,* "The Way of Succeeding" (*c.* 1610). Showing great narrative skill, Béroalde's stories are still very much in the tradition of Boccaccio; as a collection of framed stories, their main intent is to amuse and divert the reader.

As the most influential nation in Europe in the 15th and 16th centuries, Spain contributed to the proliferation of short prose fiction. Especially noteworthy are: Don Juan Manuel's collection of lively exempla *Libro de los enxiemplos del conde Lucanor et de Patronio* (1328–35), which antedates the *Decameron;* the anonymous story "The Abencerraje," which was interpolated into a pastoral novel of 1559; and, most importantly, Miguel de Cervantes' experimental *Novelas ejemplares* (1613; "Exemplary Novels"). Cervantes' short fictions vary in style and seriousness, but their single concern is clear: to explore the nature of man's secular existence. This focus was somewhat new for short fiction, heretofore either didactic or escapist.

A new concern

Despite the presence of these and other popular collections, short narrative in Spain was eventually overshadowed by a new form that began to emerge in the 16th century—the novel. Like the earlier Romans, the Spanish writers of the early Renaissance often incorporated short story material as episodes in a larger whole.

Decline of short fiction. The 17th and 18th centuries mark the temporary decline of short fiction. The causes of this phenomenon are many: the emergence of the novel; the failure of the Boccaccio tradition to produce in three centuries much more than variations or imitations of older, well-worn material; and a renaissant fascination with drama and poetry, the superior forms of classical antiquity. Another cause for the disappearance of major works of short fiction is suggested by the growing preference for journalistic sketches. The increasing awareness of other lands and the growing interest in social conditions (accommodated by a publication boom) produced a plethora of descriptive and biographical sketches. Although these journalistic elements later were incorporated in the fictional short story, for the time being fact held sway over the imagination. Travel books, criminal biographies, social description, sermons, and essays occupied the market. Only occasionally did a serious story find its way into print, and then it was usually a production of an established writer like Voltaire or Addison.

Perhaps the decline is clearest in England, where the short story had its least secure foothold. It took little to obscure the faint tradition established in the 16th and 17th centuries by the popular jestbooks, by the *Palace of Pleasure* (an anthology of stories, mostly European), and by the few rough stories written by Englishmen (*e.g.,* Barnabe Rich's *Farewell to Military Profession,* 1581).

During the Middle Ages short fiction had become primarily an amusing and diverting medium. The Renaissance and Enlightenment, however, made different demands of the form. The awakening concern with secular issues called for a new attention to actual conditions. Simply, the diverting stories were no longer relevant or viable. At first only the journalists and pamphleteers responded to the new demand. Short fiction disappeared, in effect, because it did not respond. When it did shake off its escapist trappings in the 19th century, it reappeared as the "modern short story." This was a new stage in the evolution of short fiction, one in which the short form undertook a new seriousness and gained a new vitality and respect.

Emergence of the modern short story. *The 19th century.* The modern short story emerged almost simultaneously in Germany, the United States, France, and Russia. In Germany there had been relatively little difference between the stories of the late 18th century and those in the older tradition of Boccaccio. In 1795 Goethe contributed a set of stories to Schiller's journal, *Die Horen,* that were obviously created with the *Decameron* in mind. Significantly, Goethe did not call them "short stories" (*Novellen*) although the term was available to him. Rather, he thought of them as "entertainments" for German travellers (*Unterhaltungen deutscher Ausgewanderten*). Friedrich Schlegel's early discussion of the short narrative form, appearing soon after Goethe's "entertainments," also focussed on Boccaccio (*Nachrichten von den poetischen Werken des G. Boccaccio,* 1801).

A new type of short fiction

But a new type of short fiction was near at hand—a type that accepted some of the realistic properties of popular journalism. In 1827, 32 years after publishing his own "entertainments," Goethe commented on the difference between the newly emergent story and the older kind. "What is a short story," he asked, "but an event which, though unheard of, has occurred? Many a work which passes in Germany under the title 'short story' is not a short story at all, but merely a tale or what else you would like to call it." Two influential critics, Christoph Wieland and Friedrich Schleiermacher, also argued that a short story properly concerned itself with events that actually happened or could happen. A short story, for them, had to be realistic.

Perhaps sensitive to this qualification, Heinrich von Kleist and E.T.A. Hoffmann called their short works on fabulous themes "tales" (*Erzählungen*). Somewhat like Poe, Kleist created an expression of human problems, partly metaphysical and partly psychological, by dramatizing man's confrontations with a fantastic, chaotic world. Hoffmann's intriguing tales of exotic places and of supernatural phenomena were very likely his most influential. Another important writer, Ludwig Tieck, explicitly rejected realism as the definitive element in a short story. As he noted in his preface to the 1829 collection of his works and as he demonstrated in his stories, Tieck envisioned the short story as primarily a matter of intensity and ironic inversion. A story did not have to be realistic in any outward sense, he claimed, so long as the chain of consequences was "entirely in keeping with character and circumstances." By allowing the writer to pursue an inner, and perhaps bizarre, reality and order, Tieck and the others kept the modern story open to nonjournalistic techniques.

In the United States, the short story, as in Germany, evolved in two strains. On the one hand there appeared the realistic story that sought objectively to deal with seemingly real places, events, or persons. The regionalist stories of the second half of the 19th century (including those by G.W. Cable, Bret Harte, Sarah Orne Jewett) are of this kind. On the other hand, there developed the impressionist story, a tale shaped and given meaning by the consciousness and psychological attitudes of the narrator. Predicated upon this element of subjectivity, these stories seem less objective and are less realistic in the outward sense. Of this sort are Poe's tales in which the hallucinations of a central character or narrator provide the details and facts of the story. Like the narrators in "The Tell-Tale Heart" (1843) and "The Imp of the Perverse" (1845), the narrator of "The Fall of the House of Usher" (1839) so distorts and transforms what he sees that the reader cannot hope to look objectively at the scene. Looking through an intermediary's eyes, the reader can see only the narrator's impressions of the scene.

Evolving in two strains

Some writers contributed to the development of both types of story. Washington Irving wrote several realistic sketches (*The Sketch-Book,* 1819–20; *The Alhambra,* 1832) in which he carefully recorded appearances and actions. Irving also wrote stories in which the details were taken not from ostensible reality but from within a character's mind. Much of the substance of "The Stout Gentleman" (1821), for example, is reshaped and recharged by the narrator's fertile imagination; "Rip Van Winkle" (1819) draws upon the symbolic surreality of Rip's dreams.

The short prose of Nathaniel Hawthorne illustrates that neither type of modern story, however, has exclusive rights to the use of symbol. On a few occasions, as in "My Kinsman, Major Molineux" (1832), Hawthorne's stories are about symbolic events as they are viewed subjectively by the central character. Hawthorne's greater gift, however, was for creating scenes, persons, and events that strike the reader as being actual historical facts and also as being rich in symbolic import. "Endicott and the Red Cross" (1837) may seem little more than a photographic sketch of a tableau out of history (the 17th-century Puritan leader cuts the red cross of St. George out of the colonial flag, the first act of rebellion against England), but the details are symbols of an underground of conflicting values and ideologies.

Several American writers, from Poe to James, were interested in the "impressionist" story that focusses on the impressions registered by events on the characters' minds,

The "impressionist" story

rather than the objective reality of the events themselves. In Herman Melville's "Bartleby the Scrivener" (1856) the narrator is a man who unintentionally reveals his own moral weaknesses through his telling of the story of Bartleby. Mark Twain's tales of animals ("The Celebrated Jumping Frog," 1865; "The Story of Old Ram," 1872; "Baker's Blue Jay Yarn," 1879), all impressionist stories, distort ostensible reality in a way that reflects on the men who are speaking. Ambrose Bierce's famous "An Occurrence at Owl Creek Bridge" (1891) is another example of this type of story in which the reader sees a mind at work—distorting, fabricating, and fantasizing—rather than an objective picture of actuality. In contrast, William Dean Howells usually sought an objectifying aesthetic distance. Though Howells was as interested in human psychology and behaviour as any of the impressionist writers, he did not want his details filtered through a biassed, and thus distorting, narrator. Impressionism, he felt, gave license for falsifications; in the hands of many writers of his day, it did in fact result in sentimental romanticizing.

But in other hands the impressionist technique could subtly delineate human responses. Henry James was such a writer. Throughout his prefaces to the New York edition of his works, the use of an interpreting "central intelligence" is constantly emphasized. "Again and again, on review," James observes, "the shorter things in especial that I have gathered into [the Edition] have ranged themselves not as my own impersonal account of the affair in hand, but as my account of somebody's impression of it." This use of a central intelligence, who is the "impersonal author's concrete deputy or delegate" in the story, allows James all the advantages of impressionism and, simultaneously, the freedom and mobility common to stories narrated by a disembodied voice.

In at least one way, 19th-century America resembled 16th-century Italy: there was an abundance of second- and third-rate short stories. And, yet, respect for the form grew substantially, and most of the great artists of the century were actively participating in its development. The seriousness with which many writers and readers regarded the short story is perhaps most clearly evident in the amount and kind of critical attention it received. James, Howells, Harte, Twain, Melville, and Hawthorne all discussed it as an art form, usually offering valuable insights, though sometimes shedding more light on their own work than on the art as a whole.

But the foremost American critic of the short story was Edgar Allan Poe. Himself a creator of influential impressionist techniques, Poe believed that the definitive characteristic of the short story was its unity of effect. "A skillful literary artist has constructed a tale," Poe wrote in his review of Hawthorne's *Twice-Told Tales* in 1842.

Poe's notion of the unity of effect

> If wise, he has not fashioned his thoughts to accommodate his incidents; but having conceived, with deliberate care, a certain unique or single *effect* to be wrought out, he then invents such incidents—he then combines such events as may best aid him in establishing this preconceived effect. If his very initial sentence tend not to the out-bringing of this effect, then he has failed in his first step. In the whole composition there should be no word written of which the tendency, direct or indirect, is not to the one pre-established design.

Poe's polemic primarily concerns craftsmanship and artistic integrity; it hardly prescribes limits on subject matter or dictates technique. As such, Poe's thesis leaves the story form open to experimentation and to growth while it demands that the form show evidence of artistic diligence and seriousness.

The new respect for the short story was also evident in France, as Henry James observed, when in 1844 Prosper "Mérimée with his handful of little stories was appointed to the French Academy." As illustrated by "Columbia" (1841) or "Carmen" (1845), which gained additional fame as an opera, Mérimée's stories are masterpieces of detached and dry observation, though the subject matter itself is often emotionally charged. Nineteenth-century France produced short stories as various as 19th-century America—although the impressionist tale was generally less common in France. (It is as if, not having an outstanding impressionist storyteller themselves, the French adopted Poe, who was being ignored by the critics in his own country.) The two major French impressionist writers were Charles Nodier, who experimented with symbolic fantasies, and Gérard de Nerval, whose collection *Les Filles du feu* (1854; "Daughters of Fire") grew out of recollections of his childhood. Artists primarily known for their work in other forms also attempted the short story—novelists like Honoré de Balzac and Gustave Flaubert and poets like Alfred de Vigny and Théophile Gautier.

French short stories

One of the most interesting writers of 19th-century France is Alphonse Daudet, whose stories reflect the spectrum of interest and techniques of the entire century. His earliest and most popular stories (*Lettres de mon moulin,* 1866; "Letters from My Mill") create a romantic, picturesque fantasy; his stories of the Franco-Prussian War (*Contes du Lundi,* 1873; "Monday's Tales") are more objectively realistic, and the sociological concern of his last works betrays his increasing interest in naturalistic determinism.

The greatest French storywriter, by far, is Guy de Maupassant, a master of the objective short story. Basically, Maupassant's stories are anecdotes that capture a revealing moment in the lives of middle class citizens. This crucial moment is typically recounted in a well-plotted design, though perhaps in some stories like "Boule de suif" (1880; "Ball of Tallow") and "The Necklace" (1881) the plot is too contrived, the reversing irony too neat, and the artifice too apparent. In other stories, like "The House of Madame Tellier" (1881), Maupassant's easy and fluid prose captures the innocence and the corruption of human behaviour.

A master of the objective story

During the first two decades of the 19th century in Russia, fable writing became a fad. By all accounts the most widely read fabulist was Ivan Krylov whose stories borrowed heavily from Aesop, La Fontaine, and various Germanic sources. If Krylov's tales made short prose popular in Russia, the stories of the revered poet Aleksandr Pushkin gained serious attention for the form. Somewhat like Mérimée in France (who was one of the first to translate Pushkin, Gogol, and Turgenev into French), Pushkin cultivated a detached, rather classical style for his stories of emotional conflicts (*The Queen of Spades,* 1834). Also very popular and respected was Mikhail Lermontov's "novel," *A Hero of Our Time* (1840), which actually consists of five stories that are more or less related.

Russian writers

But it is Nikolay Gogol who stands at the headwaters of the Russian short story; Dostoyevsky noted that all Russian short story writers "emerged from Gogol's overcoat," a punning allusion to the master's best known story. In a manner all his own, Gogol was developing impressionist techniques in Russia simultaneously with Poe in America. Gogol published his *Arabesques* (1835) five years before Poe collected some of his tales under a similar title. Like those of Poe, Gogol's tales of hallucination, confusing reality and dream, are among his best stories ("Nevsky Prospect" and "Diary of a Madman," both 1835). The single most influential story in the first half of the 19th century in Russia was undoubtedly Gogol's "Overcoat" (1842). Blending elements of realism (natural details from the characters' daily lives) with elements of fantasy (the central character returns as a ghost), Gogol's story seems to anticipate both the impressionism of Dostoyevsky's "Underground Man" and the realism of Tolstoy's "Ivan Ilich."

Ivan Turgenev appears, at first glance, antithetical to Gogol. In *A Sportsman's Notebook* (1852) Turgenev's simple use of language, his calm pace, and his restraint clearly differentiate him from Gogol. But like Gogol, Turgenev was more interested in capturing qualities of people and places than in building elaborate plots. A remaining difference between the two Russians, however, tends to make Turgenev more acceptable to 20th-century readers: Turgenev studiously avoided anything artificial. Though he may have brought into his realistic scenes a tale of a ghost ("Bezhin Meadow," 1852), he did not attempt to bring in a ghost (as Gogol had done in "The Overcoat"). In effect, Turgenev's allegiance was wholly to detached observation.

Developing some of the interests of Gogol, Fyodor Dostoyevsky experimented with the impressionist story. The early story "White Nights" (1848), for example, is a "Tale

of Love from the Reminiscence of a Dreamer" as the subtitle states; the title of one of his last stories, "The Dream of the Ridiculous Man" (1877), also echoes Poe and Gogol. Though sharing Dostoyevsky's interest in human motives, Leo Tolstoy used vastly different techniques. He usually sought psychological veracity through a more detached and, presumably, objective narrator ("The Death of Ivan Ilich," 1886; "The Kreutzer Sonata," 1891). Perhaps somewhat perplexed by Tolstoy's nonimpressionist means of capturing and delineating psychological impressions, Henry James pronounced Tolstoy the masterhand of the disconnection of method from matter.

The Russian master of the objective story was Anton Chekhov. No other storywriter so consistently as Chekhov turned out first-rate works. Though often compared to Maupassant, Chekhov is much less interested in constructing a well-plotted story; nothing much actually happens in Chekhov's stories, though much is revealed about his characters and the quality of their lives. While Maupassant focusses on event, Chekhov keeps his eye on character. Stories like "The Grasshopper" (1892), "The Darling" (1898), and "In the Ravine" (1900)—to name only three—all reveal Chekhov's perception, his compassion, and his subtle humour and irony. One critic says of Chekhov that he is no moralist—he simply says "you live badly, ladies and gentlemen," but his smile has the indulgence of a very wise man.

THE 20TH CENTURY

In the first half of the 20th century the appeal of the short story continued to grow. Literally hundreds of writers—including, as it seems, nearly every major dramatist, poet, and novelist—published thousands of excellent stories. William Faulkner suggested that writers often try their hand at poetry, find it too difficult, go on to the next most demanding form, the short story, fail at that, and only then settle for the novel. In the 20th century Germany, France, Russia, and the U.S. lost what had once appeared to be their exclusive domination of the form. Innovative and commanding writers emerged in countries that had previously exerted little influence on the genre: Sicily, for example, produced Luigi Pirandello; Czechoslovakia, Franz Kafka; Japan, Akutagawa Ryūnosuke; Argentina, Jorge Luis Borges. Literary journals with international circulation, such as Ford Madox Ford's *Transatlantic Review, Scribner's Magazine,* and Harriet Weaver's *Egoist,* provided a steady and prime exposure for young writers.

Increasing complexity of the short story

As the familiarity with it increased, the short story form itself became more varied and complex. The fundamental means of structuring a story underwent a significant change. The overwhelming or unique event that usually informed the 19th-century story fell out of favour with the storywriter of the early 20th century. He grew more interested in subtle actions and unspectacular events. Sherwood Anderson, one of the most influential U.S. writers of the early 20th century, observed that the common belief in his day was that stories had to be built around a plot, a notion that, in Anderson's opinion, appeared to poison all storytelling. His own aim was to achieve form, not plot, although form was more elusive and difficult. The record of the short story in the 20th century is dominated by this increased sensitivity to—and experimentation with—form. Although the popular writers of the century (like O. Henry in the U.S. and Paul Morand in France) may have continued to structure stories according to plot, the greater artists turned elsewhere for structure, frequently eliciting the response from cursory readers that "nothing *happens* in these stories." Narratives like Ernest Hemingway's "A Clean Well-Lighted Place" may seem to have no structure at all, so little physical action develops; but stories of this kind are actually structured around a psychological, rather than physical, conflict. In several of Hemingway's stories (as in many by D.H. Lawrence, Katherine Mansfield, and others), physical action and event are unimportant except insofar as the actions reveal the psychological underpinnings of the story. Stories came to be structured, also, in accordance with an underlying archetypal model: the specific plot and characters are important insofar as they allude to a traditional plot or figure, or to patterns that have recurred with wide implications in the history of mankind. Katherine Anne Porter's "Flowering Judas," for example, echoes and ironically inverts the traditional Christian legend. Still other stories are formed by means of motif, usually a thematic repetition of an image or detail that represents the dominant idea of the story. "The Dead," the final story in James Joyce's *Dubliners,* builds from a casual mention of death and snow early in the story to a culminating paragraph that links them in a profound vision. Seldom, of course, is the specific structure of one story appropriate for a different story. Faulkner, for example, used the traditional pattern of the knightly quest (in an ironic way) for his story "Was," but for "Barn Burning" he relied on a psychologically organic form to reveal the story of young Sarty Snopes.

The decline of the plot

No single form provided the 20th-century writer with the answer to structural problems. As the primary structuring agent, spectacular and suspenseful action was rather universally rejected around midcentury since motion pictures and television could present it much more vividly. As the periodicals that had supplied escapist stories to mass audiences declined, the short story became the favoured form of a smaller but intellectually more demanding readership. The Argentine Borges, for example, attracted an international following with his *Ficciones,* stories that involved the reader in dazzling displays of erudition and imagination, unlike anything previously encountered in the genre. Similarly, the American Donald Barthelme's composition consisted of bits and pieces of, *e.g.,* television commercials, political speeches, literary allusions, eavesdropped conversations, graphic symbols, dialogue from Hollywood movies—all interspersed with his own original prose in a manner that defied easy comprehension and yet compelled the full attention of the reader. The short story also lent itself to the rhetoric of student protest in the 1960s and was found in a bewildering variety of mixed-media forms in the "underground" press that publicized this life style throughout the world. In his deep concern with such a fundamental matter as form, the 20th-century writer unwittingly affirmed the maturation and popularity of the genre; only a secure and valued (not to mention flexible) genre could withstand and, moreover, encourage such experimentation. (A.J.H.)

DRAMA

Dramatic literature

The term dramatic literature implies a contradiction in that "literature" originally meant something written and "drama" meant something performed. Most of the problems, and much of the interest, in the study of dramatic literature stem from this contradiction. Even though a play may be appreciated solely for its qualities as writing, greater rewards probably accrue to those who remain alert to the volatility of the play as a whole.

In order to appreciate this complexity in drama, however, each of its elements—acting, directing, staging, etc.—should be studied, so that its relationship to all the others can be fully understood. It is the purpose of this section to study drama with particular attention to what the playwright sets down. A similar approach is taken in the following sections on the two main types of dramatic literature, *Tragedy,* and *Comedy.* The history of dramatic literature is discussed in such articles as THEATRE, THE HISTORY OF WESTERN; and EAST ASIAN ARTS; as well as in articles on the history of literature such as LITERATURE, THE HISTORY OF WESTERN. Regional studies of both the theatre and literature will be found in articles such as SOUTH ASIAN ARTS; and AFRICAN ARTS.

GENERAL CHARACTERISTICS

From the inception of a play in the mind of its author to the image of it that an audience takes away from the theatre, many hands and many physical elements help to bring it to life. Questions therefore arise as to what is and what is not essential to it. Is a play what its author thought he was writing, or the words he wrote? Is a play the way in which those words are intended to be embodied, or their actual interpretation by a director and his actors on a particular stage? Is a play in part the expectation an audience brings to the theatre, or is it the real response to what is seen and heard? Since drama is such a complex process of communication, its study and evaluation is as uncertain as it is mercurial.

Essential elements of a play

All plays depend upon a general agreement by all participants—author, actors, and audience—to accept the operation of theatre and the conventions associated with it, just as players and spectators accept the rules of a game. Drama is a decidedly unreal activity, which can be indulged only if everyone involved admits it. Here lies some of the fascination of its study. For one test of great drama is how far it can take the spectator beyond his own immediate reality and to what use this imaginative release can be put. But the student of drama must know the rules with which the players began the game before he can make this kind of judgment. These rules may be conventions of writing, acting, or audience expectation. Only when all conventions are working together smoothly in synthesis, and the make-believe of the experience is enjoyed passionately with mind and emotion, can great drama be seen for what it is: the combined work of a good playwright, good players, and a good audience who have come together in the best possible physical circumstances.

Drama in some form is found in almost every society, primitive and civilized, and has served a wide variety of functions in the community. There are, for example, records of a sacred drama in Egypt 2,000 years before Christ, and Thespis in the 6th century BC in ancient Greece is accorded the distinction of being the first known playwright. Elements of drama such as mime and dance, costume and decor long preceded the introduction of words and the literary sophistication now associated with a play. Moreover, such basic elements were not superseded by words, merely enhanced by them. Nevertheless, it is only when a playscript assumes a disciplinary control over the dramatic experience that the student of drama gains measurable evidence of what was intended to constitute the play. Only then can dramatic literature be discussed as such.

The late arrival of words in drama

The texts of plays indicate the different functions they served at different times. Some plays embraced nearly the whole community in a specifically religious celebration, as when all the male citizens of a Greek city-state came together to honour their gods; or when the annual Feast of Corpus Christi was celebrated with the great medieval Christian mystery cycles. On the other hand, the ceremonious temple ritual of the early Nō drama of Japan was performed at religious festivals only for the feudal aristocracy. But the drama may also serve a more directly didactic purpose, as did the morality plays of the later Middle Ages, some 19th-century melodramas, and the 20th-century discussion plays of George Bernard Shaw and Bertolt Brecht. Plays can satirize society, or they can gently illuminate human weakness; they can divine the greatness and the limitations of man in tragedy, or, in modern naturalistic playwriting, probe his mind. Drama is the most wide-ranging of all the arts: it not only represents life but also is a way of seeing it. And it repeatedly proves Dr. Samuel Johnson's contention that there can be no certain limit to the modes of composition open to the dramatist.

Common elements of drama. Despite the immense diversity of drama as a cultural activity, all plays have certain elements in common. For one thing, drama can never become a "private" statement—in the way a novel or a poem may be—without ceasing to be meaningful theatre. The characters may be superhuman and godlike in appearance, speech, and deed or grotesque and ridiculous, perhaps even puppets, but as long as they behave in even vaguely recognizable human ways the spectator can understand them. Only if they are too abstract do they cease to communicate as theatre. Thus, the figure of Death in medieval drama reasons like a human being, and a god in Greek tragedy or in Shakespeare talks like any mortal. A play, therefore, tells its tale by the imitation of human behaviour. The remoteness or nearness of that behaviour to the real life of the audience can importantly affect the response of that audience: it may be in awe of what it sees, or it may laugh with detached superiority at clownish antics, or it may feel sympathy. These differences of alienation or empathy are important, because it is by opening or closing this aesthetic gap between the stage and the audience that a dramatist is able to control the spectator's experience of the play and give it purpose.

The second essential is implicit in the first. Although static figures may be as meaningfully symbolic on a stage as in a painting, the deeper revelation of character, as well as the all-important control of the audience's responses, depends upon a dynamic presentation of the figures in action. A situation must be represented on the stage, one recognizable and believable to a degree, which will animate the figures as it would in life. Some argue that action is the primary factor in drama, and that character cannot emerge without it. Since no play exists without a situation, it appears impossible to detach the idea of a character from the situation in which he is placed, though it may seem possible after the experience of the whole play. Whether the playwright conceives character before situation, or vice versa, is arbitrary. More relevant are the scope and scale of the character-in-situation—whether, for example, it is man confronting God or man confronting his wife—for that comes closer to the kind of experience the play is offering its audience. Even here one must beware of passing hasty judgment, for it may be that the grandest design for heroic tragedy may be less affecting than the teasing vision of human madness portrayed in a good farce.

A third factor is style. Every play prescribes its own style, though it will be influenced by the traditions of its theatre and the physical conditions of performance. Style is not something imposed by actors upon the text after it is written, nor is it superficial to the business of the play. Rather, it is self-evident that a play will not communicate without it. Indeed, many a successful play has style and little else. By "style," therefore, is implied the whole mood and spirit of the play, its degree of fantasy or realism, its quality of ritualism or illusion, and the way in which these qualities are signalled by the directions, explicit or implicit, in the text of the play. In its finer detail, a play's style controls the kind of gesture and movement of the actor, as well as his tone of speech, its pace and inflexion. In this way the attitude of the audience is prepared also: nothing is more disconcerting than to be misled into expecting either a comedy or a tragedy and to find the opposite, although some great plays deliberately introduce elements of both. By means of signals of style, the audience may be led to expect that the play will follow known paths, and the pattern of the play will regularly echo the rhythm of response in the auditorium. Drama is a conventional game, and spectators cannot participate if the rules are constantly broken.

The role of style

By presenting animate characters in a situation with a certain style and according to a given pattern, a playwright will endeavour to communicate his thoughts and feelings and have his audience consider his ideas or reproduce the emotion that drove him to write as he did. In theatrical communication, however, audiences remain living and independent participants. In the process of performance, an actor has the duty of interpreting his author for the people watching him, and will expect to receive "feedback" in turn. The author must reckon with this in his writing. Ideas will not be accepted, perhaps, if they are offered forthrightly; and great dramatists who are intent on furthering social or political ideas, such as Henrik Ibsen, George Bernard Shaw, and Bertolt Brecht, quickly learned methods of having the spectator reason the ideas for himself as part of his response to the play. Nor will passions necessarily be aroused if overstatement of feeling

("sentimentality") is used without a due balance of thinking and even the detachment of laughter: Shakespeare and Chekhov are two outstanding examples in Western drama of writers who achieved an exquisite balance of pathos with comedy in order to ensure the affective function of their plays.

Dramatic expression. The language of drama can range between great extremes: on the one hand, an intensely theatrical and ritualistic manner; and on the other, an almost exact reproduction of real life of the kind commonly associated with motion picture and television drama. In the ritualistic drama of ancient Greece, the playwrights wrote in verse, and it may be assumed that their actors rendered this in an incantatory speech halfway between speech and song. Both the popular and the coterie drama of the Chinese and Japanese theatre were also essentially operatic, with a lyrical dialogue accompanied by music and chanted rhythmically. The effect of such rhythmical delivery of the words was to lift the mood of the whole theatre onto the level of religious worship. Verse is employed in other drama that is conventionally elevated, like the Christian drama of the Middle Ages, the tragedy of the English Renaissance, the heroic Neoclassical tragedies of 17th-century France by Pierre Corneille and Jean Racine, the Romantic lyricism of Goethe and Schiller, and modern attempts at a revival of a religious theatre like those of T.S. Eliot. Indeed, plays written in prose dialogue were at one time comparatively rare, and then associated essentially with the comic stage. Only at the end of the 19th century, when naturalistic realism became the mode, were characters in dramas expected to speak as well as behave as in real life.

The role of verse

Elevation is not the whole rationale behind the use of verse in drama. Some critics maintain that a playwright can exercise better control both over the speech and movement of his actors and over the responses of his audience by using the more subtle tones and rhythms of good poetry. The loose, idiomatic rhythms of ordinary conversation, it is argued, give both actor and spectator too much freedom of interpretation and response. Certainly, the aural, kinetic, and emotive directives in verse are more direct than prose, though, in the hands of a master of prose dialogue like Shaw or Chekhov, prose can also share these qualities. Even more certain, the "aesthetic distance" of the stage, or the degree of unreality and make-believe required to release the imagination, is considerably assisted if the play uses elements of verse, like rhythm and rhyme, not found in ordinary speech. Thus, verse drama may embrace a wide variety of nonrealistic aural and visual devices: Greek tragic choric speech provided a philosophical commentary upon the action, which at the same time drew the audience lyrically into the mood of the play. In the drama of India, a verse accompaniment made the actors' highly stylized system of symbolic gestures of head and eyes, arms and fingers a harmonious whole. The tragic soliloquy in Shakespeare permitted the hero, alone on the stage with his audience, to review his thoughts aloud in the persuasive terms of poetry; thus, the soliloquy was not a stopping place in the action but rather an engrossing moment of drama when the spectator's mind could leap forward.

Dramatic structure. The elements of a play do not combine naturally to create a dramatic experience but, rather, are made to work together through the structure of a play, a major factor in the total impact of the experience. A playwright will determine the shape of a play in part according to the conditions in which it will be performed: how long should it take to engage an audience's interest and sustain it? How long can an audience remain in their seats? Is the audience sitting in one place for the duration of performance, or is it moving from one pageant stage to the next, as in some medieval festivals? Structure is also dictated by the particular demands of the material to be dramatized: a revue sketch that turns on a single joke will differ in shape from a religious cycle, which may portray the whole history of mankind from the Creation to the Last Judgment. A realistic drama may require a good deal of exposition of the backgrounds and memories of the characters, while in a chronicle play the playwright may tell the whole story episodically from its beginning to the end. There is one general rule, as Aristotle originally suggested in his *Poetics:* a play must be long enough to supply the information an audience needs to be interested and to generate the experience of tragedy, or comedy, on the senses and imagination.

In the majority of plays it is necessary to establish a conventional code of place and time. In a play in which the stage must closely approximate reality, the location of the action will be precisely identified, and the scenic representation on stage must confirm the illusion. In such a play, stage time will follow chronological time almost exactly; and if the drama is broken into three, four, or five acts, the spectator will expect each change of scene to adjust the clock or the calendar. But the theatre has rarely expected realism, and by its nature it allows an extraordinary freedom to the playwright in symbolizing location and duration: as Dr. Samuel Johnson observed in his discussion of this freedom in Shakespeare, the spectators always allow the play to manipulate the imagination. It is sufficient for the witches in *Macbeth* to remark their "heath" with its "fog and filthy air" for their location to be accepted on a stage without scenery; and when Lady Macbeth later is seen alone reading a letter, she is without hesitation understood to be in surroundings appropriate to the wife of a Scottish nobleman. Simple stage symbolism may assist the imagination, whether the altar of the gods situated in the centre of the Greek *orchēstra,* a strip of red cloth to represent the Red Sea in a medieval miracle play, or a chair on which the Tibetan performer stands to represent a mountain. With this degree of fantasy, it is no wonder that the theatre can manipulate time as freely, passing from the past to the future, from this world to the next, and from reality to dream.

Conventions of time and place

It is questionable, therefore, whether the notion of "action" in a play describes what happens on the stage or what is recreated in the mind of the audience. Certainly it has little to do with merely physical activity by the players. Rather, anything that urges forward the audience's image of the play and encourages the growth of its imagination is a valid part of the play's action. Thus, it was sufficient for the ancient Greek dramatist Aeschylus to have only two speaking male actors who wore various masks, typed for sex, age, class, and facial expression. In the Italian 16th- and 17th-century commedia dell'arte, the standard characters Pantalone and Arlecchino, each wearing his traditional costume and mask, appeared in play after play and were immediately recognized, so that an audience could anticipate the behaviour of the grasping old merchant and his rascally servant. On a less obvious level, a speech that in reading seems to contribute nothing to the action of the play can provide in performance a striking stimulus to the audience's sense of the action, its direction and meaning. Thus, both the Greek chorus and the Elizabethan actor in soliloquy might be seen to "do" nothing, but their intimate speeches of evaluation and reassessment teach the spectator how to think and feel about the action of the main stage and lend great weight to the events of the play. For drama is a reactive art, moving constantly in time, and any convention that promotes a deep response while conserving precious time is of immeasurable value.

Conventions of action

DRAMA AS AN EXPRESSION OF A CULTURE

In spite of the wide divergencies in purpose and convention of plays as diverse as the popular kabuki of Japan and the coterie comedies of the Restoration in England, a Javanese puppet play and a modern social drama by the contemporary American dramatist Arthur Miller, all forms of dramatic literature have some points in common. Differences between plays arise from differences in conditions of performance, in local conventions, in the purpose of theatre within the community, and in cultural history. Of these, the cultural background is the most important, if the most elusive. It is cultural difference that makes the drama of the East immediately distinguishable from that of the West.

East–West differences. Oriental drama consists chiefly of the classical theatre of Hindu India and its derivatives in Malaya and of Burma, Thailand, China, Japan, Java,

and Bali. It was at its peak during the period known in the West as the Middle Ages and the Renaissance. Stable and conservative, perpetuating its customs with reverence, Oriental culture showed little of the interest in chronology and advancement shown by the West and placed little emphasis on authors and their individual achievements. Thus the origins of the drama of the Orient are lost in time, although its themes and characteristic styles probably remain much the same as before records were kept. The slow-paced, self-contained civilizations of the East have only recently been affected by Western theatre, just as the West has only recently become conscious of the theatrical wealth of the East and what it could do to fertilize the modern theatre (as in the 20th-century experimental drama of William Butler Yeats and Thornton Wilder in English, of Paul Claudel and Antonin Artaud in French, and of Bertolt Brecht in German).

Stylization of Oriental theatre

In its representation of life, classical Oriental drama is the most conventional and nonrealistic in world theatre. Performed over the centuries by actors devoted selflessly to the profession of a traditional art, conventions of performance became highly stylized, and traditions of characterization and play structure became formalized to a point of exceptional finesse, subtlety, and sophistication. In Oriental drama all the elements of the performing arts are made by usage to combine to perfection: dance and mime, speech and song, narrative and poetry. The display and studied gestures of the actors, their refined dance patterns, and the all-pervasive instrumental accompaniment to the voices of the players and the action of the play, suggest to Western eyes an exquisite combination of ballet with opera, in which the written text assumes a subordinate role. In this drama, place could be shifted with a license that would have astonished the most romantic of Elizabethan dramatists, the action could leap back in time in a way reminiscent of the "flashback" of the modern cinema, and events could be telescoped with the abandon of modern expressionism. This extreme theatricality lent an imaginative freedom to its artists and audiences upon which great theatre could thrive. Significantly, most Oriental cultures also nourished a puppet theatre, in which stylization of character, action, and staging were particularly suitable to marionettes. In the classical puppet theatre of Japan, the *bunraku,* the elocutionary art of a chanted narration and the manipulative skill with the dolls diminished the emphasis on the script except in the work of the 17th-century master Chikamatsu, who enjoyed a creative freedom in writing for puppets rather than for the actors of the Kabuki. By contrast, Western drama during and after the Renaissance has offered increasing realism, not only in decor and costume but also in the treatment of character and situation.

It is generally thought that Oriental drama, like that of the West, had its beginnings in religious festivals. Dramatists retained the moral tone of religious drama while using popular legendary stories to imbue their plays with a romantic and sometimes sensational quality. This was never the sensationalism of novelty that Western dramatists sometimes used: Eastern invention is merely a variation on what is already familiar, so that the slightest changes of emphasis could give pleasure to the cognoscenti. This kind of subtlety is not unlike that found in the repeatedly depicted myths of Greek tragedy. What is always missing in Oriental drama is that restlessness for change characteristic of modern Western drama. In the West, religious questioning, spiritual disunity, and a belief in the individual vision combined finally with commercial pressures to produce comparatively rapid changes. None of the moral probing of Greek tragedy, the character psychology of Shakespeare and Racine, the social and spiritual criticism of Ibsen and Strindberg, nor the contemporary drama of shock and argument, is imaginable in the classical drama of the East.

Drama in Western cultures. The form and style of ancient Greek tragedy, which flowered in the 5th century BC in Athens, was dictated by its ritual origins and by its performance in the great dramatic competitions of the spring and winter festivals of Dionysus. Participation in ritual requires that the audience largely knows what to expect. Ritual dramas were written on the same legendary stories of Greek heroes in festival after festival. Each new drama provided the spectators with a reassessment of the meaning of the legend along with a corporate religious exercise. Thus, the chorus of Greek tragedy played an important part in conveying the dramatist's intention. The chorus not only provided a commentary on the action but also guided the moral and religious thought and emotion of the audience throughout the play: for Aeschylus (*c.* 525–456 BC) and Sophocles (*c.* 496–406 BC) it might be said that the chorus *was* the play, and even for Euripides (*c.* 480–406 BC) it remained lyrically powerful. Other elements of performance also controlled the dramatist in the form and style he could use in these plays: in particular, the great size of the Greek arena demanded that the players make grand but simple gestures and intone a poetry that could never approach modern conversational dialogue. Today, the superhuman characters of these plays, Agamemnon and Clytemnestra, Orestes and Electra, Oedipus and Antigone, seem unreal, for they display little "characterization" in the modern sense and their fates are sealed. Nevertheless, these great operatic tableaux, built, as one critic has said, for weight and not speed, were evidently able to carry their huge audiences to a catharsis of feeling. It is a mark of the piety of those audiences that the same reverent festivals supported a leavening of satyr-plays and comedies, bawdy and irreverent comments on the themes of the tragedies, culminating in the wildly inventive satires of Aristophanes (*c.* 445–*c.* 385 BC.)

Use of familiar legends

The study of Greek drama demonstrates how the ritual function of theatre shapes both play and performance. This ritual aspect was lost when the Romans assimilated Greek tragedy and comedy. The Roman comedies of Plautus (*c.* 254–184 BC) and Terence (*c.* 186/185–159 BC) were brilliant but inoffensive entertainments, while the oratorical tragedies of Seneca (*c.* 4 BC–AD 65) on themes from the Greek were written probably only to be read by the ruling caste. Nevertheless, some of the dramatic techniques of these playwrights influenced the shape and content of plays of later times. The bold prototype characters of Plautus (the boasting soldier, the old miser, the rascally parasite), with the intricacies of his farcical plotting, and the sensational content and stoical attitudes of Seneca's drama reappeared centuries later when classical literature was rediscovered.

Establishment of prototypes of characters

Western drama had a new beginning in the medieval church, and, again, the texts reflect the ritual function of the theatre in society. The Easter liturgy, the climax of the Christian calendar, explains much of the form of medieval drama as it developed into the giant mystery cycles. From at least the 10th century the clerics of the church enacted the simple Latin liturgy of the *Quem quaeritis?* (literally "Whom do you seek?"), the account of the visit to Jesus Christ's tomb by the three Marys, who are asked this question by an angel. The liturgical form of Lent and the Passion, indeed, embodies the drama of the Resurrection to be shared mutually by actor-priest and audience-congregation. When the Feast of Corpus Christi was instituted in 1246, the great lay cycles of Biblical plays (the mystery or miracle cycles) developed rapidly, eventually treating the whole story of man from the Creation to the Last Judgment, with the Crucifixion still the climax of the experience. The other influence controlling their form and style was their manner of performance. The vast quantity of material that made up the story was broken into many short plays, and each was played on its own stage in the vernacular by members of the craft guilds. Thus, the authors of these dramas gave their audience not a mass communal experience, as the Greek dramatists had done, but rather many small and intimate dramatizations of the Bible story. In stylized and alliterative poetry, they mixed awesome events with moments of extraordinary simplicity, embodying local details, familiar touches of behaviour, and the comedy and the cruelty of medieval life. Their drama consists of strong and broad contrasts, huge in perspective but meaningful in human terms, religious and appropriately didactic in content and yet popular in its manner of reaching its simple audiences.

In an account of dramatic literature, the ebullient but

unscripted farces and romances of the commedia dell'arte properly have no place, but much in it became the basis of succeeding comedy. Two elements are worth noting.

The influence of improvisation

First, the improvisational spirit of the commedia troupes, in which the actor would invent words and comic business (*lazzi*) to meet the occasion of the play and the audience he faced, encouraged a spontaneity in the action that has affected the writing and playing of Western comedy ever since. Second, basic types of comic character derived from the central characters, who reappeared in the same masks in play after play. As these characters became well known everywhere, dramatists could rely on their audience to respond to them in predictable fashion. Their masks stylized the whole play and allowed the spectator freedom to laugh at the unreality of the action. An understanding of the commedia illuminates a great deal in the written comedies of Shakespeare in England, of Molière and Marivaux in France, and of Goldoni and Gozzi in Italy.

In the 16th century, England and Spain provided all the conditions necessary for a drama that could rival ancient Greek drama in scope and subtlety. In both nations, there were public as well as private playhouses, audiences of avid imagination, a developing language that invited its poetic expansion, a rapid growth of professional acting companies, and a simple but flexible stage. All these factors combined to provide the dramatist with an opportunity to create a varied and exploratory new drama of outstanding interest. In Elizabethan London, dramatists wrote in an extraordinary range of dramatic genres, from native comedy and farce to Senecan tragedy, from didactic morality plays to popular chronicle plays and tragicomedies, all before the advent of Shakespeare (1564–1616). Although Shakespeare developed certain genres, such as the chronicle play and the tragedy, to a high degree, Elizabethan dramatists characteristically used a medley of styles. With the exception of Ben Jonson (1572/73–1637) and a few others, playwrights mixed their ingredients without regard for classical rule. The result was a rich body of drama, exciting and experimental in character. A host of new devices were tested, mixing laughter and passion; shifting focus and perspective by slipping from verse to prose and back again; extending the use of the popular clown; exploiting the double values implicit in boy actors playing the parts of girls; exploring the role of the actor in and out of character; but, above all, developing an extraordinarily flexible dramatic poetry. These dramatists produced a visually and aurally exciting hybrid drama that could stress every subtlety of thought and feeling. It is not surprising that they selected their themes from every Renaissance problem of order and authority, of passion and reason, of good and evil and explored every comic attitude to people and society with unsurpassed vigour and vision.

Quite independently in Spain, dramatists embarked upon a parallel development of genres ranging from popular farce to chivalric tragedy. The hundreds of plays of Spain's greatest playwright, Lope de Vega (1562–1635), cover every subject from social satire to religion with equal exuberance.

French drama in the 17th century

The drama of Paris of the 17th century, however, was determined by two extremes of dramatic influence. On the one hand, some playwrights developed a tragedy rigidly based in form upon Neoclassical notions of Aristotelian unity, controlled by verse that is more regular than that of the Spanish or English dramatists. On the other hand, the French theatre developed a comedy strongly reflecting the work of the itinerant troupes of the commedia dell'arte. The Aristotelian influence resulted in the plays of Pierre Corneille (1606–84) and Jean Racine (1639–99), tragedies of honour using classical themes, highly sophisticated theatrical instruments capable of searching deeply into character and motive, and capable of creating the powerful tension of a tightly controlled plot. The other influence produced the brilliant plays of Molière (1622–73), whose training as an actor in the masked and balletic commedia tradition supplied him with a perfect mode for a more sophisticated comedy. Molière's work established the norm of French comedy, bold in plotting, exquisite in style, irresistible in comic suggestion. Soon after, upon the return of Charles II to the throne of England in 1660, a revival of theatre started the English drama on a new course. Wits such as William Wycherley (1640–1716) and William Congreve (1670–1729) wrote for the intimate playhouses of the Restoration and an unusually homogeneous coterie audience of the court circle. They developed a "comedy of manners," replete with social jokes that the actor, author, and spectator could share—a unique phase in the history of drama. These plays started a characteristic style of English domestic comedy still recognizable in London comedy today.

German dramatists of the later part of the 18th century achieved stature through a quite different type of play: Johann Wolfgang von Goethe (1749–1832), Johann Christoph Friedrich von Schiller (1759–1805), and others of the passionate, poetic Sturm und Drang ("storm and stress") movement tried to echo the more romantic tendencies in Shakespeare's plays. Dramatists of the 19th century, however, lacking the discipline of classical form, wrote derivative melodramas that varied widely in quality, often degenerating into mere sensationalism. Melodrama rapidly became the staple of the theatre across Europe and America. Bold in plotting and characterization, simple in its evangelical belief that virtue will triumph and providence always intervene, it pleased vast popular audiences and was arguably the most prolific and successful drama in the history of the theatre. Certainly, melodrama's elements of essential theatre should not be ignored by those interested in drama as a social phenomenon. At least melodramas encouraged an expansion of theatre audiences ready for the most recent phase in dramatic history.

Melodrama

The time grew ripe for a new and more adult drama at the end of the 19th century. As novelists developed greater naturalism in both content and style, dramatists too looked to new and more realistic departures: the dialectical comedies of ideas of George Bernard Shaw (1856–1950); the problem plays associated with Henrik Ibsen (1828–1906); the more lyrical social portraits of Anton Chekhov (1860–1904); the fiercely personal, social, and spiritual visions of August Strindberg (1849–1912). These dramatists began by staging the speech and behaviour of real life, in devoted detail, but became more interested in the symbolic and poetic revelation of the human condition. Where Ibsen began by modelling his tightly structured dramas of man in society upon the formula for the "well made" play, which carefully controlled the audience's interest to the final curtain, Strindberg, a generation later, developed a free psychological and religious dream play that bordered on Expressionism. As sophisticated audiences grew interested more in causes rather than in effects, the great European playwrights of the turn of the century mixed their realism increasingly with symbolism. Thus the Naturalistic movement in drama, though still not dead, had a short but vigorous life. Its leaders freed the drama of the 20th century to pursue every kind of style, and subsequent dramatists have been wildly experimental. The playwright today can adopt any dramatic mode, mixing his effects to shock the spectator into an awareness of himself, his beliefs, and his environment.

Drama in Eastern cultures. Because of its inborn conservatism, the dramatic literature of the East does not show such diversity, despite its variety of cultures and subcultures. The major features of Oriental drama may be seen in the three great classical sources of India, China, and Japan. The simplicity of the Indian stage, a platform erected for the occasion in a palace or a courtyard, like the simplicity of the Elizabethan stage, lent great freedom to the imagination of the playwright. In the plays of India's greatest playwright, Kālidāsa (probably 4th century AD), there is an exquisite refinement of detail in presentation. His delicate romantic tales leap time and place by simple suggestion and mingle courtly humour and light-hearted wit with charming sentiment and religious piety. Quite untrammelled by realism, lyrical in tone and refined in feeling, his fanciful love and adventure stories completely justify their function as pure entertainment. His plots are without the pain of reality, and his characters never descend from the ideal: such poetic drama is entirely appropriate to the Hindu aesthetic of blissful idealism in art.

Indian theatre

Some contrast may be felt between the idealistic style of the Sanskrit drama and the broader, less courtly man-

ner of the Chinese and its derivatives in Southeast Asia. These plays cover a large variety of subjects and styles, but all combine music, speech, song, and dance, as does all Oriental drama. Heroic legends, pathetic moral stories, and brilliant farces all blended spectacle and lyricism and were as acceptable to a sophisticated court audience as to a popular street audience. The most important Chinese plays stem from the Yüan dynasty (1206–1368), in which an episodic narrative is carefully structured and unified. Each scene introduces a song whose lines have a single rhyme, usually performed by one singer, with a code of symbolic gestures and intonations that has been refined to an extreme. The plays have strongly typed heroes and villains, simple plots, scenes of bold emotion, and moments of pure mime. Chinese drama avoided both the crudity of European melodrama and the esotericism of Western coterie drama.

Nō and Kabuki

The drama of Japan may be said to embrace both. There, the exquisite artistry of gesture and mime, and the symbolism of setting and costume, took two major directions. The Nō drama, emerging from religious ritual, maintained a special refinement appropriate to its origins and its aristocratic audiences; the Kabuki (its name suggesting its composition: *ka,* "singing"; *bu,* "dancing"; *ki,* "acting") in the 17th century became Japan's popular drama. Nō theatre is reminiscent of the religious tragedy of the Greeks in the remoteness of its legendary content, in its masked heroic characters, in its limit of two actors and a chorus, and in the static, oratorical majesty of its style. The Kabuki, on the other hand, finds its material in domestic stories and in popular history, and the actors, without masks, move and speak more freely, without seeming to be realistic. The Kabuki plays are less rarefied and are often fiercely energetic and wildly emotional as befitting their presentation before a broader audience. The written text of the Nō play is highly poetic and pious in tone, compressed in its imaginative ideas, fastidious and restrained in verbal expression, and formal in its sparse plotting; the text of a Kabuki play lends plentiful opportunities for spectacle, sensation, and melodrama. In the Kabuki there can be moments of realism, but also whole episodes of mime and acrobatics; there can be moments of slapstick, but also moments of violent passion. In all, the words are subordinate to performance in the Kabuki.

Drama and communal belief. The drama that is most meaningful and pertinent to its society is that which arises from it and is not imposed upon it. The religious drama of ancient Greece, the temple drama of early India and Japan, the mystery cycles of medieval Europe, all have in common more than their religious content: when the theatre is a place of worship, its drama goes to the roots of belief in a particular community. The dramatic experience becomes a natural extension of man's life both as an individual and as a social being. The content of the mystery cycles speaks formally for the orthodox dogma of the church, thus seeming to place the plays at the centre of medieval life, like the church itself. Within such a comprehensive scheme, particular needs could be satisfied by comic or pathetic demonstration; for example, such a crucial belief as that of the Virgin Birth of Jesus was presented in the York (England) cycle of mystery plays, of the 14th–16th centuries, with a nicely balanced didacticism when Joseph wonders how a man of his age could have got Mary with child and an Angel explains what has happened; the humour reflects the simplicity of the audience and at the same time indicates the perfect faith that permitted the near-blasphemy of the joke. In the tragedies Shakespeare wrote for the Elizabethan theatre, he had the same gift of satisfying deep communal needs while meeting a whole range of individual interests present in his audience.

Didactic drama

When the whole community shares a common heritage, patriotic drama and drama commemorating national heroes, as are seen almost universally in the Orient, is of this kind. Modern Western attempts at a religious didactic drama, or indeed at any drama of "ideas," have had to reckon with the disparate nature of the audience. Thus the impact of Ibsen's social drama both encouraged and divided the development of the theatre in the last years of the 19th century. Plays like *A Doll's House* (1879) and *Ghosts* (published 1881), which challenged the sanctity of marriage and questioned the loyalty a wife owed to her husband, took their audiences by storm: some violently rejected the criticism of their cherished social beliefs, and thus such plays may be said to have failed to persuade general audiences to examine their moral position; on the other hand, there were sufficient numbers of enthusiasts (so-called Ibsenites) to stimulate a new drama of ideas. "Problem" plays appeared all over Europe and undoubtedly rejuvenated the theatre for the 20th century. Shaw's early Ibsenite plays in London, attacking a negative drawing-room comedy with themes of slum landlordism (*Widowers' Houses,* 1892) and prostitution (*Mrs. Warren's Profession,* 1902) resulted only in failure, but Shaw quickly found a comic style that was more disarming. In his attack on false patriotism (*Arms and the Man,* 1894) and the motives for middle class marriage (*Candida,* 1897), he does not affront his audiences before leading them by gentle laughter and surprise to review their own positions.

INFLUENCES ON THE DRAMATIST

The author of a play is affected, consciously or unconsciously, by the conditions under which he conceives and writes, by his social and economic status as a playwright, by his personal background, by his religious or political position, by his purpose in writing. The literary form of the play and its stylistic elements will be influenced by tradition, a received body of theory and dramatic criticism, as well as by the author's innovative energy. Auxiliary theatre arts such as music and design also have their own controlling traditions and conventions, which the playwright must respect. The size and shape of the playhouse, the nature of its stage and equipment, and the actor–audience relationship it encourages also determine the character of the writing. Not least, the audience's cultural assumptions, holy or profane, local or international, social or political, may override all else in deciding the form and content of the drama. These are large considerations that can take the student of drama into areas of sociology, politics, social history, religion, literary criticism, philosophy and aesthetics, and beyond.

The role of theory. It is difficult to assess the influence of theory since theory usually is based on existing drama, rather than drama on theory. Philosophers, critics, and dramatists have attempted both to describe what happens and to prescribe what should happen in drama, but all their theories are affected by what they have seen and read.

Western theory. In Europe, the earliest extant work of dramatic theory, the fragmentary *Poetics* of Aristotle (384–322 BC), chiefly reflecting his views on Greek tragedy and his favorite dramatist, Sophocles, is still relevant to an understanding of the elements of drama. Aristotle's elliptical way of writing, however, encouraged different ages to place their own interpretation upon his statements and to take as prescriptive what many believe to have been meant only to be descriptive. There has been endless discussion of his concepts *mimēsis* ("imitation"), the impulse behind all the arts, and *katharsis* ("purgation," "purification of emotion"), the proper end of tragedy, though these notions were conceived, in part, in answer to Plato's attack on *poiēsis* (making) as an appeal to the irrational. That "character" is second in importance to "plot" is another of Aristotle's concepts that may be understood with reference to the practice of the Greeks, but not more realistic drama, in which character psychology has a dominant importance. The concept in the *Poetics* that has most affected the composition of plays in later ages has been that of the so-called unities—that is, of time, place, and action. Aristotle was evidently describing what he observed—that a typical Greek tragedy had a single plot and action that lasts one day; he made no mention at all of unity of place. Neoclassical critics of the 17th century, however, codified these discussions into rules.

"Rules" of drama

Considering the inconvenience of such rules and their final unimportance, one wonders at the extent of their influence. The Renaissance desire to follow the ancients and its enthusiasm for decorum and classification may explain it in part. Happily, the other classical work recognized at this time was Horace's *Art of Poetry* (*c.* 24 BC), with its

basic precept that poetry should offer pleasure and profit and teach by pleasing, a notion that has general validity to this day. Happily, too, the popular drama, which followed the tastes of its patrons, also exerted a liberating influence. Nevertheless, discussion about the supposed need for the unities continued throughout the 17th century (culminating in the French critic Nicolas Boileau's *Art of Poetry,* originally published in 1674), particularly in France, where a master like Racine could translate the rules into a taut, intense theatrical experience. Only in Spain, where Lope de Vega published his *New Art of Writing Plays* (1609), written out of his experience with popular audiences, was a commonsense voice raised against the classical rules, particularly on behalf of the importance of comedy and its natural mixture with tragedy. In England both Sir Philip Sidney in his *Apologie for Poetry* (1595) and Ben Jonson in *Timber* (1640) merely attacked contemporary stage practice. Jonson, in certain prefaces, however, also developed a tested theory of comic characterization (the "humours") that was to affect English comedy for a hundred years. The best of Neoclassical criticism in English is John Dryden's *Of Dramatick Poesie, an Essay* (1668). Dryden approached the rules with a refreshing honesty and argued all sides of the question; thus he questioned the function of the unities and accepted Shakespeare's practice of mixing comedy and tragedy.

The lively imitation of nature came to be acknowledged as the primary business of the playwright and was confirmed by the authoritative voices of Dr. Samuel Johnson, who said in his *Preface to Shakespeare* (1765) "there is always an appeal open from criticism to nature," and the German dramatist and critic Gotthold Ephraim Lessing, who in his *Hamburgische Dramaturgie* (or *Hamburg Dramaturgy;* 1767–69) sought to accommodate Shakespeare to a new view of Aristotle. With the classical straitjacket removed, there was a release of dramatic energies in new directions. There were still local critical skirmishes, such as Jeremy Collier's attack on the "immorality and profaneness of the English stage" in 1698; Goldoni's attacks upon the already dying Italian commedia on behalf of greater realism; and Voltaire's reactionary wish to return to the unities and to rhymed verse in French tragedy, which was challenged in turn by Diderot's call for a return to nature. But the way was open for the development of the middle class *drame* and the excursions of romanticism. Victor Hugo, in his Preface to his play *Cromwell* (1827), capitalized on the new psychological romanticism of Goethe and Schiller as well as the popularity of the sentimental *drame* in France and the growing admiration for Shakespeare; Hugo advocated truth to nature and a dramatic diversity that could yoke together the sublime and the grotesque. This view of what drama should be received support from Émile Zola in the preface to his play *Thérèse Raquin* (1873), in which he argued a theory of naturalism that called for the accurate observation of people controlled by their heredity and environment. From such sources came the subsequent intellectual approach of Ibsen and Chekhov and a new freedom for such seminal innovators of the 20th century as Luigi Pirandello, with his teasing mixtures of absurdist laughter and psychological shock; Bertolt Brecht (1898–1956), deliberately breaking the illusion of the stage; and Antonin Artaud (1896–1948), advocating a theatre that should be "cruel" to its audience, employing all and any devices that lie to hand. The modern dramatist may be grateful that he is no longer hidebound by theory and yet also regret, paradoxically, that the theatre of his time lacks those artificial limits within which an artifact of more certain efficiency can be wrought.

Twentieth-century views

Eastern theory. The Oriental theatre has always had such limits, but with neither the body of theory nor the pattern of rebellion and reaction found in the West. The Sanskrit drama of India, however, throughout its recorded existence has had the supreme authority of the *Nāṭya-śāstra,* ascribed to Bharata (*c.* 1st century AD), an exhaustive compendium of rules for all the performing arts, but particularly for the sacred art of drama with its auxiliary arts of dance and music. Not only does the *Nāṭya-śāstra* identify many varieties of gesture and movement but it also describes the multiple patterns that drama can assume, similar to a modern treatise on musical form. Every conceivable aspect of a play is treated, from the choice of metre in poetry to the range of moods a play can achieve; but perhaps its primary importance lies in its justification of the aesthetic of Indian drama as a vehicle of religious enlightenment.

In Japan, the most celebrated of early Nō writers, Zeami Motokiyo (1363–1443), left an influential collection of essays and notes to his son about his practice, and his deep knowledge of Zen Buddhism infused the Nō drama with ideals for the art that have persisted. Religious serenity of mind (*yūgen*), conveyed through an exquisite elegance in a performance of high seriousness, is at the heart of Zeami's theory of dramatic art. Three centuries later, the outstanding dramatist Chikamatsu (1653–1725) built equally substantial foundations for the Japanese puppet theatre, later known as the *bunraku.* His heroic plays for this theatre established an unassailable dramatic tradition of depicting an idealized life inspired by a rigid code of honour and expressed with extravagant ceremony and fervent lyricism. At the same time, in another vein, his pathetic "domestic" plays of middle class life and the suicides of lovers established a comparatively realistic mode for Japanese drama, which strikingly extended the range of both the *bunraku* and the Kabuki. Today, these forms, together with the more aristocratic and intellectual Nō, constitute a classical theatre based on practice rather than on theory. They may be superseded as a result of the recent invasion of Western drama, but in their perfection they are unlikely to change. The Yüan drama of China was similarly based upon a slowly evolved body of laws and conventions derived from practice, for, like the Kabuki of Japan, this too was essentially an actors' theatre, and practice rather than theory accounts for its development.

The role of music and dance. The Sanskrit treatise *Nāṭya-śāstra* suggests that drama had its origin in the art of dance, and any survey of Western theatre, too, must recognize a comparable debt to music in the classical Greek drama, which is believed to have sprung from celebratory singing to Dionysus. Similarly, the drama of the medieval church began with the chanted liturgies of the Roman mass. In the professional playhouses of the Renaissance and after, only rarely is music absent: Shakespeare's plays, particularly the comedies, are rich with song, and the skill with which he pursues dramatic ends with musical help is a study in itself. Molière conceived most of his plays as comedy-ballets, and much of his verbal style derives directly from the balletic qualities of the commedia. The popularity of opera in the 18th century led variously to John Gay's prototype for satirical ballad-opera, *The Beggar's Opera* (1728), the opera buffa in Italy, and the opéra comique in France. The development of these forms, however, resulted in the belittling of the written drama, with the notable exception of the parodistic wit of W.S. Gilbert (1836–1911). It is worth noting, however, that the most successful modern "musicals" lean heavily on their literary sources. Today, two of the strongest influences on contemporary theatre are those of Bertolt Brecht, who believed that a dialectical theatre should employ music not merely as a background embellishment but as an equal voice with the actor's, and of Antonin Artaud, who argued that the theatre experience should subordinate the literary text to mime, music, and spectacle. Since it is evident that drama often involves a balance of the arts, an understanding of their interrelationships is proper to a study of dramatic literature.

The influence of theatre design. Though apparently an elementary matter, the shape of the stage and auditorium probably offers the greatest single control over the text of the play that can be measured and tested. Moreover, it is arguable that the playhouse architecture dictates more than any other single factor the style of a play, the conventions of its acting, and the quality of dramatic effect felt by its audience. The shape of the theatre is always changing, so that to investigate its function is both to understand the past and to anticipate the future. Today, Western theatre is in the process of breaking away from the dominance of the Victorian picture-frame theatre, and therefore from the kind of experience this produced.

The contemporary English critic John Wain has called the difference between Victorian and Elizabethan theatre a difference between "consumer" and "participation" art. The difference resulted from the physical relationship between the audience and the actor in the two periods, a relationship that determined the kind of communication open to the playwright and the role the drama could play in society. Three basic playhouse shapes have emerged in the history of the theatre: the arena stage, the open stage, and the picture-frame.

Three basic playhouse types

The arena stage. To the arena, or theatre-in-the-round, belongs the excitement of the circus, the bullring, and such sports as boxing and wrestling. Arena performance was the basis for all early forms of theatre—the Druid ceremonies at Stonehenge, the Tibetan harvest-festival drama, probably early Greek ritual dancing in the *orchēstra,* the medieval rounds in 14th-century England and France, the medieval street plays on pageant wagons, the early Nō drama of Japan, the royal theatre of Cambodia. Characteristic of all these theatres is the bringing together of whole communities for a ritual experience; therefore, a sense of ritualistic intimacy and involvement is common to the content of the drama, and only the size of the audience changes the scale of the sung or spoken poetry. Clearly, the idiom of realistic dialogue would have been inappropriate both to the occasion and the manner of such theatre.

The open stage. When more narrative forms of action appeared in drama and particular singers or speakers needed to control the attention of their audience by facing them, the open, "thrust," or platform stage, with the audience on three sides of the actor, quickly developed its versatility. Intimate and ritualistic qualities in the drama could be combined with a new focus on the players as individual characters. The open stage and its variants were used by the majority of great national theatres, particularly those of China and Japan, the booths of the Italian commedia, the Elizabethan public and private playhouses, and the Spanish *corrales* (*i.e.,* the areas between town houses) of the Renaissance. While open-stage performance discouraged scenic elaboration, it stressed the actor and his role, his playing to and away from the spectators, with the consequent subtleties of empathy and alienation. It permitted high style in speech and behaviour, yet it could also accommodate moments of the colloquial and the realistic. It encouraged a drama of range and versatility, with rapid changes of mood and great flexibility of tone. It is not surprising that in the 20th century the West has seen a return to the open stage and that recent plays of Brechtian theatre and the theatre of the absurd seem composed for open staging.

The proscenium stage. The third basic theatre form is that of the proscenium-arch or picture-frame stage, which reached its highest achievements in the late 19th century. Not until public theatres were roofed, the actors withdrawn into the scene, and the stage artifically illuminated were conditions ripe in Western theatre for a new development of spectacle and illusion. This development had a revolutionary effect upon the literary drama. In the 18th and 19th centuries, plays were shaped into a new structure of acts and scenes, with intermissions to permit scene changes. Only recently has the development of lighting techniques encouraged a return to a more flexible episodic drama. Of more importance, the actor increasingly withdrew into the created illusion of the play, and his character became part of it. In the mid-19th century, when it was possible to dim the house lights, the illusion could be made virtually complete. At its best, stage illusion could produce the delicate naturalism of a Chekhovian family scene, into which the spectator was drawn by understanding, sympathy, and recognition; at its worst, the magic of spectacle and the necessary projection of the speech and acting in the largest picture-frame theatres produced a crude drama of sensation in which literary values had no place.

Audience expectations. It may be that the primary influence upon the conception and creation of a play is that of the audience. An audience allows a play to have only the emotion and meaning it chooses, or else it defends itself either by protest or by a closed mind. From the time the spectator began paying for his playgoing, during the Renaissance, the audience more and more entered into the choice of the drama's subjects and their treatment. This is not to say that the audience was given no consideration earlier; even in medieval plays there were popular non-biblical roles such as Noah's wife, or Mak the sheepthief among the three shepherds, and the antic devils of the Harrowing of Hell in the English mystery cycles. Nor, in later times, did a good playwright always give the audience only what it expected—Shakespeare's *King Lear* (*c.* 1605), for example, in the view of many the world's greatest play, had its popular elements of folktale, intrigue, disguise, madness, clowning, blood, and horror; but each was turned by the playwright to the advantage of his theme.

Any examination of the society an audience represents must illuminate not only the cultural role of its theatre but also the content, genre, and style of its plays. The exceptionally aristocratic composition of the English Restoration audience, for example, illuminates the social game its comedy represented, and the middle class composition of the subsequent Georgian audience sheds light on the moralistic elements of its "sentimental" comedy. Not unrelated is the study of received ideas in the theatre. The widespread knowledge of simple Freudian psychology has undoubtedly granted a contemporary playwright like Tennessee Williams (1911–83) the license to invoke it for character motivation; and Brecht increasingly informed his comedies with Marxist thinking on the assumption that the audiences he wrote for would appreciate his dramatized argument. Things go wrong when the intellectual or religious background of the audience does not permit a shared experience, as when Jean-Paul Sartre (1905–80) could not persuade a predominantly Christian audience with an existentialist explanation for the action of his plays, or when T.S. Eliot (1888–1965) failed to persuade an audience accustomed to the conventions of drawing-room comedy that *The Cocktail Party* (1949) was a possible setting for Christian martyrdom. Good drama persuades before it preaches, but it can only begin where the audience begins.

Special audiences. A great variety of drama has been written for special audiences. Plays have been written for children, largely in the 20th century, though Nativity plays have always been associated with children both as performers and as spectators. These plays tend to be fanciful in conception, broad in characterization, and moralistic in intention. Nevertheless, the most famous of children's plays, James Barrie's *Peter Pan* (1904), implied that the young are no fools and celebrated children in their own right. Barrie submerged his point subtly beneath the fantasy, and his play is still regularly performed, while Maurice Maeterlinck's *Blue Bird* (1908) has disappeared from the repertory because of its weighty moral tone.

Children's drama

In the wider field of adult drama, the social class of the audience often accounts for a play's form and style. Court or aristocratic drama is readily distinguished from that of the popular theatre. The veneration in which the Nō drama was held in Japan derived in large part from the feudal ceremony of its presentation, and its courtly elements ensured its survival for an upper class and intellectual elite. Although much of it derived from the Nō, the flourishing of the Kabuki at the end of the 17th century is related to the rise of a new merchant and middle class audience, which encouraged the development of less esoteric drama. The popular plays of the Elizabethan public theatres, with their broader, more romantic subjects liberally spiced with comedy, are similarly to be contrasted with those of the private theatres. The boys' companies of the private theatres of Elizabethan London played for a better paying and more sophisticated audience, which favoured the satirical or philosophical plays of Thomas Middleton (1570?–1627), John Marston (1576–1634), and George Chapman (1559?–1634). Similarly today, in all Western dramatic media—stage, film, radio, and television—popular and "commercial" forms run alongside more "cultural" and avant-garde forms, so that the drama, which in its origins brought people together, now divides them. Whether the esoteric influences the popular theatre, or vice versa, is not clear, and research remains to be done on whether

this dichotomy is good or bad for dramatic literature or the people it is written for.

THE RANGE OF DRAMATIC FORMS AND STYLES

Dramatic literature has a remarkable facility in bringing together elements from other performing and nonperforming arts: design and mime, dance and music, poetry and narrative. It may be that the dramatic impulse itself, the desire to recreate a picture of life for others through impersonation, is at the root of all the arts. Certainly, the performing arts continually have need of dramatic literature to support them. A common way of describing an opera, for example, is to say that it is a play set to music. In Wagner the music is continuous; in Verdi the music is broken into songs; in Mozart the songs are separated by recitative, a mixture of speech and song; while operettas and musical comedy consist of speech that breaks into song from time to time. All forms of opera, however, essentially dramatize a plot, even if the plot must be simplified on the operatic stage. This is because, in opera, musical conventions dominate the dramatic conventions, and the spectator who finds that the music spoils the play, or who finds that the play spoils the music, is one who has not accepted the special conventions of opera. Music is drama's natural sister; proof may be seen in the early religious music-drama of the Dionysiac festivals of Greece and the *mystères* of 14th-century France, as well as in the remarkable development of opera in 17th-century Italy spreading to the rest of the world. The librettist who writes the text of an opera, however, must usually subserve the composer, unless he is able to embellish his play with popular lyrics, as John Gay did in *The Beggar's Opera* (1728), or to work in exceptionally close collaboration with the composer, as Brecht did with Kurt Weill for his *Die Dreigroschenoper* (*The Threepenny Opera,* 1928).

Opera

Dance, with its modern, sophisticated forms of ballet, has also been traditionally associated with dramatic representation and has similarly changed its purpose from religious to secular. In ballet, the music is usually central, and the performance is conceived visually and aurally; hence, the writer does not play a dominant role. The scenario is prepared for dance and mime by the choreographer. The contemporary Irish writer Samuel Beckett, trying to reduce his dramatic statement to the barest essentials, "composed" two mimes entitled *Act Without Words I* and *II* (1957 and 1966), but this is exceptional.

In motion pictures, the script writer has a more important but still not dominant role. He usually provides a loose outline of dialogue, business, and camera work on which the director, his cameramen, and the cutting editor build the finished product. The director is usually the final artistic authority and the central creative mind in the process, and words are usually subordinate to the dynamic visual imagery. (This subject is developed at length in the article MOTION PICTURES.)

The media of radio and television both depend upon words in their drama to an extent that is not characteristic of the motion picture. Though these mass media have been dominated by commercial interests and other economic factors, they also have developed dramatic forms from the special nature of their medium. The writer of a radio play must acknowledge that the listener cannot see the actors but hears them in conditions of great intimacy. A radio script that stresses the suggestive, imaginative, or poetic quality of words and permits a more than conventional freedom with time and place can produce a truly poetic drama, perhaps making unobtrusive use of earlier devices like the chorus, the narrator, and the soliloquy: the outstanding example of radio drama is *Under Milk Wood* (1953), by the Welsh poet Dylan Thomas.

A similar kind of dramatic writing is the so-called readers' theatre, in which actors read or recite without decor before an audience. (This is not to be confused with "closet drama," often a dramatic poem that assumes dialogue form; *e.g.,* Milton's *Samson Agonistes,* 1671, written without the intention of stage performance.) The essential discipline of the circuit of communication with an audience is what distinguishes drama as a genre, however many forms it has taken in its long history. (J.L.S.)

Comedy

The classic conception of comedy, which began with Aristotle in ancient Greece of the 4th century BC and persists through the present, holds that it is primarily concerned with man as a social being, rather than as a private person, and that its function is frankly corrective. The comic artist's purpose is to hold a mirror up to society to reflect its follies and vices, in the hope that they will, as a result, be mended. The 20th-century French philosopher Henri Bergson shared this view of the corrective purpose of laughter; specifically, he felt, laughter is intended to bring the comic character back into conformity with his society, whose logic and conventions he abandons when "he slackens in the attention that is due to life." Here comedy is considered primarily as a literary genre, but also is considered for its manifestations in the other arts. The wellsprings of comedy are dealt with in the article HUMOUR AND WIT. The comic impulse in the visual arts is discussed in CARICATURE, CARTOON, AND COMIC STRIP.

The corrective purpose of laughter

ORIGINS AND DEFINITIONS

The word comedy seems to be connected by derivation with the Greek verb meaning "to revel," and comedy arose out of the revels associated with the rites of Dionysus, a god of vegetation. The origins of comedy are thus bound up with vegetation ritual. Aristotle, in his *Poetics,* states that comedy originated in phallic songs and that, like tragedy, it began in improvisation. Though tragedy evolved by stages that can be traced, the progress of comedy passed unnoticed because it was not taken seriously. When tragedy and comedy arose, poets wrote one or the other, according to their natural bent. Those of the graver sort, who might previously have been inclined to celebrate the actions of the great in epic poetry, turned to tragedy; poets of a lower type, who had set forth the doings of the ignoble in invectives, turned to comedy. The distinction is basic to the Aristotelian differentiation between tragedy and comedy: tragedy imitates men who are better than the average, and comedy men who are worse.

For centuries, efforts at defining comedy were to be along the lines set down by Aristotle: the view that tragedy deals with personages of high estate, and comedy deals with lowly types; that tragedy treats of matters of great public import, while comedy is concerned with the private affairs of mundane life; and that the characters and events of tragedy are historic and so, in some sense, true, while the humbler materials of comedy are but feigned. Implicit, too, in Aristotle is the distinction in styles deemed appropriate to the treatment of tragic and comic story. As long as there was at least a theoretical separation of comic and tragic styles, either genre could, on occasion, appropriate the stylistic manner of the other to a striking effect, which was never possible after the crossing of stylistic lines became commonplace. The ancient Roman poet Horace, who wrote on such stylistic differences, noted the special effects that can be achieved when comedy lifts its voice in pseudotragic rant and when tragedy adopts the prosaic but affecting language of comedy. Consciously combined, the mixture of styles produces the burlesque, in which the grand manner (epic or tragic) is applied to a trivial subject, or the serious subject is subjected to a vulgar treatment, to ludicrous effect. The English novelist Henry Fielding, in the preface to *Joseph Andrews* (1742), was careful to distinguish between the comic and the burlesque; the latter centres on the monstrous and unnatural and gives pleasure through the surprising absurdity it exhibits in appropriating the manners of the highest to the lowest, or vice versa. Comedy, on the other hand, confines itself to the imitation of nature, and, according to Fielding, the comic artist is not to be excused for deviating from it. His subject is the ridiculous, not the monstrous, as with the writer of burlesque; and the nature he is to imitate is human nature, as viewed in the ordinary scenes of civilized society.

Distinction between the comic and the burlesque

The human contradiction. In dealing with man as a social being, all great comic artists have known that they are in the presence of a contradiction: that behind the social being lurks an animal being, whose behaviour of-

ten accords very ill with the canons dictated by society. Comedy, from its ritual beginnings, has celebrated creative energy. The primitive revels out of which comedy arose frankly acknowledged man's animal nature; the animal masquerades and the phallic processions are the obvious witnesses to it. Comedy testifies to man's physical vitality, his delight in life, his will to go on living. Comedy is at its merriest, its most festive, when this rhythm of life can be affirmed within the civilized context of human society. In the absence of this sort of harmony between creatural instincts and the dictates of civilization, sundry strains and discontents arise, all baring witness to the contradictory nature of man, which in the comic view is a radical dualism; his efforts to follow the way of rational sobriety are forever being interrupted by the infirmities of the flesh. The duality that tragedy views as a fatal contradiction in the nature of things comedy views as one more instance of the incongruous reality that every man must live with as best he can. "Wherever there is life, there is contradiction," says Søren Kierkegaard, the 19th-century Danish Existentialist, in the *Concluding Unscientific Postscript* (1846), "and wherever there is contradiction, the comical is present." He went on to say that the tragic and the comic are both based on contradiction; but "the tragic is the suffering contradiction, comical, painless contradiction." Comedy makes the contradiction manifest along with a way out, which is why the contradiction is painless. Tragedy, on the other hand, despairs of a way out of the contradiction.

The incongruous is "the essence of the laughable," said the English essayist William Hazlitt, who also declared, in his essay "On Wit and Humour" in *English Comic Writers* (1819), that "Man is the only animal that laughs and weeps; for he is the only animal that is struck with the difference between what things are, and what they ought to be."

Comedy, satire, and romance. Comedy's dualistic view of man as an incongruous mixture of bodily instinct and rational intellect is an essentially ironic view—implying the capacity to see things in a double aspect. The comic drama takes on the features of satire as it fixes on professions of virtue and the practices that contradict them. Satire assumes standards against which professions and practices are judged. To the extent that the professions prove hollow and the practices vicious, the ironic perception darkens and deepens. The element of the incongruous points in the direction of the grotesque, which implies an admixture of elements that do not match. The ironic gaze eventually penetrates to a vision of the grotesque quality of experience, marked by the discontinuity of word and deed and the total lack of coherence between appearance and reality. This suggests one of the extreme limits of comedy, the satiric extreme, in which the sense of the discrepancy between things as they are and things as they might be or ought to be has reached to the borders of tragedy. For the tragic apprehension, as Kierkegaard states, despairs of a way out of the contradictions that life presents.

The ironic view of man

As satire may be said to govern the movement of comedy in one direction, romance governs its movement in the other. Satiric comedy dramatizes the discrepancy between the ideal and the reality and condemns the pretensions that would mask reality's hollowness and viciousness. Romantic comedy also regularly presents the conflict between the ideal shape of things as hero or heroine could wish them to be and the hard realities with which they are confronted, but typically it ends by invoking the ideal, despite whatever difficulties reality has put in its way. This is never managed without a good deal of contrivance, and the plot of the typical romantic comedy is a medley of clever scheming, calculated coincidence, and wondrous discovery, all of which contribute ultimately to making the events answer precisely to the hero's or heroine's wishes. Plotting of this sort has had a long stage tradition and not exclusively in comedy. It is first encountered in the tragicomedies of the ancient Greek dramatist Euripides (*e.g., Alcestis, Iphigeneia in Tauris, Ion, Helen*). Shakespeare explored the full range of dramatic possibilities of the romantic mode of comedy. The means by which the happy ending is accomplished in romantic comedy—the document or the bodily mark that establishes identities to the satisfaction of all the characters of goodwill—are part of the stock-in-trade of all comic dramatists, even such 20th-century playwrights as Jean Anouilh (in *Le Voyageur sans bagage*) and T.S. Eliot (in *The Confidential Clerk*).

There is nothing necessarily inconsistent in the use of a calculatedly artificial dramatic design to convey a serious dramatic statement. The contrived artifice of Shakespeare's mature comic plots is the perfect foil against which the reality of the characters' feelings and attitudes assumes the greater naturalness. The strange coincidences, remarkable discoveries, and wonderful reunions are unimportant compared with the emotions of relief and awe that they inspire. Their function, as Shakespeare uses them, is precisely to give rise to such emotions, and the emotions, thanks to the plangent poetry in which they are expressed, end by transcending the circumstances that occasioned them. But when such artifices are employed simply for the purpose of eliminating the obstacles to a happy ending—as is the case in the sentimental comedy of the 18th and early 19th centuries—then they stand forth as imaginatively impoverished dramatic clichés. The dramatists of sentimental comedy were committed to writing exemplary plays, wherein virtue would be rewarded and vice frustrated. If hero and heroine were to be rescued from the distresses that had encompassed them, any measures were apparently acceptable; the important thing was that the play's action should reach an edifying end. It is but a short step from comedy of this sort to the melodrama that flourished in the 19th-century theatre. The distresses that the hero and heroine suffer are, in melodrama, raised to a more than comic urgency, but the means of deliverance have the familiar comic stamp: the secret at last made known, the long-lost child identified, the hard heart made suddenly capable of pity. Melodrama is a form of fantasy that proceeds according to its own childish and somewhat egoistic logic; hero and heroine are pure, anyone who opposes them is a villain, and the purity that has exposed them to risks must ensure their eventual safety and happiness. What melodrama is to tragedy farce is to comedy, and the element of fantasy is equally prominent in farce and in melodrama. If melodrama provides a fantasy in which the protagonist suffers for his virtues but is eventually rewarded for them, farce provides a fantasy in which the protagonist sets about satisfying his most roguish or wanton, mischievous or destructive, impulses and manages to do so with impunity.

19th-century melodrama

THEORIES

The treatise that Aristotle is presumed to have written on comedy is lost. There is, however, a fragmentary treatise on comedy that bears an obvious relation to Aristotle's treatise on tragedy, *Poetics,* and is generally taken to be either a version of a lost Aristotelian original or an expression of the philosophical tradition to which he belonged. This is the *Tractatus Coislinianus,* preserved in a 10th-century manuscript in the De Coislin Collection in Paris. The *Tractatus* divides the substance of comedy into the same six elements that are discussed in regard to tragedy in the *Poetics:* plot, character, thought, diction, melody, and spectacle. The characters of comedy, according to the *Tractatus,* are of three kinds: the impostors, the self-deprecators, and the buffoons. The Aristotelian tradition from which the *Tractatus* derives probably provided a fourth, the churl, or boor. The list of comic characters in the *Tractatus* is closely related to a passage in Aristotle's *Nicomachean Ethics,* in which the boaster (the person who says more than the truth) is compared with the mock-modest man (the person who says less), and the buffoon (who has too much wit) is contrasted with the boor (who has too little).

Comedy as a rite. The *Tractatus* was not printed until 1839, and its influence on comic theory is thus of relatively modern date. It is frequently cited in the studies that attempt to combine literary criticism and anthropology, in the manner in which Sir James George Frazer combined studies of primitive religion and culture in *The Golden Bough* (1890–1915). In such works, comedy and tragedy alike are traced to a prehistoric death-and-resurrection

ceremonial, a seasonal pantomime in which the old year, in the guise of an aged king (or hero or god), is killed, and the new spirit of fertility, the resurrection or initiation of the young king, is brought in. This rite typically featured a ritual combat, or agon, between the representatives of the old and the new seasons, a feast in which the sacrificial body of the slain king was devoured, a marriage between the victorious new king and his chosen bride, and a final triumphal procession in celebration of the reincarnation or resurrection of the slain god. Implicit in the whole ceremony is the ancient rite of purging the tribe through the expulsion of a scapegoat, who carries away the accumulated sins of the past year. Frazer, speaking of scapegoats in *The Golden Bough,* noted that this expulsion of devils was commonly preceded or followed by a period of general license, an abandonment of the ordinary restraints of society during which all offenses except the gravest go unpunished. This quality of Saturnalia is characteristic of comedy from ancient Greece through medieval Europe.

The role of the scapegoats

The seasonal rites that celebrate the yearly cycle of birth, death, and rebirth are seen by the contemporary Canadian critic Northrop Frye as the basis for the generic plots of comedy, romance, tragedy, and irony and satire. The four prefigure the fate of a hero and the society he brings into being. In comedy (representing the season of spring), the hero appears in a society controlled by obstructing characters and succeeds in wresting it from their grasp. The movement of comedy of this sort typically replaces falsehood with truth, illusion with reality. The hero, having come into possession of his new society, sets forth upon adventures, and these are the province of romance (summer). Tragedy (autumn) commemorates the hero's passion and death. Irony and satire (winter) depict a world from which the hero has disappeared, a vision of "unidealized existence." With spring, the hero is born anew.

The moral force of comedy. The characters of comedy specified in the *Tractatus* arrange themselves in a familiar pattern: a clever hero is surrounded by fools of sundry varieties (impostors, buffoons, boors). The hero is something of a trickster; he dissimulates his own powers, while exploiting the weaknesses of those around him. The comic pattern is a persistent one; it appears not only in ancient Greek comedy but also in the farces of ancient Italy, in the commedia dell'arte that came into being in 16th-century Italy, and even in the routines involving a comedian and his straight man in the nightclub acts and the television variety shows of the present time. Implicit here is the tendency to make folly ridiculous, to laugh it out of countenance, which has always been a prominent feature of comedy.

Renaissance critics, elaborating on the brief and cryptic account of comedy in Aristotle's *Poetics,* stressed the derisive force of comedy as an adjunct to morality. The Italian scholar Gian Giorgio Trissino's account of comedy in his *Poetica,* apparently written in the 1530s, is typical: as tragedy teaches by means of pity and fear, comedy teaches by deriding things that are vile. Attention is directed here, as in other critical treatises of this kind, to the source of laughter. According to Trissino, laughter is aroused by objects that are in some way ugly and especially by that from which better qualities were hoped. His statement suggests the relation of the comic to the incongruous. Trissino was as aware as the French poet Charles Baudelaire was three centuries later that laughter betokens the fallen nature of man (Baudelaire would term it man's Satanic nature). Man laughs, says Trissino (echoing Plato's dialogue *Philebus*), because he is envious and malicious and never delights in the good of others except when he hopes for some good from it for himself.

The most important English Renaissance statement concerning comedy is that of Sir Philip Sidney in *The Defence of Poesie* (1595):

Sir Philip Sidney's definition of comedy

> comedy is an imitation of the common errors of our life, which [the comic dramatist] representeth in the most ridiculous and scornful sort that may be, so as it is impossible that any beholder can be content to be such a one.

Like Trissino, Sidney notes that, while laughter comes from delight, not all objects of delight cause laughter, and he demonstrates the distinction as Trissino had done: "we are ravished with delight to see a fair woman, and yet are far from being moved to laughter. We laugh at deformed creatures, wherein certainly we cannot delight." The element of the incongruous is prominent in Sidney's account of scornful laughter. He cites the image of the hero of Greek legend Heracles, with his great beard and furious countenance, in woman's attire, spinning at the command of his beloved queen, Omphale, and declares that this arouses both delight and laughter.

Comedy and character. Another English poet, John Dryden, in *Of Dramatick Poesie, an Essay* (1668), makes the same point in describing the kind of laughter produced by the ancient Greek comedy *The Clouds,* by Aristophanes. In it, the character of Socrates is made ridiculous by acting very unlike the true Socrates; that is, by appearing childish and absurd rather than with the gravity of the true Socrates. Dryden was concerned with analyzing the laughable quality of comedy and with demonstrating the different forms it has taken in different periods of dramatic history. Aristophanic comedy sought its laughable quality not so much in the imitation of a man as in the representation of "some odd conceit which had commonly somewhat of unnatural or obscene in it." In the so-called New Comedy, introduced by Menander late in the 4th century BC, writers sought to express the ethos, or character, as in their tragedies they expressed the pathos, or suffering, of mankind. This distinction goes back to Aristotle, who, in the *Rhetoric,* distinguished between ethos, a man's natural bent, disposition, or moral character, and pathos, emotion displayed in a given situation. And the Latin rhetorician Quintilian, in the 1st century AD, noted that ethos is akin to comedy and pathos to tragedy. The distinction is important to Renaissance and Neoclassical assumptions concerning the respective subject of comic and tragic representation. In terms of emotion, ethos is viewed as a permanent condition characteristic of the average man and relatively mild in its nature; pathos, on the other hand, is a temporary emotional state, often violent. Comedy thus expresses the characters of men in the ordinary circumstances of everyday life; tragedy expresses the sufferings of a particular man in extraordinary periods of intense emotion.

In dealing with men engaged in normal affairs, the comic dramatists tended to depict the individual in terms of some single but overriding personal trait or habit. They adopted a method based on the physiological concept of the four humours, or bodily fluids (blood, phlegm, choler, melancholy), and the belief that an equal proportion of these constituted health, while an excess or deficiency of any one of them brought disease. Since the humours governed temperament, an irregular distribution of them was considered to result not only in bodily sickness but also in derangements of personality and behaviour, as well. The resultant comedy of humours is distinctly English, as Dryden notes, and particularly identified with the comedies of Ben Jonson.

English comedy of humours

The role of wit. Humour is native to man. Folly need only be observed and imitated by the comic dramatist to give rise to laughter. Observers as early as Quintilian, however, have pointed out that, though folly is laughable in itself, such jests may be improved if the writer adds something of his own; namely, wit. A form of repartee, wit implies both a mental agility and a linguistic grace that is very much a product of conscious art. Quintilian describes wit at some length in his *Institutio oratoria;* it partakes of urbanity, a certain tincture of learning, charm, saltiness, or sharpness, and polish and elegance. In the preface (1671) to *An Evening's Love,* Dryden distinguishes between the comic talents of Ben Jonson, on the one hand, and of Shakespeare and his contemporary John Fletcher, on the other, by virtue of their excelling, respectively, in humour and wit. Jonson's talent lay in his ability "to make men appear pleasantly ridiculous on the stage"; while Shakespeare and Fletcher excelled in wit, or "the sharpness of conceit," as seen in their repartee. The distinction is noted as well in *Of Dramatick Poesie, an Essay,* where a comparison is made between the character of Morose in Jonson's play *Epicœne,* who is characterized by his

humour (namely, his inability to abide any noise but the sound of his own voice), and Shakespeare's Falstaff, who, according to Dryden, represents a miscellany of humours and is singular in saying things that are unexpected by the audience.

The distinctions that Hazlitt arrives at, then, in his essay "On Wit and Humour" are very much in the classic tradition of comic criticism:

> Humour is the describing the ludicrous as it is in itself; wit is the exposing it, by comparing or contrasting it with something else. Humour is, as it were, the growth of nature and accident; wit is the product of art and fancy.

The distinctions persist into the most sophisticated treatments of the subject. Sigmund Freud, for example, in *Wit and its Relation to the Unconscious* (1905), said that wit is made, but humour is found. Laughter, according to Freud, is aroused at actions that appear immoderate and inappropriate, at excessive expenditures of energy: it expresses a pleasurable sense of the superiority felt on such occasions.

Baudelaire on the grotesque. The view that laughter comes from superiority is referred to as a commonplace by Baudelaire, who states it in his essay "On the Essence of Laughter" (1855). Laughter, says Baudelaire, is a consequence of man's notion of his own superiority. It is a token both of an infinite misery, in relation to the absolute being of whom man has an inkling, and of infinite grandeur, in relation to the beasts, and results from the perpetual collision of these two infinities. The crucial part of Baudelaire's essay, however, turns on his distinction between the comic and the grotesque. The comic, he says, is an imitation mixed with a certain creative faculty; the grotesque is a creation mixed with a certain imitative faculty—imitative of elements found in nature. Each gives rise to laughter expressive of an idea of superiority—in the comic, the superiority of man over man, and, in the grotesque, the superiority of man over nature. The laughter caused by the grotesque has about it something more profound and primitive, something much closer to the innocent life, than has the laughter caused by the comic in man's behaviour. In France, the great master of the grotesque was the 16th-century author François Rabelais, while some of the plays of Molière, in the next century, best expressed the comic.

Baudelaire's definition of laughter

Bergson's and Meredith's theories. The French philosopher Henri Bergson (1859–1941) analyzed the dialectic of comedy in his essay "Laughter," which deals directly with the spirit of contradiction that is basic both to comedy and to life. Bergson's central concern is with the opposition of the mechanical and the living; stated in its most general terms, his thesis holds that the comic consists of something mechanical encrusted on the living. Bergson traces the implications of this view in the sundry elements of comedy: situations, language, characters. Comedy expresses a lack of adaptability to society; any individual is comic who goes his own way without troubling to get into touch with his fellow beings. The purpose of laughter is to wake him from his dream. Three conditions are essential for the comic: the character must be unsociable, for that is enough to make him ludicrous; the spectator must be insensible to the character's condition, for laughter is incompatible with emotion; and the character must act automatically (Bergson cites the systematic absentmindedness of Don Quixote). The essential difference between comedy and tragedy, says Bergson, invoking a distinction that goes back to that maintained between ethos and pathos, is that tragedy is concerned with individuals and comedy with classes. And the reason that comedy deals with the general is bound up with the corrective aim of laughter: the correction must reach as great a number of persons as possible. To this end, comedy focusses on peculiarities that are not indissolubly bound up with the individuality of a single person.

It is the business of laughter to repress any tendency on the part of the individual to separate himself from society. The comic character would, if left to his own devices, break away from logic (and thus relieve himself from the strain of thinking); give over the effort to adapt and readapt himself to society (and thus slacken in the attention that is due to life); and abandon social convention (and thus relieve himself from the strain of living).

The essay "On the Idea of Comedy and the Uses of the Comic Spirit" (1877), by Bergson's English contemporary George Meredith, is a celebration of the civilizing power of the comic spirit. The mind, he affirms, directs the laughter of comedy, and civilization is founded in common sense, which equips one to hear the comic spirit when it laughs folly out of countenance and to participate in its fellowship.

Both Bergson's and Meredith's essays have been criticized for focussing so exclusively on comedy as a socially corrective force and for limiting the scope of laughter to its derisive power. The charge is more damaging to Meredith's essay than it is to Bergson's. Whatever the limitations of the latter, it nonetheless explores the implications of its own thesis with the utmost thoroughness, and the result is a rigorous analysis of comic causes and effects for which any student of the subject must be grateful. It is with farce that Bergson's remarks on comedy have the greatest connection and on which they seem chiefly to have been founded. It is no accident that most of his examples are drawn from Molière, in whose work the farcical element is strong, and from the farces of Bergson's own contemporary Eugène Labiche. The laughter of comedy is not always derisive, however, as some of Shakespeare's greatest comedies prove; and there are plays, such as Shakespeare's last ones, which are well within an established tradition of comedy but in which laughter hardly sounds at all. These suggest regions of comedy on which Bergson's analysis of the genre sheds hardly any light at all.

The laughter of comedy

The comic as a failure of self-knowledge. Aristotle said that comedy deals with the ridiculous, and Plato, in the *Philebus,* defined the ridiculous as a failure of self-knowledge; such a failure is there shown to be laughable in private individuals (the personages of comedy) but terrible in persons who wield power (the personages of tragedy). In comedy, the failure is often mirrored in a character's efforts to live up to an ideal of self that may be perfectly worthy but the wrong ideal for him. Shakespearean comedy is rich in examples: the King of Navarre and his courtiers, who must be made to realize that nature meant them to be lovers, not academicians, in *Love's Labour's Lost;* Beatrice and Benedick, who must be made to know that nature meant them for each other, not for the single life, in *Much Ado About Nothing;* the Duke Orsino in *Twelfth Night,* who is brought to see that it is not Lady Olivia whom he loves but the disguised Viola, and Lady Olivia herself, who, when the right man comes along, decides that she will not dedicate herself to seven years of mourning for a dead brother, after all; and Angelo in *Measure for Measure,* whose image of himself collapses when his lust for Isabella makes it clear that he is not the ascetic type. The movement of all these plays follows a familiar comic pattern, wherein characters are brought from a condition of affected folly amounting to self-delusion to a plain recognition of who they are and what they want. For the five years or so after he wrote *Measure for Measure,* in 1604, Shakespeare seems to have addressed himself exclusively to tragedy, and each play in the sequence of masterpieces he produced during this period—*Othello, King Lear, Macbeth, Antony and Cleopatra,* and *Coriolanus*—turns in some measure on a failure of self-knowledge. This is notably so in the case of *Lear,* which is the tragedy of a man who (in the words of one of his daughters) "hath ever but slenderly known himself," and whose fault (as the Fool suggests) is to have grown old before he grew wise.

The plots of Shakespeare's last plays (*Pericles, Cymbeline, The Winter's Tale, The Tempest*) all contain a potential tragedy but one that is resolved by nontragic means. They contain, as well, an element of romance of the kind purveyed from Greek New Comedy through the plays of the ancient Roman comic dramatists Plautus and Terence. Children lost at birth are miraculously restored, years later, to their parents, thereby providing occasion for a recognition scene that functions as the denouement of the plot. Characters find themselves—they come to know themselves—in all manner of ways by the ends of these

plays. Tragic errors have been made, tragic losses have been suffered, tragic passions—envy, jealousy, wrath—have seemed to rage unchecked, but the miracle that these plays celebrate lies in the discovery that the errors can be forgiven, the losses restored, and the passions mastered by man's godly spirit of reason. The near tragedies experienced by the characters result in the ultimate health and enlightenment of the soul. What is learned is of a profound simplicity: the need for patience under adversity, the need to repent of one's sins, the need to forgive the sins of others. In comedy of this high and sublime sort, patience, repentance, and forgiveness are opposed to the viciously circular pattern of crime, which begets vengeance, which begets more crime. Comedy of this sort deals in regeneration and rebirth. There is always about it something of the religious, as humankind is absolved of its guilt and reconciled one to another and to whatever powers that be.

Distinction between comedy and tragedy

Divine comedies in the West and East. The 4th-century Latin grammarian Donatus distinguished comedy from tragedy by the simplest terms: comedies begin in trouble and end in peace, while tragedies begin in calms and end in tempest. Such a differentiation of the two genres may be simplistic, but it provided sufficient grounds for Dante to call his great poem *La Commedia* (*The Comedy;* later called *The Divine Comedy*), since, as he says in his dedicatory letter, it begins amid the horrors of hell but ends amid the pleasures of heaven. This suggests the movement of Shakespeare's last plays, which begin amid the distresses of the world and end in a supernal peace. Comedy conceived in this sublime and serene mode is rare but recurrent in the history of the theatre. The Spanish dramatist Calderón's *Vida es sueño* (1635; "Life Is a Dream") is an example; so, on the operatic stage, is Mozart's *Magic Flute* (1791), in spirit and form so like Shakespeare's *Tempest,* to which it has often been compared. In later drama, Henrik Ibsen's *Little Eyolf* (1894) and August Strindberg's *To Damascus* (1898–1904)—both of which are among the late works of these Scandinavian dramatists—have affinities with this type, and this is the comic mode in which T.S. Eliot's last play, *The Elder Statesman* (1958), is conceived. It may represent the most universal mode of comedy. The American philosopher Susanne K. Langer writes:

> In Asia the designation "Divine Comedy" would fit numberless plays; especially in India triumphant gods, divine lovers united after various trials [as in the perennially popular romance of Rama and Sita], are the favourite themes of a theater that knows no "tragic rhythm." The classical Sanskrit drama was heroic comedy—high poetry, noble action, themes almost always taken from the myths—a serious, religiously conceived drama, yet in the "comic" pattern, which is not a complete organic development reaching a foregone, inevitable conclusion, but is episodic, restoring a lost balance, and implying a new future. The reason for this consistently "comic" image of life in India is obvious enough: both Hindu and Buddhist regard life as an episode in the much longer career of the soul which has to accomplish many incarnations before it reaches its goal, nirvana. Its struggles in the world do not exhaust it; in fact they are scarcely worth recording except in entertainment theater, "comedy" in our sense—satire, farce, and dialogue. The characters whose fortunes are seriously interesting are the eternal gods; and for them there is no death, no limit of potentialities, hence no fate to be fulfilled. There is only the balanced rhythm of sentience and emotion, upholding itself amid the changes of material nature. (From *Feeling and Form;* Charles Scribner's Sons, 1953.)

KINDS OF COMEDY IN DIVERSE HISTORICAL PERIODS

Old and New Comedy in ancient Greece. The 11 surviving plays of Aristophanes represent the earliest extant body of comic drama; what is known of Greek Old Comedy is derived from these plays, the earliest of which, *The Acharnians,* was produced in 425 BC. Aristophanic comedy has a distinct formal design but displays very little plot in any conventional sense. Rather, it presents a series of episodes aimed at illustrating, in humorous and often bawdy detail, the implications of a deadly serious political issue: it is a blend of invective, buffoonery, and song and dance. Old Comedy often used derision and scurrility, and this may have proved its undoing; though praised by all, the freedom it enjoyed degenerated into license and violence and had to be checked by law.

In New Comedy, which began to prevail around 336 BC, the Aristophanic depiction of public personages and events was replaced by a representation of the private affairs (usually amorous) of imaginary men and women. New Comedy is known only from the fragments that have survived of the plays of Menander (*c.* 342–*c.* 292 BC) and from plays written in imitation of the form by the Romans Plautus (*c.* 254–184 BC) and Terence (195 or 185–159 BC). A number of the stock comic characters survived from Old Comedy into New: an old man, a young man, an old woman, a young woman, a learned doctor or pedant, a cook, a parasite, a swaggering soldier, a comic slave. New Comedy, on the other hand, exhibits a degree of plot articulation never achieved in the Old. The action of New Comedy is usually about plotting; a clever servant, for example, devises ingenious intrigues in order that his young master may win the girl of his choice. There is satire in New Comedy: on a miser who loses his gold from being overcareful of it (the *Aulularia* of Plautus); on a father who tries so hard to win the girl from his son that he falls into a trap set for him by his wife (Plautus' *Casina*); and on an overstern father whose son turns out worse than the product of an indulgent parent (in the *Adelphi* of Terence). But the satiric quality of these plays is bland by comparison with the trenchant ridicule of Old Comedy. The emphasis in New Comic plotting is on the conduct of a love intrigue; the love element per se is often of the slightest, the girl whom the hero wishes to possess sometimes being no more than an offstage presence or, if onstage, a mute.

Development of stock characters

New Comedy provided the model for European comedy through the 18th century. During the Renaissance, the plays of Plautus and, especially, of Terence were studied for the moral instruction that young men could find in them: lessons on the need to avoid the snares of harlots and the company of braggarts, to govern the deceitful trickery of servants, to behave in a seemly and modest fashion to parents. Classical comedy was brought up to date in the plays of the "Christian Terence," imitations by schoolmasters of the comedies of the Roman dramatist. They added a contemporary flavour to the life portrayed and displayed a somewhat less indulgent attitude to youthful indiscretions than did the Roman comedy. New Comedy provided the basic conventions of plot and characterization for the *commedia erudita*—comedy performed from written texts—of 16th-century Italy, as in the plays of Machiavelli and Ariosto. Similarly, the stock characters that persisted from Old Comedy into New were taken over into the improvisational commedia dell'arte, becoming such standard masked characters as Pantalone, the Dottore, the vainglorious Capitano, the young lovers, and the servants, or *zanni.*

Rise of realistic comedy in 17th-century England. The early part of the 17th century in England saw the rise of a realistic mode of comedy based on a satiric observation of contemporary manners and mores. It was masterminded by Ben Jonson, and its purpose was didactic. Comedy, said Jonson in *Every Man Out of his Humour* (1599), quoting the definition that during the Renaissance was attributed to Cicero, is an imitation of life, a glass of custom, an image of truth. Comedy holds the mirror up to nature and reflects things as they are, to the end that society may recognize the extent of its shortcomings and the folly of its ways and set about its improvement. Jonson's greatest plays—*Volpone* (1606), *Epicœne* (1609), *The Alchemist* (1610), *Bartholomew Fair* (1614)—offer a richly detailed contemporary account of the follies and vices that are always with us. The setting (apart from *Volpone*) is Jonson's own London, and the characters are the ingenious or the devious or the grotesque products of the human wish to get ahead in the world. The conduct of a Jonsonian comic plot is in the hands of a clever manipulator who is out to make reality conform to his own desires. Sometimes he succeeds, as in the case of the clever young gentleman who gains his uncle's inheritance in *Epicœne* or the one who gains the rich Puritan widow for his wife in *Bartholomew Fair.* In *Volpone* and *The Alchemist,* the schemes eventually fail, but this is the fault of the manipulators, who will never stop when they are ahead, and not at all due to any

The rise of a realistic comedy

insight on the part of the victims. The victims are almost embarrassingly eager to be victimized. Each has his ruling passion—his humour—and it serves to set him more or less mechanically in the path that he will undeviatingly pursue, to his own discomfiture.

English comedy of the later 17th century is cast in the Jonsonian mold. Restoration comedy is always concerned with the same subject—the game of love—but the subject is treated as a critique of fashionable society. Its aim is distinctly satiric, and it is set forth in plots of Jonsonian complexity, where the principal intriguer is the rakish hero, bent on satisfying his sexual needs, outside the bonds of marriage, if possible. In the greatest of these comedies—Sir George Etherege's *Man of Mode* (1676), for example, or William Wycherley's *Country-Wife* (1675) or William Congreve's *Way of the World* (1700)—the premium is on the energy and the grace with which the game is played, and the highest dramatic approval is reserved for those who take the game seriously enough to play it with style but who have the good sense to know when it is played out. The satiric import of Restoration comedy resides in the dramatist's awareness of a familiar incongruity: that between the image of man in his primitive nature and the image of man amid the artificial restraints that society would impose upon him. The satirist in these plays is chiefly concerned with detailing the artful dodges that ladies and gentlemen employ to satisfy nature and to remain within the pale of social decorum. Inevitably, then, hypocrisy is the chief satiric target. The animal nature of man is taken for granted, and so is his social responsibility to keep up appearances; some hypocrisy must follow, and, within limits, society will wink at indiscretions so long as they are discreetly managed. The paradox is typical of those in which the Restoration comic dramatists delight; and the strongly rational and unidealistic ethos of this comedy has its affinities with the naturalistic and skeptical cast of late-17th-century philosophical thought.

Sentimental comedy of the 17th and 18th centuries. The Restoration comic style collapsed around the end of the 17th century, when the satiric vision gave place to a sentimental one. Jeremy Collier's *Short view of the Profaneness and Immorality of the English Stage,* published in 1698, signalled the public opposition to the real or fancied improprieties of plays staged during the previous three decades. "The business of plays is to recommend Vertue, and discountenance Vice": so runs the opening sentence of Collier's attack. No Restoration comic dramatist ever conceived of his function in quite these terms. "It is the business of a comic poet to paint the vices and follies of humankind," Congreve had written a few years earlier (in the dedication to *The Double-Dealer*). Though Congreve may be assumed to imply—in accordance with the time-honoured theory concerning the didactic end of comedy—that the comic dramatist paints the vices and follies of humankind for the purpose of correcting them through ridicule, he is, nonetheless, silent on this point. Collier's assumption that all plays must recommend virtue and discountenance vice has the effect of imposing on comedy the same sort of moral levy that critics such as Thomas Rymer were imposing on tragedy in their demand that it satisfy poetic justice.

The vogue of tragicomedy

At the beginning of the 18th century, there was a blending of the tragic and comic genres that, in one form or another, had been attempted throughout the preceding century. The vogue of tragicomedy may be said to have been launched in England with the publication of John Fletcher's *Faithfull Shepheardesse* (*c.* 1608), an imitation of the *Pastor fido,* by the Italian poet Battista Guarini. In his *Compendium of Tragicomic Poetry* (1601), Guarini had argued the distinct nature of the genre, maintaining it to be a third poetic kind, different from either the comic or the tragic. Tragicomedy, he wrote, takes from tragedy its great persons but not its great action, its movement of the feelings but not its disturbance of them, its pleasure but not its sadness, its danger but not its death; from comedy it takes laughter that is not excessive, modest amusement, feigned difficulty, and happy reversal. Fletcher adapted this statement in the address "To the Reader" that prefaces *The Faithfull Shepheardesse.*

The form quickly established itself on the English stage, and, through the force of such examples as Beaumont and Fletcher's *Phylaster* (1610) and *A King and No King* (1611) and a long sequence of Fletcher's unaided tragicomedies, it prevailed during the 20 years before the closing of the theatres in 1642. The taste for tragicomedy continued unabated at the Restoration, and its influence was so pervasive that during the closing decades of the century the form began to be seen in plays that were not, at least by authorial designation, tragicomedies. Its effect on tragedy can be seen not only in the tendency, always present on the English stage, to mix scenes of mirth with more solemn matters but also in the practice of providing tragedy with a double ending (a fortunate one for the virtuous, an unfortunate one for the vicious), as in Dryden's *Aureng-Zebe* (1675) or Congreve's *Mourning Bride* (1697). The general lines separating the tragic and comic genres began to break down, and that which is high, serious, and capable of arousing pathos could exist in the same play with what is low, ridiculous, and capable of arousing derision. The next step in the process came when Sir Richard Steele, bent on reforming comedy for didactic purposes, produced *The Conscious Lovers* (1722) and provided the English stage with an occasion when the audience at a comedy could derive its chief pleasure not from laughing but from weeping. It wept in the delight of seeing virtue rewarded and young love come to flower after parental opposition had been overcome. Comedy of the sort inaugurated by *The Conscious Lovers* continued to represent the affairs of private life, as comedy had always done, but with a seriousness hitherto unknown; and the traditionally low personages of comedy now had a capacity for feeling that bestowed on them a dignity previously reserved for the personages of tragedy.

This trend in comedy was part of a wave of egalitarianism that swept through 18th-century political and social thought. It was matched by a corresponding trend in tragedy, which increasingly selected its subjects from the affairs of private men and women in ordinary life, rather than from the doings of the great. The German dramatist Gotthold Lessing wrote that the misfortunes of those whose circumstances most resemble those of the audience must naturally penetrate most deeply into its heart, and his own *Minna von Barnhelm* (1767) is an example of the new serious comedy. The capacity to feel, to sympathize with, and to be affected by the plight of a fellow human being without regard for his rank in the world's esteem became the measure of one's humanity. It was a bond that united the fraternity of mankind in an aesthetic revolution that preceded the political revolutions of the 18th century. In literature, this had the effect of hastening the movement toward a more realistic representation of reality, whereby the familiar events of common life are treated "seriously and problematically" (in the phrase of the critic Erich Auerbach, who traced the process in his book *Mimesis* [1946]). The results may be seen in novels such as Samuel Richardson's *Pamela* and *Clarissa* and in middle-class tragedies such as George Lillo's *The London Merchant* (1731) in England; in the *comédie larmoyante* ("tearful comedy") in France; in Carlo Goldoni's efforts to reform the commedia dell'arte and replace it with a more naturalistic comedy in the Italian theatre; and in the English sentimental comedy, exemplified in its full-blown state by plays such as Hugh Kelly's *False Delicacy* (1768) and Richard Cumberland's *West Indian* (1771). Concerning the sentimental comedy it must be noted that it is only in the matter of appropriating for the bourgeoisie a seriousness of tone and a dignity of representational style previously considered the exclusive property of the nobility that the form can be said to stand in any significant relationship to the development of a more realistic mimetic mode than the traditional tragic and comic ones. The plots of sentimental comedy are as contrived as anything in Plautus and Terence (which with their fondness for foundling heroes who turn out to be long-lost sons of rich merchants, they often resemble); and with their delicate feelings and genteel moral atmosphere, comedies of this sort seem as affected in matters of sentiment as Restoration comedy seems in matters of wit.

The trend to egalitarianism

Goldsmith's views on sentimental comedy

Oliver Goldsmith, in his "A Comparison Between Laughing and Sentimental Comedy" (1773), noted the extent to which the comedy in the England of his day had departed from its traditional purpose, the excitation of laughter by exhibiting the follies of the lower part of mankind. He questioned whether an exhibition of its follies would not be preferable to a detail of its calamities. In sentimental comedy, Goldsmith continued, the virtues of private life were exhibited, rather than the vices exposed; and the distresses rather than the faults of mankind generated interest in the piece. Characters in these plays were almost always good; if they had faults, the spectator was expected not only to pardon but to applaud them, in consideration of the goodness of their hearts. Thus, according to Goldsmith, folly was commended instead of being ridiculed. Goldsmith concluded by labelling sentimental comedy a "species of bastard tragedy," "a kind of *mulish* production": a designation that ironically brings to mind Guarini's comparison of tragicomedy in its uniqueness (a product of comedy and tragedy but different from either) to the mule (the offspring of the horse and the ass but itself neither one nor the other). The production of Goldsmith's *She Stoops to Conquer* (1773) and of Richard Brinsley Sheridan's *Rivals* (1775) and *The School for Scandal* (1777) briefly reintroduced comic gaiety to the English stage; by the end of the decade, Sheridan's dramatic burlesque, *The Critic* (first performed 1779), had appeared, with its parody of contemporary dramatic fashions, the sentimental included. But this virtually concluded Sheridan's career as a dramatist; Goldsmith had died in 1774; and the sentimental play was to continue to govern the English comic stage for over a century to come.

The comic outside the theatre. The great comic voices of the 18th century in England were not those in the theatre. No dramatic satire of the period can exhibit anything comparable to the furious ridicule of man's triviality and viciousness that Jonathan Swift provided in *Gulliver's Travels* (1726). His *Modest Proposal* (1729) is a masterpiece of comic incongruity, with its suave blend of rational deliberation and savage conclusion. The comic artistry of Alexander Pope is equally impressive. Pope expressed his genius in the invective of his satiric portraits and in the range of moral and imaginative vision that was capable, at one end of his poetic scale, of conducting that most elegant of drawing-room epics, *The Rape of the Lock* (1712–14), to its sublimely inane conclusion and, at the other, of invoking from the scene that closes *The Dunciad* (1728) an apocalyptic judgment telling what will happen when the vulgarizers of the word have carried the day.

Censorship

When the voice of comedy did sound on the 18th-century English stage with anything approaching its full critical and satiric resonance, the officials soon silenced it. John Gay's *Beggar's Opera* (1728) combined hilarity with a satiric fierceness worthy of Swift (who may have suggested the original idea for it). The officials tolerated its spectacularly successful run, but no license from the lord chamberlain could be secured for Gay's sequel, *Polly,* which was not staged until 1777. The Licensing Act of 1737 ended the theatrical career of Henry Fielding, whose comedies had come under constant fire from the authorities for their satire on the government. Fielding's comic talents were perforce directed to the novel, the form in which he parodied the sentiment and the morality of Richardson's *Pamela*—in his *Shamela* and *Joseph Andrews* (1742)—as brilliantly as he had earlier burlesqued the rant of heroic tragedy in *Tom Thumb* (1730).

Comedy of the sort that ridicules the follies and vices of society to the end of laughing them out of countenance entered the English novel with Fielding. His statement in *Joseph Andrews* concerning the function of satire is squarely in the Neoclassic tradition of comedy as a corrective of manners and mores: the satirist holds

> the glass to thousands in their closets, that they may contemplate their deformity, and endeavour to reduce it, and thus by suffering private mortification may avoid public shame.

Fielding's scenes of contemporary life display the same power of social criticism as that which distinguishes the engravings of his great fellow artist William Hogarth, whose "Marriage à la Mode" (1745) depicts the vacuity and the casual wantonness of the fashionable world that Fielding treats of in the final books of *Tom Jones.* Hogarth's other series, such as "A Rake's Progress" (1735) or "A Harlot's Progress" (1732), also make a didactic point about the wages of sin, using realistic details heightened with grotesquerie to expose human frailty and its sinister consequences. The grotesque is a recurrent feature of the satiric tradition in England, where comedy serves social criticism. Artists such as Hogarth and Thomas Rowlandson worked in the tradition of Jonson and the Restoration dramatists in the preceding century.

The novel, with its larger scope for varied characters, scenes, and incidents, rather than the drama, afforded the 19th-century artist in comedy a literary form adequate to his role as social critic. The spectacle of man and his society is regularly presented by the 19th-century novelist in comedic terms, as in *Vanity Fair* (1848), by William Makepeace Thackeray or the *Comédie humaine* (1842–55) of Honoré de Balzac, and with the novels of Jane Austen, Anthony Trollope, Charles Dickens, and George Meredith.

20th-century tragicomedy. The best that the comic stage had to offer in the late 19th century lay in the domain of farce. The masters of this form were French, but it flourished in England as well; what the farces of Eugène Labiche and Georges Feydeau and the operettas of Jacques Offenbach were to the Parisian stage the farces of W.S. Gilbert and the young Arthur Wing Pinero and the operettas that Gilbert wrote in collaboration with Arthur Sullivan were to the London stage. As concerns comedy, the situation in England improved at the end of the century, when Oscar Wilde and George Bernard Shaw turned their talents to it. Wilde's *Importance of Being Earnest* (1895) is farce raised to the level of high comic burlesque. Shaw's choice of the comic form was inevitable, given his determination that the contemporary English stage should deal seriously and responsibly with the issues that were of crucial importance to contemporary English life. Serious subjects could not be resolved by means of the dramatic clichés of Victorian melodrama. Rather, the prevailing stereotypes concerning the nature of honour, courage, wisdom, and virtue were to be subjected to a hail of paradox, to the end of making evident their inner emptiness or the contradictions they concealed.

French farce

Shaw dealt with what, in the preface to *Major Barbara* (1905), he called "the tragi-comic irony of the conflict between real life and the romantic imagination," and his use of the word tragicomic is a sign of the times. The striking feature of modern art, according to the German novelist Thomas Mann, was that it had ceased to recognize the categories of tragic and comic or the dramatic classifications of tragedy and comedy but saw life as tragicomedy. The sense that tragicomedy is the only adequate dramatic form for projecting the unreconciled ironies of modern life mounted through the closing decades of the 19th century. Ibsen had termed *The Wild Duck* (published 1884) a tragicomedy; it was an appropriate designation for this bitter play about a young man blissfully ignorant of the lies on which he and his family have built their happy life until an outsider who is committed to an ideal of absolute truth exposes all their guilty secrets with disastrous results. The plays of the Russian writer Anton Chekhov, with their touching and often quite humorous figures leading lives of quiet desperation, reflect precisely that mixture of inarticulate joy and dull pain that is the essence of the tragicomic view of life.

A dramatist such as August Strindberg produces a kind of tragicomedy peculiarly his own, one that takes the form of bourgeois tragedy; it lacerates its principals until they become a parody of themselves. Strindberg's *Dance of Death* (1901), with its cruelty and pain dispensed with robust pleasure by a fiercely battling husband and wife, is a significant model of the grotesque in the modern theatre; it is reflected in such mid-20th-century examples of what came to be called black comedy as Eugène Ionesco's *Victims of Duty* (1953) and Edward Albee's *Who's Afraid of Virginia Woolf?* (1962). Almost equally influential as a turn-of-the-century master of the grotesque is Frank Wedekind, whose

Earth Spirit (1895) and its sequel, *Pandora's Box* (written 1892–1901), though both are termed tragedies by their author, are as much burlesques of tragedy as *The Dance of Death.* Their grotesquerie consists chiefly in their disturbing combination of innocence and depravity, of farce and horror, of passionate fervour issuing in ludicrous incident that turns deadly. Wedekind's celebration of primitive sexuality and the varied ways in which it manifests itself in an oversophisticated civilization distorts the tragic form to achieve its own grotesque beauty and power.

The great artist of the grotesque and of tragicomedy in the 20th century is the Italian Luigi Pirandello. His drama is explicitly addressed to the contradictoriness of experience: appearances collide and cancel out each other; the quest of the absolute issues in a mind-reeling relativism; infinite spiritual yearnings are brought up hard against finite physical limits; rational purpose is undermined by irrational impulse; and with the longing for permanence in the midst of change comes the ironic awareness that changelessness means death. Stated thus, Pirandello's themes sound almost forbiddingly intellectual, but one of his aims was to convert intellect into passion. Pirandello's characters suffer from intellectual dilemmas that give rise to mental and emotional distress of the most anguished kind, but their sufferings are placed in a satiric frame. The incongruities that the characters are furiously seeking to reconcile attest to the comic aspect of this drama, but there is nothing in it of the traditional movement of comedy, from a state of illusion into the full light of reality. Pirandello's characters dwell amid ambiguities and equivocations that those who are wise in the tragicomic nature of life will accept without close inquiry. The logic of comedy implies that illusions exist to be dispelled; once they are dispelled, everyone will be better off. The logic of Pirandello's tragicomedy demonstrates that illusions make life bearable; to destroy them is to destroy the basis for any possible happiness.

The role of illusions in Pirandello's tragicomedy

The absurd. In their highly individual ways, both Samuel Beckett and Ionesco have employed the forms of comedy—from tragicomedy to farce—to convey the vision of an exhausted civilization and a chaotic world. The very endurance of life amid the grotesque circumstances that obtain in Beckett's plays is at once a tribute to the human power of carrying on to the end and an ironic reflection on the absurdity of doing so. Beckett's plays close in an uneasy silence that is the more disquieting because of the uncertainty as to just what it conceals: whether it masks sinister forces ready to spring or is the expression of a universal indifference or issues out of nothing at all.

Silence seldom reigns in the theatre of Ionesco, which rings with voices raised in a usually mindless clamour. Some of Ionesco's most telling comic effects have come from his use of dialogue overflowing with clichés and non sequiturs, which make it clear that the characters do not have their minds on what they are saying and, indeed, do not have their minds on anything at all. What they say is often at grotesque variance with what they do. Beneath the moral platitudes lurks violence, which is never far from the surface in Ionesco's plays, and the violence tells what happens to societies in which words and deeds have become fatally disjunct. Ionesco's comic sense is evident as well in his depiction of human beings as automata, their movements decreed by forces they have never questioned or sought to understand. There is something undeniably farcical in Ionesco's spectacles of human regimentation, of men and women at the mercy of things (*e.g.,* the stage full of chairs in *The Chairs* or the growing corpse in *Amédée*); the comic quality here is one that Bergson would have appreciated. But the comic in Ionesco's most serious work, as in so much of the contemporary theatre, has ominous implications that give to it a distinctly grotesque aspect. In Ionesco's *Victims of Duty* and *The Killer* (1959), as in the works of his Swiss counterparts—*Der Besuch der alten Dame* (performed 1956; *The Visit,* 1958) and *The Physicists* (1962), by Friedrich Dürrenmatt, and *The Firebugs* (1958), by Max Frisch—the grotesquerie of the tragicomic vision delineates a world in which the humane virtues are dying, and casual violence is the order of the day.

Ionesco's comic sense

The radical reassessment of the human image that the 20th century has witnessed is reflected in the novel as well as in drama. Previous assumptions about the rational and divine aspects of man have been increasingly called into question by the evidences of man's irrationality, his sheer animality. These are qualities of human nature that writers of previous ages (Swift, for example) have always recognized, but hitherto they have been typically viewed as dark possibilities that could overtake humanity if the rule of reason did not prevail. It is only in the mid-20th century that the savage and the irrational have come to be viewed as part of the normative condition of humanity rather than as tragic aberrations from it. The savage and the irrational amount to grotesque parodies of human possibility, ideally conceived. Thus it is that 20th-century novelists as well as dramatists have recognized the tragicomic nature of the contemporary human image and predicament, and the principal mode of representing both is the grotesque. This may take various forms: the apocalyptic nightmare of tyranny and terror in Kafka's novels *The Trial* (1925) and *The Castle* (1926); the tragic farce in terms of which the Austrian novelist Robert Musil describes the slow collapse of a society into anarchy and chaos, in *The Man Without Qualities* (1930–43); the brilliant irony whereby Thomas Mann represents the hero as a confidence man in *The Confessions of Felix Krull* (1954); the grimly parodic account of Germany's descent into madness in Günter Grass's novel *The Tin Drum* (1959). The English novel contains a rich vein of the comic grotesque that extends at least back to Dickens and Thackeray and persisted in the 20th century in such varied novels as Evelyn Waugh's *Decline and Fall* (1928), Angus Wilson's *Anglo-Saxon Attitudes* (1956), and Kingsley Amis' *Lucky Jim* (1954). What novelists such as these have in common is the often disturbing combination of hilarity and desperation. It has its parallel in a number of American novels—John Barth's *Giles Goat-Boy* (1966), Kurt Vonnegut, Jr.'s *Slaughterhouse Five* (1969)—in which shrill farce is the medium for grim satire. And the grotesque is a prominent feature of modern poetry, as in some of the "Songs and Other Musical Pieces" of W.H. Auden.

THE COMIC IN OTHER ARTS

The visual arts. The increasing use of the affairs of common life as the subject matter of dramatic comedy through the Middle Ages and the Renaissance is also seen in painting of that time. Scenes from medieval mystery cycles, such as the comic episodes involving Noah's stubborn wife, have counterparts in medieval pictures in the glimpses of everyday realities that are caught through the windows or down the road from the sites where the great spiritual mysteries are in progress: the angel Gabriel may appear to the Virgin in the foreground, while a man is chopping wood in the yard outside. Medieval artists had never neglected the labours and the pleasures of the mundane world, but the treatment of them is often literally marginal, as in the depiction of men and women at work or play in the ornamental borders of an illuminated manuscript page. The seasonal round of life, with its cycle of plowing, sowing, mowing, and reaping interspersed with hawking, hunting, feasts, and weddings (the cycle of life, indeed, which comedy itself celebrates), is depicted in series after series of exquisite miniatures, such as those in the *Très Riches Heures du Duc de Berry.* By the mid-16th century, however, in Pieter Bruegel's famous painting "Landscape with the Fall of Icarus," mundane reality has taken over the foreground; the plowman tills the soil, and the shepherd attends his flock, while, unnoticed by both, the legs of Icarus disappear inconspicuously into the sea. Bruegel is not a comic artist, but his art bears witness to what all great comic art celebrates: the basic rhythm of life. "Peasant Wedding" and "Peasant Dance" endow their heavy men and women with an awkward grace and dignity that bear comparison with Shakespeare's treatment of his comic characters. Paintings like Bruegel's "Children's Games" and his "Fight Between Carnival and Lent" are joyous representations of human energy. The series of "The Labours of the Months"—"Hunters in the Snow" for January, "Haymaking" for July, "Harvesters" for August, "Return of the Herd" for November—give

Bruegel's mastery of the grotesque

pictorial treatment to a favourite subject of the medieval miniaturists. Finally, allusion must be made to Bruegel's mastery of the grotesque, notably in "The Triumph of Death" and in the "Dulle Griet," in which demons swarm over a devastated landscape.

It is through the art of caricature that the spirit of comedy enters most directly into painting. The style derives from the portraits with ludicrously exaggerated features made by the Carracci, an Italian family of artists, early in the 17th century (Italian *caricare,* "to overload"). In defiance of the theory of ideal beauty, these portraits emphasized the features that made one man different from another. This method of character portrayal—the singling out of one distinctive feature and emphasizing it over all others—is not unlike the practice of characterizing the personages of the comic stage by means of some predominant humour, which Ben Jonson was developing at about the same time in the London theatre. The use of exaggeration for comic effect was as evident to painters as it was to dramatists. Its usefulness as a means of social and political satire is fully recognized by Hogarth. Hogarth's counterpart in mid-19th-century Paris was Honoré Daumier. His caricatures portray a human comedy as richly detailed and as shrewdly observed as the one portrayed in fiction by his contemporary Balzac. But Daumier's sense of the comic goes beyond caricature; his numerous treatments of scenes from Molière's plays and, most notably, his drawings and canvases of Don Quixote and Sancho Panza attest to the pathos that can lie beneath the comic mask.

Modern art has abstracted elements of comedy to aid it in the representation of a reality in which the mechanical is threatening to win out over the human. Bergson's contention that the essence of comedy consists of something mechanical encrusted on the living may be said to have achieved a grotesque apotheosis in the French Dadaist Marcel Duchamp's painting "Bride" (1912), in which the female figure has been reduced to an elaborate piece of plumbing. The highly individual Swiss Expressionist Paul Klee's pen-and-ink drawing tinted with watercolour and titled "Twittering Machine" (1922) represents an ingenious device for imitating the sound of birds. The delicacy of the drawing contrasts with the sinister implications of the mechanism, which, innocent though it may appear at first glance, is almost certainly a trap.

The grotesque is a constant stylistic feature of the artist's representation of reality in its brutalized or mechanized aspects. The carnival masks worn by the figures in the painting "Intrigue" (1890), by the Belgian James Ensor, make manifest the depravity and the obscenity that lurk beneath the surface of conventional appearances; Ensor's paintings make much the same point about the persistence of the primitive and the savage into modern life as Wedekind's plays were to do a few years later. German artists after World War I invoked the grotesque with particular power, depicting the inhuman forces that bear upon the individual, as in George Grosz's savage cartoon titled "Germany, a Winter's Tale" (1918), in which the puppet-like average citizen sits at table surrounded by militarist, capitalist, fatuous clergyman and all the violent and dissolute forces of a decadent society. The mutilated humanity in Max Beckmann's "Dream" (1921) and "Departure" (1932–33) is a further testament to human viciousness, 20th-century variety.

Fantasy in modern paintings

Rather more explicitly comic is the element of fantasy in modern paintings, in which seemingly unrelated objects are brought together in a fine incongruity, as in the French primitive Henri Rousseau's famous "Dream" (1910), with its nude woman reclining on a red-velvet sofa amid the flora and fauna of a lush and exotic jungle. The disparate figures that float (in defiance of all the laws of gravity) through the paintings of the Russian Surrealist Marc Chagall are individually set forth in a nimbus of memory and in the landscape of dream. But fantasy can take on a grotesquerie of its own, as in some of Chagall's work, such as the painting "I and The Village" (1911).

The purest expression of the comic in modern painting must surely be Henri Matisse's "Joy of Life" (1905–06), a picture that might be taken as a visual expression of the precept that the rhythm of comedy is the basic rhythm of life. But Matisse's painting was not to be the last word on the subject: "Joy of Life" produced, as a counterstatement, Pablo Picasso's "Demoiselles d'Avignon" (1906–07), in which the daughters of joy, in their grim and aggressive physical tension, stand as a cruel parody of the delight in the senses that Matisse's picture celebrates. "Les Demoiselles d'Avignon" and such a later Picasso masterpiece as the "Three Dancers" (1925) suggest that, for the visual as well as the literary artist of the 20th century, the joy of life tends to issue in grotesque shapes.

Music. Given the wide range of imitative sounds of which musical instruments and the human voice are capable, comic effects are readily available to the composer who wants to use them. At the simplest level, these may amount to nothing more than humorous adjuncts to a larger composition, such as the loud noise with which the 18th-century Austrian composer Joseph Haydn surprises his listeners in *Symphony No. 94* or the sound of the ticking clock in *No. 101.* The scherzo, which Ludwig van Beethoven introduced into symphonic music in the early 19th century, may be said to have incorporated in it a musical joke but one of a highly abstract kind; its nervous jocularity provides a contrast and a commentary (both heavy with irony) on the surrounding splendour. A century after Beethoven, the jocularity grew more desperate and the irony more profound in the grim humour that rises out of the grotesque scherzos of Gustav Mahler. A more sustained and a more explicit musical exposition of comic themes and attitudes comes when a composer draws his inspiration directly from a work of comic literature, as Richard Strauss does in his orchestral variations based on *Don Quixote* and on the merry pranks of *Till Eulenspiegel.*

The form imposed by music

It is, however, opera that provides the fullest form for comedy to express itself in music, and some of the most notable achievements of comic art have been conceived for the operatic stage. High on any list of comic masterpieces must come the four principal operas of Mozart: *The Marriage of Figaro* (1786), *Don Giovanni* (1787), *Così fan tutte* (1790), and *The Magic Flute* (1791), and there are countless others worthy of mention. Operatic comedy has an advantage over comedy in the spoken theatre in its ability to impose a coherent form on the complexities of feeling and action that are often of the essence in comedy. The complex feeling experienced by different characters must be presented in spoken comedy seriatim; operatic comedy can present them simultaneously. When three or four characters talk simultaneously in the spoken theatre, the result is an incoherent babble. But the voices of three or four or even more characters can be blended together in an operatic ensemble, and, while most of the words may be lost, the vocal lines will serve to identify the individual characters and the general nature of the emotions they are expressing. The complexities of action in the spoken theatre are the chief source of the comic effect, which increases as the confusion mounts; such complexities of action operate to the same comic end in opera but here with the added ingredient of music, which provides an overarching design of great formal coherence. In the music, all is manifestly ordered and harmonious, while the events of the plot appear random and chaotic; the contrast between the movement of the plot and the musical progression provides a Mozart or a Rossini with some of his wittiest and most graceful comic effects. Finally, it should be noted that operatic comedy can probe psychological and emotional depths of character that spoken comedy would scarcely attempt. The Countess in Mozart's *Figaro* is a very much more moving figure than she is in Beaumarchais' play; the Elvira of *Don Giovanni* exhibits a fine extravagance that is little more than suggested in Molière's comedy.

Television and cinema. When comedy is dependent on the favour of a large part of the public, as reflected in box-office receipts or the purchase of a television sponsor's product, it seldom achieves a high level of art. There is nothing innocent about laughter at the whims and inconsistencies of humankind, and radio and television and film producers have always been wary of offending their audiences with it. On radio and television, the laughter is

usually self-directed (as in the performances of comedians such as Jack Benny or Red Skelton), or it is safely contained within the genial confines of a family situation (*e.g.*, the "Fibber McGee and Molly" radio show or "I Love Lucy" on television). Much the same attitude has obtained with regard to comedy in the theatre in the United States. Satire has seldom succeeded on Broadway, which instead has offered pleasant plays about the humorous behaviour of basically nice people, such as the eccentric family in George S. Kaufmann and Moss Hart's *You Can't Take It with You* (1936) or the lovable head of the household in Howard Lindsay and Russel Crouse's *Life with Father* (1939) or the indefatigable Dolly Levi in Thornton Wilder's *Matchmaker* (1954) and in her later reincarnation in the musical *Hello, Dolly!*

The American public and comedy

The American public has never been quite comfortable in the presence of comedy. The calculated ridicule and the relentless exposure often seem cruel or unfair to a democratic public. If all men are created equal, then it ill becomes anyone to laugh at the follies of his fellows, especially when they are follies that are likely to be shared, given the common background of social opportunity and experience of the general public. There is an insecurity in the mass audience that is not compatible with the high self-assurance of comedy as it judges between the wise and the foolish of the world. The critical spirit of comedy has never been welcome in American literature; in both fiction and drama, humour, not comedy, has raised the laughter. American literature can boast an honorable tradition of humorists, from Mark Twain to James Thurber, but has produced no genuinely comic writer. As American social and moral tenets were subjected to increasing critical scrutiny from the late 1950s onward, however, there were some striking achievements in comedy in various media: Edward Albee's *American Dream* (1961) and *Who's Afraid of Virginia Woolf?* (1962), on the stage; novels such as those of Saul Bellow and Joseph Heller's *Catch-22* (1961); and films such as *Dr. Strangelove* (1964).

This last example is remarkable, because comedy in the medium of film, in America, had been conceived as entertainment and not much more. This is not to say that American film comedies lacked style. The best of them always displayed verve and poise and a thoroughly professional knowledge of how to amuse the public without troubling it. Their shortcoming has always been that the amusement they provide lacks resonance.

If films have seldom explored comedy with great profundity, they have, nonetheless, produced it in great variety. There have been comedies of high sophistication, the work of directors such as Ernst Lubitsch, George Cukor, Frank Capra, Joseph L. Mankiewicz, and Billy Wilder and of actors and actresses such as Greta Garbo (in Lubitsch's *Ninotchka,* 1939), Katharine Hepburn and Cary Grant (in Cukor's *Philadelphia Story,* 1940), Bette Davis (in Mankiewicz' *All About Eve,* 1950), Clark Gable and Claudette Colbert (in Capra's *It Happened One Night,* 1934), Gary Cooper and Jean Arthur (in Capra's *Mr. Deeds Goes to Town,* 1936), and Marilyn Monroe and Jack Lemmon (in Billy Wilder's *Some Like It Hot,* 1959). There have been comedies with music, built around the talents of singers and dancers such as Ruby Keeler and Dick Powell and Ginger Rogers and Fred Astaire; there are the classic farces of Charlie Chaplin and Buster Keaton and, later, of W.C. Fields and the Marx Brothers and Laurel and Hardy; and there is a vast, undistinguished field of comedies dealing with the humours of domestic life. The varieties of comedy in Hollywood films have always been replicas of those on the New York stage; as often as not, they were products of the same talents: in the 1930s, of dramatists such as Philip Barry or S.N. Behrman and composers such as Cole Porter, Richard Rodgers, and Irving Berlin; in the 1960s, of the dramatist Neil Simon and the composer Burt Bacharach.

Comedies produced by European film makers

European film makers, with an older and more intellectual tradition of comedy available to them, produced comedies of more considerable stature. Among French directors, Jean Renoir, in his *The Rules of the Game* (1939), conveyed a moving human drama and a profoundly serious vision of French life on the eve of World War II in a form, deriving from the theatre, that blends the comic and the tragic. His disciple François Truffaut, in *Jules and Jim* (1961), directed a witty and tender but utterly clear-sighted account of how gaiety and love turn deadly. Though not generally regarded as a comic artist, the Swedish film maker Ingmar Bergman produced a masterpiece of film comedy in *Smiles of a Summer Night* (1955), a wise, wry account of the indignities that must sometimes be endured by those who have exaggerated notions of their wisdom or virtue. The films of the Italian director and writer Federico Fellini represent a comic vision worthy of Pirandello. *La strada* (1954), with its Chaplinesque waif (played by Fellini's wife, Giulietta Masina) as central figure, is a disturbing compound of pathos and brutality. Comedy's affirmation of the will to go on living has had no finer portrayal than in Giulietta Masina's performance in the closing scene of *Nights of Cabiria* (1956). *La dolce vita* (1960) is a luridly satiric vision of modern decadence, where ideals are travestied by reality, and everything is illusion and disillusionment; the vision is carried to even more bizarre lengths in Fellini's *Satyricon* (1969), in which the decadence of the modern world is grotesquely mirrored in the ancient one. *8½* (1963) and *Juliet of the Spirits* (1965) are Fellini's most brilliantly inventive films, but their technical exuberance is controlled by a profoundly serious comic purpose. The principals in both films are seeking—through the phantasmagoria of their past and present, of their dreams and their delusions, all of which seem hopelessly mixed with their real aspirations—to know themselves. (C.H.Ho.)

Tragedy

Although the word tragedy is often used loosely to describe any sort of disaster or misfortune, it more precisely refers to a work of art, usually a play or novel, that probes with high seriousness questions concerning the role of man in the universe. The Greeks of Attica, the ancient state whose chief city was Athens, first used the word in the 5th century BC to describe a specific kind of play, which was presented at festivals in Greece. Sponsored by the local governments, these plays were attended by the entire community, a small admission fee being provided by the state for those who could not afford it themselves. The atmosphere surrounding the performances was more like that of a religious ceremony than entertainment. There were altars to the gods, with priests in attendance, and the subjects of the tragedies were the misfortunes of the heroes of legend, religious myth, and history. Most of the material was derived from the works of Homer and was common knowledge in the Greek communities. So powerful were the achievements of the three greatest Greek dramatists—Aeschylus (525–456 BC), Sophocles (*c.* 496–406 BC), and Euripides (*c.* 480–406 BC)—that the word they first used for their plays survived and came to describe a literary genre that, in spite of many transformations and lapses, has proved its viability through 25 centuries.

Four great periods of tragedy

Historically, tragedy of a high order has been created in only four periods and locales: Attica, in Greece, in the 5th century BC; England in the reigns of Elizabeth I and James I, from 1558 to 1625; 17th-century France; and Europe and America during the second half of the 19th century and the first half of the 20th. Each period saw the development of a special orientation and emphasis, a characteristic style of theatre. In the modern period, roughly from the middle of the 19th century, the idea of tragedy found embodiment in the collateral form of the novel.

This section focusses primarily on the development of tragedy as a literary genre. Further information on the relationship of tragedy to other types of drama will be found in the section above on *Dramatic literature.* The role of tragedy in the growth of theatre is discussed in THEATRE, THE HISTORY OF WESTERN.

DEVELOPMENT

Origins in Greece. The questions of how and why tragedy came into being and of the bearing of its origins on its development in subsequent ages and cultures have been

investigated by historians, philologists, archaeologists, and anthropologists with results that are suggestive but conjectural. Even the etymology of the world tragedy is far from established. The most generally accepted source is the Greek *tragōidia,* or "goat-song," from *tragos* ("goat") and *aeidein* ("to sing"). The word could have referred either to the prize, a goat, that was awarded to the dramatists whose plays won the earliest competitions or to the dress (goat skins) of the performers, or to the goat that was sacrificed in the primitive rituals from which tragedy developed.

In these communal celebrations, a choric dance may have been the first formal element and perhaps for centuries was the principal element. A speaker was later introduced into the ritual, in all likelihood as an extension of the role of the priest, and dialogue was established between him and the dancers, who became the chorus in the Athenian drama. Aeschylus is usually regarded as the one who, realizing the dramatic possibilities of the dialogue, first added a second speaker and thus invented the form of tragedy. That so sophisticated a form could have been fully developed by a single artist, however, is scarcely credible. Hundreds of early tragedies have been lost, including some by Aeschylus himself. Of some 90 plays attributed to him, only seven have survived.

Four Dionysia, or feasts of the Greek God Dionysus, were held annually in Athens. Since Dionysus once held place as the god of vegetation and the vine, and the goat was believed sacred to him, it has been conjectured that tragedy originated in fertility feasts to commemorate the harvest and the vintage and the associated ideas of the death and renewal of life. The purpose of such rituals is to exercise some influence over these vital forces. Whatever the original religious connections of tragedy may have been, two elements have never entirely been lost: (1) its high seriousness, befitting matters in which survival is at issue and (2) its involvement of the entire community in matters of ultimate and common concern. When either of these elements diminishes, when the form is overmixed with satiric, comic, or sentimental elements, or when the theatre of concern succumbs to the theatre of entertainment, then tragedy falls from its high estate and is on its way to becoming something else.

Two vital elements

As the Greeks developed it, the tragic form, more than any other, raised questions about man's existence. Why must man suffer? Why must man be forever torn between the seeming irreconcilables of good and evil, freedom and necessity, truth and deceit? Are the causes of his suffering outside himself, in blind chance, in the evil designs of others, in the malice of the gods? Are its causes within him, and does he bring suffering upon himself through arrogance, infatuation, or the tendency to overreach himself? Why is justice so elusive?

Aeschylus: the first great tragedian. It is this last question that Aeschylus asks most insistently in his two most famous works, the *Oresteia* (a trilogy comprising *Agamemnon, Choephoroi,* and *Eumenides*) and *Prometheus Bound* (the first part of a trilogy of which the last two parts have been lost): is it right that Orestes, a young man in no way responsible for his situation, should be commanded by a god, in the name of justice, to avenge his father by murdering his mother? Is there no other way out of his dilemma than through the ancient code of blood revenge, which will only compound the dilemma? Again: was it right that in befriending mankind with the gifts of fire and the arts, Prometheus should offend the presiding god Zeus and himself be horribly punished? Aeschylus opened questions whose answers in the Homeric stories had been taken for granted. In Homer, Orestes' patricide is regarded as an act of filial piety, and Prometheus' punishment is merely the inevitable consequence of defying the reigning deity. All of the materials of tragedy, all of its cruelty, loss, and suffering, are present in Homer and the ancient myths but are dealt with as absolutes—self-sufficient and without the questioning spirit that was necessary to raise them to the level of tragedy. It remained for Aeschylus and his fellow tragedians first to treat these "absolutes" critically and creatively in sustained dramatic form. They were true explorers of the human spirit.

In addition to their remarkable probing into the nature of existence, their achievements included a degree of psychological insight for which they are not generally given credit. Though such praise is usually reserved for Shakespeare and the moderns, the Athenian dramatists conveyed a vivid sense of the living reality of their characters' experience: of what it felt like to be caught, like Orestes, in desperately conflicting loyalties or to be subjected, like Prometheus, to prolonged and unjust punishment. The mood of the audience as it witnessed the acting out of these climactic experiences has been described as one of impassioned contemplation. From their myths and epics and from their history in the 6th century, the people of Athens learned that they could extend an empire and lay the foundations of a great culture. From their tragedies of the 5th century, they learned who they were, something of the possibilities and limitations of the spirit, and of what it meant, not merely what it felt like, to be alive in a world both beautiful and terrible.

Aeschylus has been called the most theological of the Greek tragedians. His *Prometheus* has been compared to the Book of Job of the Bible both in its structure (*i.e.,* the immobilized heroic figure maintaining his cause in dialogues with visitors) and in its preoccupation with the problem of suffering at the hands of a seemingly unjust deity. Aeschylus tended to resolve the dramatic problem into some degree of harmony, as scattered evidence suggests he did in the last two parts of the *Promethiad* and as he certainly did in the conclusion of the *Oresteia.* This tendency would conceivably lead him out of the realm of tragedy and into religious assurance. But his harmonies are never complete. In his plays evil is inescapable, loss is irretrievable, suffering is inevitable. What the plays say positively is that man can learn through suffering. The chorus in *Agamemnon,* the first play of the *Oresteia,* says this twice. The capacity to learn through suffering is a distinguishing characteristic of the tragic hero, preeminently of the Greek tragic hero. He has not merely courage, tenacity, and endurance but also the ability to grow, by means of these qualities, into an understanding of himself, of his fellows, and of the conditions of existence. Suffering, says Aeschylus, need not be embittering but can be a source of knowledge. The moral force of his plays and those of his fellow tragedians can hardly be exaggerated. They were shaping agents in the Greek notion of education. It has been said that from Homer the Greeks learned how to be good Greeks; from the tragedies they learned an enlarged humanity. If it cannot be proved that Aeschylus "invented" tragedy, it is clear that he at least set its tone and established a model that is still operative. Even in the 20th century, the *Oresteia* has been acclaimed as the greatest spiritual work of man, and dramatists such as T.S. Eliot, in *The Family Reunion* (1939), and Jean-Paul Sartre, in *The Flies* (1943), found modern relevance in its archetypal characters, situations, and themes.

Aeschylus' view of the problem of suffering

Sophocles: the purest artist. Sophocles' life spanned almost the whole of the 5th century. He is said to have written his last play, *Oedipus at Colonus,* at the age of 90. Only seven of his plays, of some 125 attributed to him, survive. He won the prize in the tragic competitions 20 times and never placed lower than second.

Sophocles has been called the great mediating figure between Aeschylus and Euripides. Of the three, it might be said that Aeschylus tended to resolve tragic tensions into higher truth, to look beyond, or above, tragedy; that Euripides' irony and bitterness led him the other way to fix on the disintegration of the individual; and that Sophocles, who is often called the "purest" artist of the three, was truest to the actual state of human experience. Unlike the others, Sophocles seems never to insinuate himself into his characters or situations, never to manipulate them into preconceived patterns. He sets them free on a course seemingly of their own choosing. He neither preaches nor rails. If life is hard and often destructive, the question Sophocles asks is not how did this come to be or why did such a misfortune have to happen but rather, given the circumstances, how must a man conduct himself, how should he act, what must he do?

His greatest play, *Oedipus the King,* may serve as a model of his total dramatic achievement. Embodied in it,

and suggested with extraordinary dramatic tact, are all the basic questions of tragedy, which are presented in such a way as almost to define the form itself. It is not surprising that Aristotle, a century later, analyzed it for his definition of tragedy in the *Poetics*. It is the nuclear Greek tragedy, setting the norm in a way that cannot be claimed for any other work, not even the *Oresteia*.

Sophocles' emphasis on action

In *Oedipus,* as in Sophocles' other plays, the chorus is much less prominent than in Aeschylus' works. The action is swifter and more highly articulated; the dialogue is sharper, more staccato, and bears more of the meaning of the play. Though much has been made of the influence of fate on the action of the play, later critics emphasize the freedom with which Oedipus acts throughout. Even before the action of the play begins, the oracle's prediction that Oedipus was doomed to kill his father and marry his mother had long since come true, though he did not realize it. Though he was fated, he was also free throughout the course of the play—free to make decision after decision, to carry out his freely purposed action to its completion. In him, Sophocles achieved one of the enduring definitions of the tragic hero—that of a man for whom the liberation of the self is a necessity. The action of the play, the purpose of which is to discover the murderer of Oedipus' father and thereby to free the city from its curse, leads inevitably to Oedipus' suffering—the loss of his wife, his kingdom, his sight. The messenger who reports Oedipus' self-blinding might well have summarized the play with "All ills that there are names for, all are here." And the chorus' final summation deepens the note of despair: "Count no man happy," they say in essence, "until he is dead."

But these were not Sophocles' ultimate verdicts. The action is so presented that the final impression is not of human helplessess at the hands of malign gods nor of man as the pawn of fate. Steering his own course, with great courage, Oedipus has ferreted out the truth of his identity and administered his own punishment, and, in his suffering, learned a new humanity. The final impression of the *Oedipus,* far from being one of unmixed evil and nihilism, is of massive integrity, powerful will, and magnanimous acceptance of a horribly altered existence.

Some 50 years later, Sophocles wrote a sequel to *Oedipus the King.* In *Oedipus at Colonus,* the old Oedipus, further schooled in suffering, is seen during his last day on earth. He is still the same Oedipus in many ways: hot-tempered, hating his enemies, contentious. Though he admits his "pollution" in the murder of his father and the marriage to his mother, he denies that he had sinned, since he had done both deeds unwittingly. Throughout the play, the theme of which has been described as the "heroization" of Oedipus, he grows steadily in nobility and awesomeness. Finally, sensing the approach of the end, he leaves the scene, to be elevated in death to a demigod, as the messenger describes the miraculous event. In such manner Sophocles leads his tragedy toward an ultimate assertion of values. His position has been described as "heroic humanism," as making a statement of belief in the human capacity to transcend evils, within and without, by means of the human condition itself.

Tragedy must maintain a balance between the higher optimisms of religion or philosophy, or any other beliefs that tend to explain away the enigmas and afflictions of existence, on the one hand, and the pessimism that would reject the whole human experiment as valueless and futile on the other. Thus the opposite of tragedy is not comedy but the literature of cynicism and despair, and the opposite of the tragic artist's stance, which is one of compassion and involvement, is that of the detached and cynical ironist.

Euripides: the dark tragedian. The tragedies of Euripides test the Sophoclean norm in this direction. His plays present in gruelling detail the wreck of human lives under the stresses that the gods often seem willfully to place upon them. Or, if the gods are not willfully involved through jealousy or spite, they sit idly by while man wrecks himself through passion or heedlessness. No Euripidean hero approaches Oedipus in stature. The margin of freedom is narrower, and the question of justice, so central and absolute an ideal for Aeschylus, becomes a subject for irony. In *Hippolytus,* the goddess Artemis never thinks of justice as she takes revenge on the young Hippolytus for neglecting her worship; she acts solely out of personal spite. In *Medea,* Medea's revenge on Jason through the slaughter of their children is so hideously unjust as to mock the very question. In the *Bacchae,* when the frenzied Agave tears her son, Pentheus, to pieces and marches into town with his head on a pike, the god Dionysus, who had engineered the situation, says merely that Pentheus should not have scorned him. The Euripidean gods, in short, cannot be appealed to in the name of justice. Euripides' tendency toward moral neutrality, his cool tacking between sides (*e.g.,* between Pentheus versus Dionysus and the bacchantes) leave the audience virtually unable to make a moral decision. In Aeschylus' *Eumenides* (the last play of the *Oresteia*) the morals of the gods improve. Athena is there, on the stage, helping to solve the problem of justice. In Sophocles, the gods are distant, but their moral governance is not questioned. *Oedipus* ends as if with a mighty "So be it." In Euripides the gods are destructive, wreaking their capricious wills on defenseless man. Aristotle called Euripides the most tragic of the three dramatists; surely his depiction of the arena of human life is the grimmest.

Euripides' view of justice

Many qualities, however, keep his tragedies from becoming literature of protest, of cynicism, or of despair. He reveals profound psychological insight, as in the delineation of such antipodal characters as Jason and Medea, or of the forces, often subconscious, at work in the group frenzy of the *Bacchae.* His Bacchic odes reveal remarkable lyric power. And he has a deep sense of human values, however external and self-conscious. Medea, even in the fury of her hatred for Jason and her lust for revenge, must steel herself to the murder of her children, realizing the evil of what she is about to do. In this realization, Euripides suggests a saving hope: here is a great nature gone wrong—but still a great nature.

Later Greek drama. After Euripides, Greek drama reveals little that is significant to the history of tragedy. Performances were given during the remainder of the pre-Christian era in theatres throughout the Mediterranean world, but with the decline of Athens as a city-state, the tradition of tragedy eroded. As external affairs deteriorated, the high idealism, the exalted sense of human capacities depicted in tragedy at its height yielded more and more to the complaints of the skeptics. The Euripidean assault on the gods ended in the debasement of the original lofty conceptions. A 20th-century British classical scholar, Gilbert Murray, used the phrase "the failure of nerve" to describe the late Greek world. It may, indeed, provide a clue to what happened. On the other hand, according to the 19th-century German philosopher Friedrich Nietzsche, in *The Birth of Tragedy* (1872), a quite different influence may have spelled the end of Greek tragedy: the so-called Socratic optimism, the notion underlying the dialogues of Plato that man could "know himself" through the exercise of his reason in patient, careful dialectic—a notion that diverted questions of man's existence away from drama and into philosophy. In any case, the balance for tragedy was upset, and the theatre of Aeschylus, Sophocles, and Euripides gave way to what seems to have been a theatre of diatribe, spectacle, and entertainment.

The long hiatus. The Roman world did not succeed in reviving tragedy. Seneca (4 BC–AD 65) wrote nine tragedies, mostly adaptations of Greek materials, such as the stories of Oedipus, Hippolytus, Medea, and Agamemnon, but with little of the Greek tragic feeling for character and theme. The emphasis is on sensation and rhetoric, tending toward melodrama and bombast. The plays are of interest in this context mainly as the not entirely healthy inspiration for the precursors of Elizabethan tragedy in England.

The long hiatus in the history of tragedy between the Greeks and the Elizabethans has been variously explained. In the Golden Age of Roman literature, roughly from the birth of Virgil in 70 BC to the death of Ovid in AD 17, the Roman poets followed the example of Greek literature; although they produced great lyric and epic verse, their tragic drama lacked the probing freshness and directness fundamental to tragedy.

Tragedy and the Christian mass

With the collapse of the Roman world and the invasions of the barbarians came the beginnings of the long, slow development of the Christian Church. Churchmen and philosophers gradually forged a system, based on the Christian revelation, of the nature and destiny of man. The mass, with its daily reenactment of the sacrifice of Jesus Christ, its music, and its dramatic structure, may have provided something comparable to tragic drama in the lives of the people.

With the coming of the Renaissance, the visual arts more and more came to represent the afflictive aspects of life, and the word tragedy again came into currency. Chaucer (1340–1400) used the word in *Troilus and Criseyde,* and in *The Canterbury Tales* it is applied to a series of stories in the medieval style of *de casibus virorum illustrium,* meaning "the downfalls" (more or less inevitable) "of princes." Chaucer used the word to signify little more than the turn of the wheel of fortune, against whose force no meaningful effort of man is possible. It remained for the Elizabethans to develop a theatre and a dramatic literature that reinstated the term on a level comparable to that of the Greeks.

Elizabethan. The long beginning of the Elizabethan popular theatre, like that of the Greek theatre, lay in religious ceremonials, probably in the drama in the liturgy of the two greatest events in the Christian year, Christmas and Easter. In the Early Church, exchanges between two groups of choristers, or between the choir and a solo voice, led to the idea of dialogue, just as it had in the development of Greek tragedy. The parts became increasingly elaborate, and costumes were introduced to individualize the characters. Dramatic gestures and actions were a natural development. More and more of the biblical stories were dramatized, much as the material of Homer was used by the Greek tragedians, although piously in this instance, with none of the tragic skepticism of the Greeks. In the course of generations, the popularity of the performances grew to such an extent that, to accommodate the crowds, they were moved, from inside the church to the porch, or square, in front of the church. The next step was the secularization of the management of the productions, as the towns and cities took them over. Day-long festivals were instituted, involving, as in the Greek theatre, the whole community. Cycles of plays were performed at York,

Cycles of miracle and mystery plays

Chester, and other English religious centres, depicting in sequences of short dramatic episodes the whole human story, from the Fall of Lucifer and the Creation to the Day of Doom. Each play was assigned to an appropriate trade guild (the story of Noah and the Ark, for example, went to the shipwrights), which took over complete responsibility for the production. Hundreds of actors and long preparation went into the festivals. These "miracle" and "mystery" plays, however crude they may now seem, dealt with the loftiest of subjects in simple but often powerful eloquence. Although the audience must have been a motley throng, it may well have been as involved and concerned as those of the Greek theatre.

Once the drama became a part of the secular life of the communities, popular tastes affected its religious orientation. Comic scenes, like those involving Noah's nagging wife, a purely secular creation who does not appear in the Bible, became broader. The "tragic" scenes—anything involving the Devil or Doomsday—became more and more melodramatic. With the Renaissance came the rediscovery of the Greek and Roman cultures and the consequent development of a world view that led away from moral and spiritual absolutes and toward an increasingly skeptical individualism. The high poetic spirits of the mid-16th century began to turn the old medieval forms of the miracles and mysteries to new uses and to look to the ancient plays, particularly the lurid tragedies of Seneca, for their models. A bloody play, *Gorboduc,* by Thomas Sackville and Thomas Norton, first acted in 1561, is now known as the first formal tragedy in English, though it is far from fulfilling the high offices of the form in tone, characterization, and theme. Thomas Kyd's *Spanish Tragedie* (*c.* 1589) continued the Senecan tradition of the "tragedy of blood" with somewhat more sophistication than *Gorboduc* but even more bloodletting. Elizabethan tragedy never freed itself completely from certain melodramatic aspects of the influence of Seneca.

Marlowe and the first Christian tragedy. The first tragedian worthy of the tradition of the Greeks was Christopher Marlowe (1564–93). Of Marlowe's tragedies, *Tamburlaine* (1587), *Doctor Faustus* (*c.* 1588), *The Jew of Malta* (1589), and *Edward II* (*c.* 1593), the first two are the most famous and most significant. In *Tamburlaine,* the material was highly melodramatic—Tamburlaine's popular image was that of the most ruthless and bloody of conquerors. In a verse prologue, when Marlowe invites the audience to "View but his [Tamburlaine's] picture in this tragic glass," he had in mind little more, perhaps, than the trappings and tone of tragedy: "the stately tent of war," which is to be his scene, and "the high astounding terms," which will be his rhetoric. But he brought such imaginative vigour and sensitivity to bear that melodrama is transcended, in terms reminiscent of high tragedy. Tamburlaine, a Scythian shepherd of the 14th century, becomes the spokesman, curiously enough, for the new world of the Renaissance—iconoclastic, independent, stridently ambitious. Just as the Greek tragedians challenged tradition, Tamburlaine shouts defiance at all the norms, religious and moral, that Marlowe's generation inherited. But Tamburlaine, although he is an iconoclast, is also a poet. No one before him on the English stage had talked with such magnificent lyric power as he does, whether it be on the glories of conquest or on the beauties of Zenocrate, his beloved. When, still unconquered by any enemy, he sickens and dies, he leaves the feeling that something great, however ruthless, has gone. Here once again is the ambiguity that was so much a part of the Greek tragic imagination—the combination of awe, pity, and fear that Aristotle defined.

Conflict between tradition and individualism

In *Doctor Faustus* the sense of conflict between the tradition and the new Renaissance individualism is much greater. The claims of revealed Christianity are presented in the orthodox spirit of the morality and mystery plays, but Faustus' yearnings for power over space and time are also presented with a sympathy that cannot be denied. Here is modern man, tragic modern man, torn between the faith of tradition and faith in himself. Faustus takes the risk in the end and is bundled off to hell in true mystery-play fashion. But the final scene does not convey that justice has been done, even though Faustus admits that his fate is just. Rather, the scene suggests that the transcendent human individual has been caught in the consequences of a dilemma that he might have avoided but that no imaginative man *could* have avoided. The sense of the interplay of fate and freedom is not unlike that of *Oedipus.* The sense of tragic ambiguity is more poignant in Faustus than in *Oedipus* or *Tamburlaine* because Faustus is far more introspective than either of the other heroes. The conflict is inner; the battle is for Faustus' soul, a kind of conflict that neither the Greeks nor Tamburlaine had to contend with. For this reason, and not because it advocates Christian doctrine, the play has been called the first Christian tragedy.

Shakespearean tragedy. Shakespeare was a long time coming to his tragic phase, the six or seven years that produced his five greatest tragedies, *Hamlet* (*c.* 1601), *Othello* (*c.* 1602), *King Lear* (*c.* 1605), *Macbeth* (*c.* 1605), and *Antony and Cleopatra* (*c.* 1606). These were not the only plays written during those years. *Troilus and Cressida* may have come about the same time as *Hamlet; All's Well That Ends Well,* shortly after *Othello;* and *Measure for Measure,* shortly before *King Lear.* But the concentration of tragedies is sufficient to distinguish this period from that of the comedies and history plays before and of the so-called romances afterward. Although the tragic period cannot entirely be accounted for in terms of biography, social history, or current stage fashions, all of which have been adduced as causes, certain questions should be answered, at least tentatively: What is Shakespeare's major tragic theme and method? How do they relate to classical, medieval, and Renaissance traditions? In attempting to answer these questions, this proviso must be kept in mind: the degree to which he was consciously working in these traditions, consciously shaping his plays on early models, adapting Greek and Roman themes to his own purpose,

or following the precepts of Aristotle must always remain conjectural. On the one hand, there is the comment by Ben Jonson that Shakespeare had "small Latin and less Greek," and Milton in "L'Allegro" speaks of him as "fancy's child" warbling "his native wood-notes wild," as if he were unique, a sport of nature. On the other hand, Shakespeare knew Jonson (who knew a great deal of Latin and Greek) and is said to have acted in Jonson's *Sejanus* in 1603, a very classical play, published in 1605 with a learned essay on Aristotle as preface. It can be assumed that Shakespeare knew the tradition. Certainly the Elizabethan theatre could not have existed without the Greek and Roman prototype. For all of its mixed nature—with comic and melodramatic elements jostling the tragic—the Elizabethan theatre retained some of the high concern, the sense of involvement, and even the ceremonial atmosphere of the Greek theatre. When tragedies were performed, the stage was draped in black. Modern studies have shown that the Elizabethan theatre retained many ties with both the Middle Ages and the tradition of the Greeks.

Tragic elements in Shakespeare's comedies and histories

Shakespeare's earliest and most lighthearted plays reveal a sense of the individual, his innerness, his reality, his difference from every other individual, and, at times, his *plight.* Certain stock characters, to be sure, appear in the early comedies. Even Falstaff, that triumphant individual, has a prototype in the braggadocio of Roman comedy, and even Falstaff has his tragic side. As Shakespeare's art developed, his concern for the plight or predicament or dilemma seems to have grown. His earliest history plays, for instance (*Henry VI,* Parts I, II, III), are little more than chronicles of the great pageant figures—kingship in all its colour and potency. *Richard III,* which follows them, focusses with an intensity traditionally reserved for the tragic hero on one man and on the sinister forces, within and without, that bring him to destruction. From kingship, that is, Shakespeare turned to the king, the symbolic individual, the focal man, to whom whole societies look for their values and meanings. Thus Richard III is almost wholly sinister, though there exists a fascination about him, an all but tragic ambiguity.

Although Shakespeare's developing sense of the tragic cannot be summed up adequately in any formula, one might hazard the following: he progressed from the *individual* of the early comedies; to the *burdened* individual, such as, in *Henry IV,* Prince Hal, the future Henry V, who manipulates, rather than suffers, the tragic ambiguities of the world; and, finally, in the great tragedies, to (in one critic's phrase) the *overburdened* individual, Lear being generally regarded as the greatest example. In these last plays, man is at the limits of his sovereignty as a human being, where everything that he has lived by, stood for, or loved is put to the test. Like Prometheus on the crag, or Oedipus as he learns who he is, or Medea deserted by Jason, the Shakespearean tragic heroes are at the extremities of their natures. Hamlet and Macbeth are thrust to the very edge of sanity; Lear and, momentarily, Othello are thrust beyond it. In every case, as in the Greek plays, the destructive forces seem to combine inner inadequacies or evils, such as Lear's temper or Macbeth's ambition, with external pressures, such as Lear's "tiger daughters," the witches in *Macbeth,* or Lady Macbeth's importunity. Once the destructive course is set going, these forces operate with the relentlessness the Greeks called Moira, or Fate.

Total vision of Shakespeare's tragedies

At the height of his powers, Shakespeare's tragic vision comprehended the totality of possibilities for good and evil as nearly as the human imagination ever has. His heroes are the vehicles of psychological, societal, and cosmic forces that tend to ennoble and glorify humanity or infect it and destroy it. The logic of tragedy that possessed him demanded an insistence upon the latter. Initially, his heroes make free choices and are free time after time to turn back, but they move toward their doom as relentlessly as did Oedipus. The total tragic statement, however, is not limited to the fate of the hero. He is but the centre of an action that takes place in a context involving many other characters, each contributing a point of view, a set of values or antivalues to the complex dialectic of the play. In Macbeth's demon-ridden Scotland, where weird things happen to men and horses turn cannibal, there is the virtuous Malcolm, and society survives. Hamlet had the trustworthy friend Horatio, and, for all the bloodletting, what was "rotten" was purged. In the tragedies, most notably *Lear,* the Aeschylean notion of "knowledge through suffering" is powerfully dramatized; it is most obvious in the hero, but it is also shared by the society of which he is the focal figure. The flaw in the hero may be a *moral* failing or, sometimes, an excess of virtue; the flaw in society may be the rottenness of the Danish court in *Hamlet* or the corruption of the Roman world in *Antony and Cleopatra;* the flaw or fault or dislocation may be in the very universe itself, as dramatized by Lear's raving at the heavens or the ghosts that walk the plays or the witches that prophesy. All these faults, Shakespeare seems to be saying, are inevitabilities of the human condition. But they do not spell rejection, nihilsm, or despair. The hero may die, but in the words of the novelist E.M. Forster to describe the redeeming power of tragedy, "he has given us life."

Such is the precarious balance a tragedian must maintain: the cold, clear vision that sees the evil but is not maddened by it, a sense of the good that is equally clear but refuses the blandishments of optimism or sentimentalism. Few have ever sustained the balance for long. Aeschylus tended to slide off to the right, Euripides to the left, and even Sophocles had his hero transfigured at Colonus. Marlowe's early death should perhaps spare him the criticism his first plays warrant. Shakespeare's last two tragedies, *Macbeth* and *Antony and Cleopatra,* are close to the edge of a valueless void. The atmosphere of *Macbeth* is murky with evil; the action moves with almost melodramatic speed from horror to horror. The forces for good rally at last, but Macbeth himself steadily deteriorates into the most nihilistic of all Shakespeare's tragic heroes, saved in nothing except the sense of a great nature, like Medea, gone wrong. *Antony,* in its ambiguities and irony, has been considered close to the Euripidean line of bitterness and detachment. Shakespeare himself soon modulated into another mood in his last plays, *Cymbeline* (*c.* 1609), *The Winter's Tale* (*c.* 1610), and *The Tempest* (*c.* 1611). Each is based on a situation that could have been developed into major tragedy had Shakespeare followed out its logic as he had done with earlier plays. For whatever reason, however, he chose not to. The great tragic questions are not pressed. *The Tempest,* especially, for all Prospero's charm and magnanimity, gives a sense of brooding melancholy over the ineradicable evil in mankind, a patient but sad acquiescence. All of these plays end in varying degrees of harmony and reconciliation. Shakespeare willed it so.

Decline in 17th-century England. From Shakespeare's tragedies to the closing of the theatres in England by the Puritans in 1642, the quality of tragedy is steadily worse, if the best of the Greek and Shakespearean tragedies are taken as a standard. Among the leading dramatists of the period—John Webster, Thomas Middleton, Francis Beaumont, John Fletcher, Cyril Tourneur, and John Ford—there were some excellent craftsmen and brilliant poets. Though each of them has a rightful place in the history of English drama, tragedy suffered a transmutation in their hands.

The Jacobean transmutation

The Jacobean dramatists—those who flourished in England during the reign of James I—failed to transcend the negative tendencies they inherited from Elizabethan tragedy: a sense of defeat, a mood of spiritual despair implicit in Marlowe's tragic thought; in the nihilistic broodings of some of Shakespeare's characters in their worst moods—Hamlet, Gloucester in *Lear,* Macbeth; in the metaphoric implication of the theme of insanity, of man pressed beyond the limit of endurance, that runs through many of these tragedies; most importantly, perhaps, in the moral confusion ("fair is foul and foul is fair") that threatens to unbalance even the staunchest of Shakespeare's tragic heroes. This sinister tendency came to a climax about 1605 and was in part a consequence of the anxiety surrounding the death of Queen Elizabeth I and the accession of James I. Despite their negative tendencies, the Elizabethans, in general, had affirmed life and celebrated it; Shakespeare's moral balance, throughout even his darkest plays, remained firm. The Jacobeans, on the

other hand, were possessed by death. They became superb analysts of moral confusion and of the darkened vision of humanity at cross purposes, preying upon itself; of lust, hate, and intrigue engulfing what is left of beauty, love, and integrity. There is little that is redemptive or that suggests, as had Aeschylus, that evil might be resolved by the enlightenment gained from suffering. As in the tragedies of Euripides, the protagonist's margin of freedom grows ever smaller. "You are the deed's creature," cries a murderer to his unwitting lady accomplice in Middleton's *Changeling* (1622), and a prisoner of her deed she remains. Many of the plays maintained a pose of ironic, detached reportage, without the sense of sympathetic involvement that the greatest tragedians have conveyed from the beginning.

Some of the qualities of the highest tragedians have been claimed for John Webster. One critic points to his search for a moral order as a link to Shakespeare and sees in his moral vision a basis for renewal. Webster's *Duchess of Malfi* (*c.* 1613) has been interpreted as a final triumph of life over death. Overwhelmed by final unleashed terror, the Duchess affirms the essential dignity of man. Despite such vestiges of greatness, however, the trend of tragedy was downward. High moral sensitivity and steady conviction are required to resist the temptation to resolve the intolerable tensions of tragedy into either the comfort of optimism or the relaxed apathy of despair. Periods of the creation of high tragedy are therefore few and shortlived. The demands on artist and audience alike are very great. Forms wear out, and public taste seems destined to go through inevitable cycles of health and disease. What is to one generation powerful and persuasive rhetoric becomes bombast and bathos to the next. The inevitable materials of tragedy—violence, madness, hate, and lust—soon lose their symbolic role and become perverted to the uses of melodrama and sensationalism, mixed, for relief, with the broadest comedy or farce.

These corruptions had gone too far when John Milton, 29 years after the closing of the theatres, attempted to bring back the true spirit and tone of tragedy, which he called "the gravest, moralest, and most profitable of all other Poems." His *Samson Agonistes* (1671), however, is magnificent "closet tragedy"—drama more suitable for reading than for popular performance. Modelled on the *Prometheus,* it recalls Aeschylus' tragedy both in its form, in which the immobilized hero receives a sequence of visitors, and in its theme, in which there is a resurgence of the hero's spirit under stress. With Restoration comedy in full swing, however, and with the "heroic play" (an overly moralized version of tragedy) about to reach its crowning achievement in John Dryden's *All for Love* only seven years later (published 1678), *Samson Agonistes* was an anachronism.

Neoclassical. *Corneille and Racine.* Another attempt to bring back the ancient form had been going on for some time across the English Channel, in France. The French Classical tragedy, whose monuments are Pierre Corneille's *Cid* (1637) and Jean Racine's *Bérénice* (1670) and *Phèdre* (1677), made no attempt to be popular in the way of the Elizabethan theatre. The plays were written by and for intellectual aristocrats, who came together in an elite theatre, patronized by royalty and nobility. Gone were the bustle and pageantry of the Elizabethan tragedies, with their admixtures of whatever modes and moods the dramatists thought would work. The French playwrights submitted themselves to the severe discipline they derived from the Greek models and especially the "rules," as they interpreted them, laid down by Aristotle. The unities of place, time, and action were strictly observed. One theme, the conflict between Passion and Reason, was uppermost. The path of Reason was the path of Duty and Obligation (noblesse oblige), and that path had been clearly plotted by moralists and philosophers, both ancient and modern. In this sense there was nothing exploratory in the French tragedy; existing moral and spiritual norms were insisted upon. The norms are never criticized or tested as Aeschylus challenged the Olympians or as Marlowe presented, with startling sympathy, the Renaissance overreacher. Corneille's *Cid* shows Duty triumphant over Passion, and, as a reward, hero and heroine are happily united. By the time of *Phèdre,* Corneille's proud affirmation of the power of the will and the reason over passion had given way to what Racine called "stately sorrow," with which he asks the audience to contemplate Phèdre's heroic, but losing, moral struggle. Her passion for her stepson, Hippolyte, bears her down relentlessly. Her fine principles and heroic will are of no avail. Both she and Hippolyte are destroyed. The action is limited to one terrible day; there is no change of scene; there is neither comic digression nor relief—the focus on the process by which a great nature goes down is sharp and intense. Such is the power of Racine's poetry (it is untranslatable), his conception of character, and his penetrating analysis of it, that it suggests the presence of Sophoclean "heroic humanism." In this sense it could be said that Racine tested the norms, that he uncovered a cruel injustice in the nature of a code that could destroy such a person as Phèdre. Once again, here is a world of tragic ambiguity, in which no precept or prescription can answer complicated human questions.

The return to Greek models

The English "heroic play." This ambiguity was all but eliminated in the "heroic play" that vied with the comedy of the Restoration stage in England in the latter part of the 17th century. After the vicissitudes of the Civil War, the age was hungry for heroism. An English philosopher of the time, Thomas Hobbes, defined the purpose of the type: "The work of an heroic poem is to raise admiration, principally for three virtues, valor, beauty, and love." Moral concern, beginning with Aeschylus, has always been central in tragedy, but in the works of the great tragedians this concern was exploratory and inductive. The moral concern of the heroic play is the reverse. It is deductive and dogmatic. The first rule, writes Dryden (following the contemporary French critic, René Le Bossu) in his preface to his *Troilus and Cressida* (1679), is "to make the moral of the work; that is, to lay down to yourself what that precept of morality shall be, which you would insinuate into the people . . . ." In *All for Love* the moral is all too clear: Antony must choose between the path of honour and his illicit passion for Cleopatra. He chooses Cleopatra, and they are both destroyed. Only Dryden's poetry, with its air of emotional argumentation, manages to convey human complexities in spite of his moral bias and saves the play from artificiality—makes it, in fact, the finest near-tragic production of its age.

Inductive and deductive morality

The eclipse of tragedy. Although the annals of the drama from Dryden onward are filled with plays called tragedies by their authors, the form as it has been defined here went into an eclipse during the late 17th, the 18th, and the early 19th centuries. Reasons that have been suggested for the decline include the politics of the Restoration in England; the rise of science and, with it, the optimism of the Enlightenment throughout Europe; the developing middle class economy; the trend toward reassuring deism in theology; and, in literature, the rise of the novel and the vogue of satire. The genius of the age was discursive and rationalistic. In France and later in England, belief in Evil was reduced to the perception of evils, which were looked upon as institutional and therefore remediable. The nature of man was no longer the problem; rather, it was the better organization and management of men. The old haunting fear and mystery, the sense of ambiguity at the centre of man's nature and of dark forces working against him in the universe, were replaced by a new and confident dogma. Tragedy never lost its high prestige in the minds of the leading spirits. Theorizing upon it were men of letters as diverse as Dr. Samuel Johnson, David Hume, Samuel Taylor Coleridge, and Percy Bysshe Shelley and German philosophers from Gotthold Lessing in the 18th century to Friedrich Nietzsche in the 19th. Revivals of Shakespeare's tragedies were often bowdlerized or altered, as in the happy ending for *Lear* in a production of 1681. Those who felt themselves called upon to write tragedies produced little but weak imitations. Shelley tried it once, in *The Cenci* (1819), but, as his wife wrote, "the bent of his mind went the other way"—which way may be seen in his *Prometheus Unbound* (1820), in which Zeus is overthrown and man enters upon a golden age, ruled by the power of love. Goethe had the sense to stay away from tragedy: "The mere attempt to write tragedy," he

said, "might be my undoing." He concluded his two-part *Faust* (1808, 1832) in the spirit of the 19th-century optimistic humanitarianism. It was not until the latter part of the 19th century, with the plays of a Norwegian, Henrik Ibsen, a Russian, Anton Chekhov, a Swede, August Strindberg, and, later, an American, Eugene O'Neill, that something of the original vision returned to inspire the tragic theatre.

A new vehicle: the novel. The theme and spirit of tragedy, meanwhile, found a new vehicle in the novel. This development is important, however far afield it may seem from the work of the formal dramatists. The English novelist Emily Brontë's *Wuthering Heights* (1847), in its grim Yorkshire setting, reflects the original concerns of tragedy: *i.e.,* the terrifying divisions in nature and human nature, love that creates and destroys, character at once fierce and pitiable, destructive actions that are willed yet seemingly destined, as if by a malicious fate, yet the whole controlled by an imagination that learns as it goes. Another English novelist, Thomas Hardy, in the preface to his *Woodlanders* (1887), speaks of the rural setting of this and other of his novels as being comparable to the stark and simple setting of the Greek theatre, giving his novels something of that drama's intensity and sharpness of focus. His grimly pessimistic view of man's nature and destiny and of the futility of human striving, as reflected in his novels *The Return of the Native* (1878), *Tess of the D'Urbervilles* (1891), and *Jude the Obscure* (1895), is barely redeemed for tragedy by his sense of the beauty of nature and of the beauty and dignity of human character and effort, however unavailing.

Thomas Hardy's pessimism

The work of the Polish-born English novelist Joseph Conrad (1857–1924) provides another kind of setting for novels used as vehicles of the tragic sense. *Lord Jim* (1900), originally conceived as a short story, grew to a full-length novel as Conrad found himself exploring in ever greater depth the perplexing, ambiguous problem of lost honour and guilt, expiation and heroism. Darkness and doubt brood over the tale, as they do over his long story "Heart of Darkness" (1899), in which Conrad's narrator, Marlow, again leads his listeners into the shadowy recesses of the human heart, with its forever unresolved and unpredictable capacities for good and evil.

Dostoyevsky's tragic view. In Russia, the novels of Fyodor Dostoyevsky, particularly *Crime and Punishment* (1866) and *The Brothers Karamazov* (1880), revealed a world of paradox, alienation, and loss of identity, prophetic of the major tragic themes of the 20th century. More than any earlier novelist, Dostoyevsky appropriated to his fictions the realm of the subconscious and explored in depth its shocking antinomies and discontinuities. Sigmund Freud, the founder of psychoanalysis, frequently acknowledged his indebtedness to Dostoyevsky's psychological insights. Dostoyevsky's protagonists are reminiscent of Marlowe's Doctor Faustus, caught between the old world of orthodox belief and the new world of intense individualism, each with its insistent claims and justifications. The battleground is once more the soul of man, and the stakes are survival. Each of his major heroes—Raskolnikov in *Crime and Punishment* and the three Karamazovs—wins a victory, but it is in each case morally qualified, partial, or transient. The harmonious resolutions of the novels seem forced and are neither decisive of the action nor definitive of Dostoyevsky's total tragic view.

Paradox, alienation, loss of identity in Dostoyevsky

The American tragic novel. In America, Nathaniel Hawthorne's novel *The Scarlet Letter* (1850) and Herman Melville's *Moby Dick* (1851) are surprisingly complete embodiments of the tragic form, written as they were at a time of booming American optimism, materialistic expansion, and sentimentalism in fiction—and no tragic theatre whatever. In *The Scarlet Letter,* a story of adultery set in colonial New England, the heroine's sense of sin is incomplete; her spirited individualism insists (as she tells her lover) that "what we did had a consecration of its own." The resulting conflict in her heart and mind is never resolved, and, although it does not destroy her, she lives out her life in gray and tragic isolation. Melville said that he was encouraged by Hawthorne's exploration of "a certain tragic phase of humanity," by his deep broodings and by the "blackness of darkness" in him, to proceed with similar explorations of his own in *Moby Dick,* which he dedicated to Hawthorne. Its protagonist, Captain Ahab, represents a return to what Melville called (defending Ahab's status as tragic hero) a "mighty pageant creature, formed for noble tragedies," whose "ponderous heart," "globular brain," and "nervous lofty language" prove that even an old Nantucket sea captain can take his place with kings and princes of the ancient drama. Shakespearean echoes abound in the novel; some of its chapters are written in dramatic form. Its theme and central figure, reminiscent of Job and Lear in their search for justice and of Oedipus in his search for the truth, all show what Melville might have been—a great tragic dramatist had there been a tragic theatre in America.

Some American novelists of the 20th century carried on, however partially, the tragic tradition. Theodore Dreiser's *American Tragedy* (1925) is typical of the naturalistic novel, which is also represented by the work of Stephen Crane, James T. Farrell, John Steinbeck. Though showing great sensitivity to environmental or sociological evils, such works fail to embody the high conception of character (as Melville describes it above) and are concerned mainly with externals, or reportage. The protagonists are generally "good" (or weak) and beaten down by society. The novels of Henry James, which span the period from 1876 to 1904, are concerned with what has been called the tragedy of manners. The society James projects is sophisticated, subtle, and sinister. The innocent and the good are destroyed, like Milly Theale in *The Wings of the Dove* (1902), who in the end "turns her face to the wall" and dies but in her death brings new vision and new values to those whose betrayals had driven her to her death.

The trend in American fiction, as in the drama, continued in the 20th century, toward the pathos of the victim—the somehow inadequate, the sometimes insignificant figure destroyed by such vastly unequal forces that the struggle is scarcely significant. F. Scott Fitzgerald's Gatsby in his novel *The Great Gatsby* (1925) is betrayed by his own meretricious dream, nurtured by a meretricious society. The hero of Ernest Hemingway's novel *A Farewell to Arms* (1929), disillusioned by war, makes a separate peace, deserts, and joins his beloved in neutral Switzerland. When she dies in childbirth, he sees it as still another example of how "they"—society, the politicians who run the war, or the mysterious forces destroying Catherine—get you in the end. The tone is lyric and pathetic rather than tragic (though Hemingway called the novel his *Romeo and Juliet*). Grief turns the hero away from, rather than toward, a deeper examination of life.

Only the novels of William Faulkner, in their range and depth and in their powerful assault on the basic tragic themes, recall unmistakably the values of the tragic tradition. His "saga of the South," as recounted in a series of novels (notably *Sartoris,* 1929; *The Sound and the Fury,* 1929; *As I Lay Dying,* 1930; *Sanctuary,* 1931; *Light in August,* 1932; *Absalom, Absalom!* 1936; *Intruder in the Dust,* 1948; *Requiem for a Nun,* 1951), incorporates some 300 years of Southern history from Indian days to the present. At first regarded as a mere exploiter of decadence, he can now be seen as gradually working beyond reportage and toward meaning. His sociology became more and more the "sin" of the South—the rape of the land, slavery, the catastrophe of the Civil War and its legacy of a cynical and devitalized materialism. Increasingly he saw the conflict as internal. The subject of art, Faulkner said in his 1949 Nobel Prize speech, is "the human heart in conflict with itself." His insistence is on guilt as the evidence of man's fate, and on the possibility of expiation as the assertion of man's freedom. Compassion, endurance, and the capacity to learn are seen to be increasingly effective in his characters. In the veiled analogies to Christ as outcast and redeemer in *Light in August* and in the more explicit Christology of *A Fable* (1954), in the pastoral serenity following the anguish and horror in *Light in August,* and in the high comedy of the last scene of *Intruder in the Dust,* Faulkner puts into tragic fiction the belief he stated in his Nobel speech: "I decline to accept the end of man."

Guilt and expiation in Faulkner

TRAGEDY AND MODERN DRAMA

Tragic themes in Ibsen, Strindberg, and Chekhov. The movement toward naturalism in fiction in the latter decades of the 19th century did much to purge both the novel and the drama of the sentimentality and evasiveness that had so long emasculated them. In Norway Henrik Ibsen incorporated in his plays the smug and narrow ambitiousness of his society. The hypocrisy of overbearing men and women replace, in their fashion, the higher powers of the old tragedy. His major tragic theme is the futility, leading to catastrophe, of the idealist's effort to create a new and better social order. The "Problem play"—one devoted to a particular social issue—is saved in his hand from the flatness of a sociological treatise by a sense of doom, a pattern of retribution, reminiscent of the ancient Greeks. In *Pillars of Society* (1877), *The Wild Duck* (published 1884), *Rosmersholm* (published 1886), and *The Master Builder* (published 1892), for example, one sacrifice is expiated by another.

In Sweden, August Strindberg, influenced by Ibsen, was a powerful force in the movement. *The Father* (1887) and *Miss Julie* (published 1888) recall Ibsen's attacks on religious, moral, and political orthodoxies. Strindberg's main concern, however, is with the destructive effects of sexual maladjustment and psychic imbalance. Not since Euripides' *Medea* or Racine's *Phèdre* had the tragic aspects of sex come under such powerful analysis. In this respect, his plays look forward to O'Neill's.

Anton Chekhov, the most prominent Russian dramatist of the period, wrote plays about the humdrum life of inconspicuous, sensitive people (*Uncle Vanya,* 1899; *The Three Sisters,* 1901; and *The Cherry Orchard,* 1904, are typical), whose lives fall prey to the hollowness and tedium of a disintegrating social order. They are a brood of lesser Hamlets without his compensating vision of a potential greatness. As in the plays of the Scandinavian dramatists, Chekhov's vision of this social evil is penetrating and acute, but the powerful, resistant counterthrust that makes for tragedy is lacking. It is a world of victims.

American tragic dramatists. In little of the formal drama between the time of Ibsen, Strindberg, and Chekhov and the present are the full dimensions of tragedy presented. Some critics suggested that it was too late for tragedy, that modern man no longer valued himself highly enough, that too many sociological and ideological factors were working against the tragic temperament. The long and successful career of Eugene O'Neill may be a partial answer to this criticism. He has been called the first American to succeed in writing tragedy for the theatre, a fulfillment of his avowed purpose, for he had declared that in the tragic, alone, lay the meaning of life—and the hope. He sought in Freud's concept of the subconscious the equivalent of the Greek idea of fate and modelled his great trilogy, *Mourning Becomes Electra* (1931), on Aeschylus' *Oresteia.* Although the hovering sense of an ancient evil is powerful, the psychological conditioning controls the characters too nakedly. They themselves declare forces that determine their behaviour, so that they seem almost to connive in their own manipulation. *Desire Under the Elms* (1924) presents a harsh analysis of decadence in the sexual and avaricious intrigues of a New England farmer's family, unrelieved by manifestations of the transcendent human spirit. *The Great God Brown* (1926) and *Long Day's Journey into Night* (1939–41; first performance, 1956) come closer to true tragedy. In the latter, the capacity for self-knowledge is demonstrated by each member of the wrangling Tyrone family (actually, O'Neill's own; the play is frankly autobiographical). The insistent theme of the "death wish" (another example of Freud's influence), however, indicates too radical a pessimism for tragedy; even the character of Edmund Tyrone, O'Neill's own counterpart, confesses that he has always been a little in love with death, and in another late play, *The Iceman Cometh* (1939), the death wish is more strongly expressed. Although he never succeeded in establishing a tragic theatre comparable to the great theatres of the past, O'Neill made a significant contribution in his sustained concentration on subjects at least worthy of such a theatre. He made possible the significant, if slighter, contributions of Arthur Miller, whose *Death of a Salesman* (1949) and *A View from the Bridge* (1955) contain material of tragic potential that is not fully realized. Tennessee Williams' *Streetcar Named Desire* (1947) is a sensitive study of the breakdown of a character under social and psychological stress. As with Miller's plays, however, it remains in the area of pathos rather than tragedy.

O'Neill: pessimism too radical for tragedy

Other serious drama. The 20th century has produced much serious and excellent drama, which, though not in the main line of the tragic tradition, deserves mention. In British theatre, George Bernard Shaw's *Saint Joan* (1923) and T.S. Eliot's *Murder in the Cathedral* (1935) dramatized with great power both doubt and affirmation, the ambiguity of human motives, and the possibility of fruitless suffering that are true of the human condition as reflected by tragedy. During the Irish literary revival, the work of J.M. Synge (*Riders to the Sea,* 1904) and Sean O'Casey (*Shadow of a Gunman,* 1923), like Faulkner's work, sought a tragic theme in the destiny of a whole people. The masterpiece of this movement, however, is not a tragedy but a comic inversion of the ancient tragedy of *Oedipus*—Synge's *Playboy of the Western World* (1907).

Tragedy in works of social protest

The drama of social protest—exemplified in such works as the Russian Maksim Gorky's *Lower Depths* (1902), the German Bertolt Brecht's *Threepenny Opera* (1928) and *Mother Courage* (1941), and the American Clifford Odets' *Waiting for Lefty* (1935)—shares the tragedians' concern for evils that frustrate or destroy human values. The evils, however, are largely external, identifiable, and, with certain recommended changes in the social order, remediable. The type shows how vulnerable tragedy is to dogma or programs of any sort. A British author, George Orwell, suggested in *Nineteen Eighty-four* that tragedy would cease to exist under pure Marxist statism. Brecht's fine sense of irony and moral paradox redeem him from absolute dogmatism but give his work a hard satiric thrust that is inimical to tragedy. Traditional values and moral imperatives are all but neutralized in the existentialist worlds of the dramas and novels of Jean-Paul Sartre and Albert Camus, two outstanding philosopher-dramatists of the post World War II era. In their works, the protagonist is called upon to forge his own values, if he can, in a world in which the disparity between the ideal (what man longs for) and the real (what he gets) is so great as to reduce the human condition to incoherence and absurdity. Plays that led to the coinage of the term the theatre of the absurd are exemplified by *Waiting for Godot* (1952) and *The Killer* (1959), respectively by the Irish writer Samuel Beckett and the Romanian Eugène Ionesco, both of whom pursued their careers in Paris. Here, the theme of victimization is at its extreme, the despair and defeat almost absolute.

A coherent and affirmative view of man, society, and the cosmos is vital to tragedy—however tentative the affirmation may be. Unresolved questions remain at the end of every tragedy. There is always an irrational factor, disturbing, foreboding, not to be resolved by the sometime consolations of philosophy and religion or by any science of the mind or body; there is irretrievable loss, usually though not necessarily symbolized by the death of the hero. In the course of the action, however, in the development of character, theme, and situation and in the conceptual suggestiveness of language, tragedy presents the positive terms in which these questions might be answered. The human qualities are manifest, however limited; man's freedom is real, however marginal. The forces that bear him down may be mysterious but actual—fate, the gods, chance, the power of his own or the race's past working through his soul. Though never mastered, they can be contended with, defied, and, at least in spirit, transcended. The process is cognitive; man can learn.

Absence of tragedy in Oriental drama. In no way can the importance of a conceptual basis for tragedy be better illustrated than by a look at other drama-producing cultures with radically different ideas of the individual, his nature, and his destiny. While the cultures of India, China, and Japan have produced significant and highly artistic drama, there is little here to compare in magnitude, intensity, and freedom of form to the tragedies of the West.

In Buddhist teaching, the aim of the individual is to sup-

press and regulate all those questioning, recalcitrant, rebellious impulses that first impel the Western hero toward his tragic course. The goal of Nirvāṇa is the extinction of those impulses, the quieting of the passions, a kind of *quietus* in which worldly existence ceases. Western tragedy celebrates life, and the tragic hero clings to it: to him, it is never "sweet to die" for his country or for anything else, and the fascination for Western audiences is to follow the hero—as it were, *from the inside*—as he struggles to assert himself and his values against whatever would deny them. In Oriental drama, there is no such intense focus on the individual. In the Japanese Nō plays, for instance, the hero may be seen in moments of weariness and despair, of anger or confusion, but the mood is lyric, and the structure of the plays is ritualistic, with a great deal of choral intoning, dancing, and stylized action. Although a number of Nō plays can be produced together to fill a day's performance, the individual plays are very short, hardly the length of a Western one-act play. Nō plays affirm orthodoxy, rather than probing and questioning it, as Western tragedies do.

The drama in India has a long history, but there too the individual is subordinated to the mood of the idyll or romance or epic adventure. Perhaps one reason why the drama of India never developed the tragic orientation of the West is its removal from the people; it has never known the communal involvement of the Greek and Elizabethan theatres. Produced mainly for court audiences, an upper class elite, it never reflected the sufferings of common (or uncommon) humanity. Only recently has the drama in China embraced the vigour and realism of the common people, but the drama is in the service not of the individual but of a political ideology, which replaces the traditional themes of ancestor worship and filial piety. In all this, the mighty pageant figure—Oedipus, Prometheus, Lear, or Ahab standing for the individual as he alone sees and feels the workings of an unjust universe—is absent.

Tragedy in Nō drama

An example from the Nō plays will illustrate these generalizations. In *The Hoka Priests,* by Zenchiku Ujinobu (1414–99), a son is confronted with Hamlet's problem—*i.e.,* that of avenging the death of his father. He is uncertain how to proceed, since his father's murderer has many bold fellows to stand by him, while he is all alone. He persuades his brother, a priest, to help him, and disguising themselves as priests, they concoct a little plot to engage the murderer in religious conversation. There are a few words of lament—"Oh why,/ Why back to the bitter World/ Are we borne by our intent?"—and the Chorus sings lyrically about the uncertainties of life. The theme of the conversation is the unreality of the World and the reality of Thought. At an appropriate moment, the brothers cry, "Enough! Why longer hide our plot?" The murderer places his hat on the floor and exits. The brothers mime the killing of the murderer in a stylized attack upon the hat, while the Chorus describes and comments on the action: "So when the hour was come/ Did these two brothers/ By sudden resolution/ Destroy their father's foe./ For valour and piety are their names remembered/ Even in this aftertime" (translated by Arthur Waley, *The Nō Plays of Japan,* 1921).

Thus the Nō avoids directly involving the audience in the emotions implicit in the events portrayed on the stage. It gives only a slight hint of the spiritual struggle in the heart of the protagonist—a struggle that is always speedily resolved in favour of traditional teaching. In play after play the action does not take place before our eyes but is reenacted by the ghost of one of the participants. Thus, the events presented are tinged with memory or longing—hardly the primary emotions that surge through and invigorate Western tragedy at its best.

Loss of viability in the West. The absence, even in the West, of a great tragic theatre in the 20th century may be explained by the pantheon of panaceas to which modern man has subscribed. Politics, psychology, social sciences, physical sciences, nationalism, the occult—each offered a context in terms of which he might act out his destiny, were it not crowded out by the others. Modern man is not tested but harried and not by gods but, too often, by demons. In the dramas of Athens and England, tragedy was born of the impossibility of a clear-cut victory in man's struggle with powers greater than himself. In the modern drama, the struggle itself seems impossible.

The would-be hero is saved from a meaningful death by being condemned to a meaningless life. This, too, however, has its tragic dimension, in its illustration of the power of evil to survive from millennium to millennium in the presence or the absence of the gods.

Tragedy is a means of coming to terms with that evil. To assume that tragedy has lost viability is to forget that this viability was seriously questioned by the first Western philosopher to address himself to the problem. An account of the development of the theory of tragedy will reveal a resourcefulness in man's critical powers that can help to compensate, or occasionally even supersede, his lapsing creative powers. (R.B.S.)

THEORY OF TRAGEDY

Classical theories. As the great period of Athenian drama drew to an end at the beginning of the 4th century BC, Athenian philosophers began to analyze its content and formulate its structure. In the thought of Plato (*c.* 427–347 BC), the history of the criticism of tragedy began with speculation on the role of censorship. To Plato (in the dialogue on the *Laws*) the state was the noblest work of art, a representation (*mimēsis*) of the fairest and best life. He feared the tragedians' command of the expressive resources of language, which might be used to the detriment of worthwhile institutions. He feared, too, the emotive effect of poetry, the Dionysian element that is at the very basis of tragedy. Therefore, he recommended that the tragedians submit their works to the rulers, for approval, without which they could not be performed. It is clear that tragedy, by nature exploratory, critical, independent, could not live under such a regimen.

Ethics and art—Plato and Aristotle

Plato is answered, in effect and perhaps intentionally, by Aristotle's *Poetics.* Aristotle (384–322 BC) defends the purgative power of tragedy and, in direct contradiction to Plato, makes moral ambiguity the essence of tragedy. The tragic hero must be neither a villain nor a virtuous man but a "character between these two extremes, . . . a man who is not eminently good and just, yet whose misfortune is brought about not by vice or depravity, but by some error or frailty [*hamartia*]." The effect on the audience will be similarly ambiguous. A perfect tragedy, he says, should imitate actions that excite "pity and fear." He uses Sophocles' *Oedipus the King* as a paradigm. Near the beginning of the play, Oedipus asks how his stricken city (the counterpart of Plato's state) may cleanse itself, and the world he uses for the purifying action is a form of the word catharsis. The concept of catharsis provides Aristotle with his reconciliation with Plato, a means by which to satisfy the claims of both ethics and art. "Tragedy," says Aristotle, "is an imitation [*mimēsis*] of an action that is serious, complete and of a certain magnitude . . . through pity and fear effecting the proper purgation [catharsis] of these emotions." Ambiguous means may be employed, Aristotle maintains in contrast to Plato, to a virtuous and purifying end.

To establish the basis for a reconciliation between ethical and artistic demands, Aristotle insists that the principal element in the structure of tragedy is not character but plot. Since the erring protagonist is always in at least partial opposition to the state, the importance of tragedy lies not in him but in the enlightening event. "Most important of all," Aristotle said, "is the structure of the incidents. For tragedy is an imitation not of men but of an action and of life, and life consists in action, and its end is a mode of action, not a quality . . . ." Aristotle considered the plot to be the soul of a tragedy, with character in second place. The goal of tragedy is not suffering but the knowledge that issues from it, as the denouement issues from a plot. The most powerful elements of emotional interest in tragedy, according to Aristotle, are reversal of intention or situation (*peripeteia*) and recognition scenes (*anagnōrisis*), and each is most effective when it is coincident with the other. In *Oedipus,* for example, the messenger who brings Oedipus news of his real parentage, intending to allay his fears, brings about a sudden reversal of his fortune, from

happiness to misery, by compelling him to recognize that his wife is also his mother.

Later critics found justification for their own predilections in the authority of Greek drama and Aristotle. For example, the Roman poet Horace (65–8 BC), in his *Ars poetica* (*Art of Poetry*), elaborated the Greek tradition of extensively narrating offstage events into a dictum on decorum forbidding events such as Medea's butchering of her boys from being performed on stage. And where Aristotle had discussed tragedy as a separate genre, superior to epic poetry, Horace discussed it as a genre with a separate style, again with considerations of decorum foremost. A theme for comedy may not be set forth in verses of tragedy; each style must keep to the place alloted it.

On the basis of this kind of stylistic distinction, the *Aeneid,* the epic poem of Virgil, Horace's contemporary, is called a tragedy by the fictional Virgil in Dante's *Divine Comedy,* on the grounds that the *Aeneid* treats only of lofty things. Dante (1265–1321) calls his own poem a comedy partly because he includes "low" subjects in it. He makes this distinction in his *De vulgari eloquentia* (1304–05; "Of Eloquence in the Vulgar") in which he also declares the subjects fit for the high, tragic style to be salvation, love, and virtue. Despite the presence of these subjects in this poem, he calls it a comedy because his style of language is "careless and humble" and because it is in the vernacular tongue rather than Latin. Dante makes a further distinction:

> Comedy . . . differs from tragedy in its subject matter, in this way, that tragedy in its beginning is admirable and quiet, in its ending or catastrophe fould and horrible . . . From this it is evident why the present work is called a comedy.

Fate and will in comedy and tragedy

Dante's emphasis on the outcome of the struggle rather than on the nature of the struggle is repeated by Chaucer and for the same reason: their belief in the providential nature of human destiny. Like Dante, he was under the influence of *De consolatione philosophiae* (*Consolation of Philosophy*), the work of the 6th-century Roman philosopher Boethius (*c.* 480–524) that he translated into English. Chaucer considered Fortune to be beyond the influence of the human will. In his *Canterbury Tales,* he introduces "The Monk's Tale" by defining tragedy as "a certeyn storie . . . / of him that stood in greet prosperitee,/ And is y-fallen out of heigh degree/ Into miserie, and endeth wrecchedly." Again, he calls his *Troilus and Criseyde* a tragedy because, in the words of Troilus, "all that comth, comth by necessitee . . . / That forsight of divine purveyaunce/ Hath seyn alwey me to forgon Criseyde."

Elizabethan approaches. The critical tradition of separating the tragic and comic styles is continued by the Elizabethan English poet Sir Philip Sidney, whose *Defence of Poesie* (also published as *An Apologie for Poetrie*) has the distinction of containing the most extended statement on tragedy in the English Renaissance and the misfortune of having been written in the early 1580s (published 1595), before the first plays of Shakespeare, or even of Marlowe. Nevertheless, Sidney wrote eloquently of "high and excellent tragedy, that . . . with stirring the affects of admiration and commiseration teacheth the uncertainty of this world and upon how weak foundations gilden roofs are builded."

Since the word admiration here means awe, Sidney's "admiration and commiseration" are similar to Aristotle's "pity and fear." He differs from Aristotle, however, in preferring epic to tragic poetry. The Renaissance was almost as concerned as Plato with the need to justify poetry on ethical grounds, and Sidney ranks epic higher than tragedy because it provides morally superior models of behaviour.

Sidney goes further than mere agreement with Aristotle, however, in championing the unities of time and place. Aristotle had asserted the need for a unity of time: "Tragedy endeavors, as far as possible, to confine itself to a single revolution of the sun, or but slightly to exceed this limit." Sidney, following the lead of a 16th-century Italian Neoclassicist, Ludovico Castelvetro, added the unity of place: "the stage should always represent but one place, and the uttermost time presupposed in it should be, both by Aristotle's precept and common reason, but one day . . ." Sidney also seconds Horace's disapproval of the mingling of styles, which Sidney says produces a "mongrel tragicomedy."

Shakespeare's opinion of the relative merits of the genres is unknown, but his opinion of the problem itself may be surmised. In *Hamlet* he puts these words in the mouth of the foolish old pedant Polonius: "The best actors in the world, either for tragedy, comedy, history, pastoral, pastoral-comical, historical-pastoral, tragical-historical, tragical-comical-historical-pastoral; scene individable, or poem unlimited . . ." (Act II, scene 2). As to the classical unities, Shakespeare adheres to them only twice and neither time in a tragedy, in *The Comedy of Errors* and *The Tempest.* And through the mouths of his characters, Shakespeare, like Aristotle, puts himself on both sides of the central question of tragic destiny—that of freedom and necessity. Aristotle says that a tragic destiny is precipitated by the hero's tragic fault, his "error or frailty" (hamartia), but Aristotle also calls this turn of events a change of "fortune." Shakespeare's Cassius in *Julius Caesar* says, "The fault, dear Brutus, is not in our stars, / But in ourselves . . . ," and in King Lear, Edmund ridicules a belief in fortune as the "foppery of the world." But Hamlet, in a comment on the nature of hamartia, is a fatalist when he broods on the "mole of nature," the "one defect" that some men are born with, "wherein they are not guilty," and that brings them to disaster (Act I, scene 4). Similarly, Sophocles' Oedipus, though he says, "It was Apollo who brought my woes to pass," immediately adds, "it was my hand that struck my eyes." These ambiguities are a powerful source of the tragic emotion of Athenian and Elizabethan drama, unequalled by traditions that are more sure of themselves, such as French Neoclassicism, or less sure of themselves, such as 20th-century drama.

Neoclassical theory. In the Neoclassical period Aristotle's reasonableness was replaced by rationality, and his moral ambiguity by the mechanics of "poetic justice." In the 17th century, under the guise of a strict adherence to Classical formulas, additional influences were brought to bear on the theory of tragedy. In France, the theological doctrine of Jansenism, which called for an extreme orthodoxy, exercised a strong influence. In England, the restoration of the monarchy in 1660, with the reopening of the theatres, introduced a period of witty and lusty literature. In both nations, the influence of natural law—the idea that laws binding upon humanity are inferable from nature—increased, along with the influence of the exact sciences. Critics in both nations declared that Aristotle's "rules" were made to reduce nature into a method.

In his 1679 preface to Shakespeare's *Troilus and Cressida,* Dryden says, "we lament not, but detest a wicked man, we are glad when we behold his crimes are punished, and that Poetical justice is done upon him." Similar sentiments, calling for the punishment of crimes and the reward of virtue, were expressed in France. Catharsis had become vindication. Thomas Rymer, one of the most influential English critics of the time, in *The Tragedies of The Last Age* (1678), wrote that

Vindication rather than catharsis

> besides the purging of the passions, something must stick by observing . . . that necessary relation and chain, whereby the causes and the effects, the vertues and rewards, the vices and their punishments are proportion'd and link'd together, how deep and dark soever are laid the Springs, and however intricate and involv'd are their operations.

The effect was to rob tragedy of a great deal of its darkness and depth. The temper of the age demanded that mystery be brought to the surface and to the light, a process that had effects not merely different from but in part antipathetic to tragedy. Nicolas Boileau, the chief spokesman of the French Neoclassical movement, in his discussion of pity and fear in *Art Poétique* (1674), qualified these terms with the adjectives "beguiling" and "pleasant" (*pitié charmante, douce terreur*), which radically changed their meaning. The purged spectator became a grateful patient.

In his preface to *Phèdre* (1677), Racine subscribed to the *quid pro quo* view of retribution.

> I have written no play in which virtue has been more celebrated than in this one. The smallest faults are here severely punished; the mere idea of a crime is looked upon with as much horror as the crime itself.

Of Phèdre herself, his greatest heroine, he says,

> I have taken the trouble to make her a little less hateful than she is in the ancient versions of this tragedy, in which she herself resolves to accuse Hippolytus. I judged that that calumny had about it something too base and black to be put into the mouth of a Princess . . . This depravity seemed to me more appropriate to the character of a nurse, whose inclinations might be supposed to be more servile . . . .

For Aristotle, pity and fear made a counterpoint typical of Classicism, each tempering the other to create a balance. For Racine, pity and fear each must be tempered in itself. In the marginalia to his fragmentary translation of Aristotle's *Poetics,* Racine wrote that in arousing the passions of pity and fear, tragedy

> removes from them whatever they have of the excessive and the vicious and brings them back to a moderated condition and conformable to reason.

Corneille contradicted Aristotle outright. Discussing *Le Cid* he said, in *A Discourse on Tragedy* (1660),

> Our pity ought to give us fear of falling into similar misfortune, and purge us of that excess of love which is the cause of their disaster . . . but I do not know that it gives us that, or purges us, and I am afraid that the reasoning of Aristotle on this point is but a pretty idea . . . it is not requisite that these two passions always serve together . . . it suffices . . . that one of the two bring about the purgation. . . .

The accommodation of tragedy to Neoclassical ideas of order demanded a simplification of tragedy's complexities and ambiguities. The simplifying process was now inspired, however, by the fundamental tenet of all primitive scientific thought namely, that orderliness and naturalness are in a directly proportionate relationship. Racine declared the basis of the naturalistic effect in drama to be a strict adherence to the unities, which now seem the opposite of naturalistic. In his preface to *Bérénice* (1670), he asked what probability there could be when a multitude of things that would scarcely happen in several weeks are made to happen in a day. The illusion of probability, which is the Aristotelian criterion for the verisimilitude of a stage occurrence, is made to sound as if it were the result of a strict dramaturgical determinism, on the grounds that necessity is the truest path to freedom.

Racine and Corneille both contradicted Dante and Chaucer on the indispensability of a catastrophic final scene. "Blood and deaths," said Racine, are not necessary, for "it is enough that the action be grand, that the actors be heroic, that the passions be aroused" to produce "that stately sorrow that makes the whole pleasure of tragedy" (preface to *Bérénice*).

Milton was artistically much more conservative. He prefaced his *Samson Agonistes* (1671) with a warning against the

> error of intermixing Comic stuff with Tragic sadness and gravity; or introducing trivial and vulgar persons: which by all judicious hath been counted absurd; and brought in without discretion, corruptly to gratify the people.

He bypassed Shakespeare for the ancients and ranked Aeschylus, Sophocles, and Euripides as tragic poets unequalled yet by any others. Part of the rule, for Milton, was that which affirmed the unities. In his concurrence with the Classical idea of the purgative effect of pity and fear, Milton combined reactionary aesthetics with the scientific spirit of the recently formed Royal Society.

> Nor is Nature wanting in her own effects to make good his assertion [Aristotle on catharsis]: for so, in Physic things of melancholic hue and quality are used against melancholy, sour against sour, salt to remove salt humours.

Dryden spoke against a delimiting conception of either the genres or the unities. Speaking in the guise of Neander in *Of Dramatick Poesie, an Essay* (1668), he said that it was

> to the honour of our nation, that we have invented, increased, and perfected a more pleasant way of writing for the stage, than was ever known to the ancients or moderns of any nation, which is tragi-comedy.

The French dramatists, he felt, through their observance of the unities of time and place, wrote plays characterized by a dearth of plot and narrowness of imagination. Racine's approach to the question of probability was turned completely around by Dryden, who asked:

> How Many beautiful accidents might naturally happen in two or three days, which cannot arrive with any probability in the compass of twenty-four hours?

Johnson's definitive critique

The definitive critique of Neoclassical restrictions was not formulated, however, until the following century, when it was made by Samuel Johnson and was, significantly, part of his 1765 preface to Shakespeare, the first major step in the long process of establishing Shakespeare as the preeminent tragic poet of post-Classical drama. On genre he wrote:

> Shakespeare's plays are not in the rigorous and critical sense either tragedies or comedies, but compositions of a distinct kind; . . . expressing the course of the world, in which the loss of one is the gain of another; in which, at the same time, the reveller is hasting to his wine, and the mourner burying his friend . . . . That this is a practice contrary to the rules of criticism will be readily allowed; but there is always an appeal open from criticism to nature.

And on the unities:

> The necessity of observing the unities of time and place arises from the supposed necessity of making the drama credible. [But] the objection arising from the impossibility of passing the first hour at Alexandria, and the next at Rome, supposes, that when the play opens, the spectator really imagines himself at Alexandria . . . Surely he that imagines this may imagine more.

Johnson's appeal to nature was the essence of subsequent Romantic criticism.

Romantic theories. Lessing was the first important Romantic critic. He stated one of Romanticism's chief innovations in his *Hamburg Dramaturgy* (1767–69):

> The names of princes and heroes can lend pomp and majesty to a play, but they contribute nothing to our emotion. The misfortune of those whose circumstances most resemble our own, must naturally penetrate most deeply into our hearts, and if we pity kings, we pity them as human beings, not as kings.

Within a generation, revolutions in Europe and America offered social expression of this literary precept, and a dramatic tradition dominant for 22 centuries was upturned. From the time of Aristotle, who thought that the tragic hero should be highly renowned and prosperous, the tragic hero had been an aristocrat, if not a man of royal blood. With the exception of their minor or peripheral characters, the tragic dramas of Athens, England, and France told nothing of the destinies of the mass of mankind. All this was now changed.

But it is not certain that what was good for the revolution was good for tragedy. Coleridge in his critical writings of 1808–18 said that:

> there are two forms of disease most preclusive of tragic worth. The first [is] a sense and love of the ludicrous, and a diseased sensibility of the assimilating power . . . that in the boldest bursts of passion will lie in wait, or at once kindle into jest . . . . The second cause is matter of exultation to the philanthropist and philosopher, and of regret to the poet, . . . namely, the security, comparative equability, and ever-increasing sameness of human life.

In accord with this distaste for an excess of the mundane, Coleridge attacked the new German tragedies in which "the dramatist becomes a novelist *in his directions to the actors,* and degrades tragedy to pantomime." To describe, or rather indicate, what tragedy should ideally be, Coleridge said "it is not *a copy* of nature; but it is an imitation."

Coleridge's operative words and phrases in his discussions of tragedy were "innate," "from within," "implicit," "the being within," "the inmost heart," "our inward nature," "internal emotions," and "retired recesses." The new philosophical dispensation in Coleridge, like the new social dispensation in Lessing, reversed the old priorities; and where there were once princes there were now burghers, and where there were once the ordinances of God and the state there were now the dictates of the heart. By means of this reversal, Coleridge effected a reconciliation of the "tragedy of fate" and the "tragedy of character" in his description of the force of fate as merely the embodiment

Coleridge's reconciliation of fate and character

of an interior compulsion different in scale but not in kind from the interior compulsions of character. In Classical tragedy, he said the human "will" was "exhibited as struggling with fate, a great and beautiful instance and illustration of which is the Prometheus of Aeschylus; and the deepest effect is produced, when the fate is represented as a higher and intelligent will. . . ."

According to Coleridge, Shakespeare used the imaginative "variety" that characterizes man's inward nature in place of the mechanical regularity of the Neoclassical unities to produce plays that were "neither tragedies nor comedies, nor both in one, but a different genus, diverse in kind, not merely different in degree,—romantic dramas or dramatic romances." In his preoccupation with the mixture of genres and his distinction between the "mechanical" (Neoclassicism) and the "organic" (Shakespeare), Coleridge was influenced by *Lectures on Dramatic Art and Literature* (delivered 1808–09, published 1809–11), by August Wilhelm von Schlegel, perhaps the most influential of German Romantic critics.

Like Coleridge and most Romantic critics of tragedy, Schlegel found his champion in Shakespeare, and, also like them, he was preoccupied with the contrast between Classic and Romantic. Like Coleridge, Schlegel emphasized Shakespeare's inwardness, what Coleridge called his "implicit wisdom deeper even than our consciousness." It is in Shakespeare's most profound insights that Schlegel locates one of the principal distinctions between Classical and Shakespearean tragedy, in what he calls Shakespeare's "secret irony." The irony in *Oedipus the King* consists in the relation between the audience's knowledge of the protagonist's situation and his own ignorance of it. But Shakespeare's "readiness to remark the mind's fainter and involuntary utterances" is so great, says Schlegel, that "nobody ever painted so truthfully as he has done the facility of self-deception, the half self-conscious hypocrisy towards ourselves, with which even noble minds attempt to disguise the almost inevitable influence of selfish motives in human nature."

The irony Schlegel sees in Shakespeare's characterizations also extends to the whole of the action, as well as to the separate characters. In his discussion of it he suggests the reason for the difficulty of Shakespeare's plays and for the quarrelsome, irreconcilable "interpretations" among Shakespeare's commentators:

> Most poets who portray human events in a narrative or dramatic form take themselves apart, and exact from their readers a blind approbation or condemnation of whatever side they choose to support or oppose . . . . When, however, by a dexterous manoeuvre, the poet allows us an occasional glance at the less brilliant reverse of the medal, then he makes, as it were, a sort of secret understanding with the select circle of the more intelligent of his readers or spectators; he shows them that he had previously seen and admitted the validity of their tacit objections; that he himself is not tied down to the represented subject but soars freely above it. . . .

In Greek tragedy, the commentary by the chorus was an explicit and objective fact of the drama itself. In the presentation of Shakespeare's plays, such a commentary is carried on in the separate minds of the spectators, where it is diffused, silent, and not entirely sure of itself. When the spectators speak their minds after the curtain falls, it is not surprising that they often disagree.

In *Oedipus the King,* which Aristotle cited as the model of Classical tragedy, the irony of the protagonist's situation is evident to the spectator. In *Hamlet,* however, according to the American philosopher George Santayana, writing in 1908, it is the secret ironies, half-lights, and self-contradictions that make it the central creation of Romantic tragedy. As has been noted, Coleridge objected to the dramatist's giving directions to the actors, but part of the price of not having them is to deny to the audience as well an explicit indication of the playwright's meaning.

George Wilhelm Friedrich Hegel (1770–1831), the immensely influential German philospher, in his *Aesthetik* (1820–29), proposed that the sufferings of the tragic hero are merely a means of reconciling opposing moral claims. The operation is a success because of, not in spite of, the fact that the patient dies. According to Hegel's account of Greek tragedy, the conflict is not between good and evil but between goods that are each making too exclusive a claim. The heroes of ancient tragedy, by adhering to the *one* ethical system by which they molded their own personality, must come into conflict with the ethical claims of another. It is the moral one-sidedness of the tragic actor, not any negatively tragic fault in his morality or in the forces opposed to him, that proves his undoing, for both sides of the contradiction, if taken by themselves, are justified.

The nuclear Greek tragedy for Hegel is, understandably, Sophocles' *Antigone,* with its conflict between the valid claims of conscience (Antigone's obligation to give her brother a suitable burial) and law (King Creon's edict that enemies of the state should not be allowed burial). The two claims represent what Hegel regards as essentially concordant ethical claims. Antigone and Creon are, in this view, rather like pawns in the Hegelian dialectic—his theory that thought progresses from a thesis (*i.e.,* an idea), through an antithesis (an idea opposing the original thesis), to a synthesis (a more comprehensive idea that embraces both the thesis and antithesis), which in turn becomes the thesis in a further progression. At the end of *Antigone,* something of the sense of mutually appeased, if not concordant, forces does obtain after Antigone's suicide and the destruction of Creon's family. Thus, in contrast to Aristotle's statement that the tragic actors should represent not an extreme of good or evil but something between, Hegel would have them too good to live; that is, too extreme an embodiment of a particular good to survive in the world. He also tends to dismiss other traditional categories of tragic theory. For instance, he prefers his own kind of catharsis to Aristotle's—the feeling of *reconciliation.*

Tragedy as Hegelian dialectic

Hegel's emphasis on the correction of moral imbalances in tragedy is reminiscent of the "poetic justice" of Neoclassical theory, with its similar dialectic of crime and punishment. He sounds remarkably like Racine when he claims that, in the tragic denouement, the necessity of all that has been experienced by particular individuals is seen to be in complete accord with reason and is harmonized on a true ethical basis. But where the Neoclassicists were preoccupied with the unities of time and place, Hegel's concerns, like those of other Romantics, are inward. For him, the final issue of tragedy is not the misfortune and suffering of the tragic antagonists but rather the satisfaction of spirit arising from "reconciliation." Thus, the workings of the spirit, in Hegel's view, are subject to the rationalistic universal laws.

Hegel's system is not applicable to Shakespearean or Romantic tragedy. Such Shakespearean heroes as Macbeth, Richard III, and Mark Antony cannot be regarded as embodiments of any transcendent good. They behave as they do, says Hegel, now speaking outside of his scheme of tragedy, simply because they are the kind of men they are. In a statement pointing up the essence of uninhibited romantic lust and willfulness Hegel said: "it is the inner experience of their heart and individual emotion, or the particular qualities of their personality, which insist on satisfaction."

The traditional categories of tragedy are nearly destroyed in the deepened subjectivities of Romanticism of the 19th-century German philosophers, Arthur Schopenhauer and his disciple Friedrich Nietzsche. In Schopenhauer's *Die Welt als Wille und Vorstellung* (1819; *The World as Will and Idea*), much more than the social or ethical order is upturned. In place of God, the good, reason, soul, or heart, Schopenhauer installs the will, as reality's true inner nature, the metaphysical to everything physical in the world. In Schopenhauer, there is no question of a Hegelian struggle to achieve a more comprehensive good. There is rather the strife of will with itself, manifested by fate in the form of chance and error and by the tragic personages themselves. Both fate and men represent one and the same will, which lives and appears in them all. Its individual manifestations, however, in the form of such phenomena as chances, errors, or men, fight against and destroy each other.

Schopenhauer accordingly rejects the idea of poetic justice: "the demand for so-called poetical justice rests on en-

tire misconception of the nature of tragedy, and, indeed, of the nature of the world itself . . . . The true sense of tragedy is the deeper insight, that it is not his own individual sins that the hero atones for, but original sin, *i.e.,* the crime of existence itself. . . ." Schopenhauer distinguishes three types of tragic representation: (1) "by means of a character of extraordinary wickedness . . . who becomes the author of the misfortune"; (2) "blind fate—*i.e.,* chance and error" (such as the title characters in Shakespeare's *Romeo and Juliet* and "most of the tragedies of the ancients"); and (3) when "characters of ordinary morality . . . are so situated with regard to each other that their position compels them, knowingly and with their eyes open, to do each other the greatest injury, without any one of them being entirely in the wrong" (such as, "to a certain extent," *Hamlet*).

Schopenhauer's three types

This last kind of tragedy seems to Schopenhauer far to surpass the other two. His reason, almost too grim to record, is that it provides the widest possible play to the destructive manifestations of the will. It brings tragedy, so to speak, closest to home.

Schopenhauer finds tragedy to be the summit of poetical art, because of the greatness of its effect and the difficulty of its achievement. According to Schopenhauer, the egoism of the protagonist is purified by suffering almost to the purity of nihilism. His personal motives become dispersed as his insight into them grows; "the complete knowledge of the nature of the world, which has a quieting effect on the will, produces resignation, the surrender not merely of life, but of the very will to live."

Schopenhauer's description has limited application to tragic denouements in general. In the case of his own archetypal hero, the hero's end seems merely the mirror image of his career, an oblivion of resignation or death that follows an oblivion of violence. Instead of a dialogue between higher and lower worlds of morality or feeling (which take place even in Shakespeare's darkest plays), Schopenhauer posits a succession of states as helpless in knowledge as in blindness. His "will" becomes a synonym for all that is possessed and necessity-ridden.

Nietzsche's division of tragedy in two elements

Nietzsche's *Geburt der Tragödie aus dem Geiste der Musik* (1872) was deeply influenced by Schopenhauer. The two elements of tragedy, says Nietzsche, are the Apollonian (related to the Greek god Apollo, here used as a symbol of measured restraint) and the Dionysian (from Dionysus, the Greek god of ecstasy). His conception of the Apollonian is the equivalent of what Schopenhauer called the individual phenomenon—the particular chance, error, or man, the individuality of which is merely a mask for the essential truth of reality which it conceals. The Dionysian element is a sense of universal reality, which, according to Schopenhauer, is experienced after the loss of individual egoism. The "Dionysian ecstasy," as defined by Nietzsche, is experienced "not as individuals but as the *one* living being, with whose creative joy we are united."

Nietzsche dismisses out of hand one of the most venerable features of the criticism of tragedy, the attempt to reconcile the claims of ethics and art. He says that the events of a tragedy are "supposed" to discharge pity and fear and are "supposed" to elevate and inspire by the triumph of noble principles at the sacrifice of the hero. But art, he says, must demand purity within its own sphere. To explain tragic myth, the first requirement is to seek the pleasure that is peculiar to it in the purely aesthetic sphere, without bringing in pity, fear, or the morally sublime.

The essence of this specifically aesthetic tragic effect is that it both reveals and conceals, causing both pain and joy. The drama's exhibition of the phenomena of suffering individuals (Apollonian elements) forces upon the audience "the struggle, the pain, the destruction of phenomena," which in turn communicates "the exuberant fertility of the universal." The spectators then "become, as it were, one with the infinite primordial joy in existence, and . . . we anticipate, in Dionysian ecstasy, the indestructibility and eternity of this joy." Thus, he says, there is a desire "to see tragedy and at the same time to get beyond all seeing . . . to hear and at the same time long to get beyond all hearing."

The inspired force of Nietzsche's vision is mingled with a sense of nihilism:

> "only after the spirit of science has been pursued to its limits, . . . may we hope for a rebirth of tragedy . . . I understand by the spirit of science the faith that first came to light in the person of Socrates—the faith in the explicability of nature and in knowledge as a panacea."

Nietzsche would replace the spirit of science with a conception of existence and the world as an aesthetic phenomenon and justified only as such. Tragedy would enjoy a prominent propagandistic place. It is "precisely the tragic myth that has to convince us that even the ugly and disharmonic are part of an artistic game that the will in the eternal amplitude of its pleasure plays with itself." And, consummately: "we have art in order that we may not perish through truth."

Tragedy in music. Musical dissonance was Nietzsche's model for the double effect of tragedy. The first edition of his book was titled *The Birth of Tragedy out of the Spirit of Music,* another influence from Schopenhauer, for whom music differed from all the other arts in that it is not a copy of a phenomenon but the direct copy of the will itself. He even called the world "embodied music , . . . embodied will." Nietzsche's theorizing on the relation of the tragic theme to art forms other than the drama was in fact confirmed in such operas as Mussorgsky's version of Pushkin's tragedy *Boris Godunov,* Verdi's of *Macbeth* and *Othello,* and Gounod's *Faust.* In contrast to these resettings of received forms, Wagner, Verdi, and Bizet achieved a new kind of tragic power for Romanticism in the theme of the operatic love-death in, respectively, *Tristan and Isolde, Aida,* and *Carmen.* Thus, the previous progression of the genre from tragedy to tragicomedy to romantic tragedy continued to a literary–musical embodiment of what Nietzsche called "tragic dithyrambs."

An earlier prophecy than Nietzsche's regarding tragedy and opera was made by the German poet Friedrich von Schiller in a letter of 1797 to Goethe:

> I have always trusted that out of opera, as out of the choruses of the ancient festival of Bacchus, tragedy would liberate itself and develop in a nobler form. In opera, servile imitation of nature is dispensed with . . . here is . . . the avenue by which the ideal can steal its way back into the theatre.

20th-century critical theory. In the 20th century, discussion of tragedy was sporadic until the aftermath of World War II. Then it enjoyed new vigour, perhaps to compensate for, or help explain, the dearth of genuine tragic literature, either in the novel or in the theatre. In the 1950s and 1960s countless full-length studies, articles, and monographs variously sought the essence, the vision, the view of life, or the spirit of tragedy out of a concern for the vital culture loss were the death of tragedy to become a reality. They also attempted to mediate the meaning of tragedy to a public that was denied its reality, save in revivals or an occasional approximation. Since the Romantic critics first ventured beyond the Aristotelean categories to consider tragedy, or the tragic, as a sense of life, there was an increasing tendency to regard tragedy not merely as drama but as a philosophical form. It is noteworthy that the Spanish philosopher Miguel de Unamuno's influential book, *The Tragic Sense of Life* (1921), barely mentions the formal drama.

From the time of Aristotle, tragedy has achieved importance primarily as a medium of self-discovery—the discovery of man's place in the universe and in society. That is the main concern of Aristotle in his statements about reversal, recognition, and catharsis, though it remained for the Romantic critics to point it out. The loss of this concern in the facile plays of the 19th and 20th centuries resulted in the reduction of tragic mystery to confused sentimentalism. Critics of the 20th century, being less certain even than Schopenhauer or Nietzsche of what man's place in the scheme of things may be, experimented with a variety of critical approaches, just as contemporary dramatists experimented with various "theatres." Although these critics lacked the philosophical certainties of earlier theorists, they had a richer variety of cultures and genres to instruct them. The hope of both critics and dramatists was that this multiplicity would produce not mere impressionism or haphazard eclecticism but new form and new meaning.

(L.C.)

OTHER GENRES

Satire

"Satire" is a protean term. Together with its derivatives, it is one of the most heavily worked literary designations and one of the most imprecise. The great English lexicographer Samuel Johnson defined satire as "a poem in which wickedness or folly is censured," and more elaborate definitions are rarely more satisfactory. No strict definition can encompass the complexity of a word that signifies, on one hand, a kind of literature—as when one speaks of the satires of the Roman poet Horace or calls the American novelist Nathanael West's *A Cool Million* a satire—and, on the other, a mocking spirit or tone that manifests itself in many literary genres but can also enter into almost any kind of human communication. Wherever wit is employed to expose something foolish or vicious to criticism, there satire exists, whether it be in song or sermon, in painting or political debate, on television or in the movies. In this sense satire is everywhere. Although this section deals primarily with satire as a literary phenomenon, it records its manifestations in a number of other areas of human activity.

THE NATURE OF SATIRE

Historical definitions. The terminological difficulty is pointed up by a phrase of the Roman rhetorician Quintilian: "satire is wholly our own" ("satura tota nostra est"). Quintilian seems to be claiming satire as a Roman phenomenon, although he had read the Greek dramatist Aristophanes and was familiar with a number of Greek forms that one would call satiric. But the Greeks had no specific word for satire; and by *satura* (which meant originally something like "medley" or "miscellany" and from which comes the English "satire") Quintilian intended to specify that kind of poem "invented" by Lucilius, written in hexameters on certain appropriate themes, and characterized by a Lucilian–Horatian tone. *Satura* referred, in short, to a poetic form, established and fixed by Roman practice. (Quintilian mentions also an even older kind of satire written in prose by Marcus Terentius Varro and, one might add, by Menippus and his followers Lucian and Petronius.) After Quintilian's day *satura* began to be used metaphorically to designate works that were "satirical" in tone but not in form. As soon as a noun enters the domain of metaphor, as one modern scholar has pointed out, it clamours for extension; and *satura* (which had had no verbal, adverbial, or adjectival forms) was immediately broadened by appropriation from the Greek word for "satyr" (*satyros*) and its derivatives. The odd result is that the English "satire" comes from the Latin *satura;* but "satirize," "satiric," etc., are of Greek origin. By about the 4th century AD the writer of satires came to be known as *satyricus;* St. Jerome, for example, was called by one of his enemies "a satirist in prose" ("satyricus scriptor in prosa"). Subsequent orthographic modifications obscured the Latin origin of the world satire: *satura* becomes *satyra,* and in England, by the 16th century, it was written "satyre."

Satire and satyr

Elizabethan writers, anxious to follow Classical models but misled by a false etymology, believed that "satyre" derived from the Greek satyr play: satyrs being notoriously rude, unmannerly creatures, it seemed to follow that "satyre" should be harsh, coarse, rough. The English author Joseph Hall wrote:

> The Satyre should be like the Porcupine,
> That shoots sharpe quils out in each angry line,
> And wounds the blushing cheeke, and fiery eye,
> Of him that heares, and readeth guiltily.
> (*Virgidemiarum,* V, 3, 1–4.)

The false etymology that derives satire from satyrs was finally exposed in the 17th century by the Classical scholar Isaac Casaubon; but the old tradition has aesthetic if not etymological appropriateness and has remained strong.

In the prologue to his book, Hall makes a claim that has caused confusion like that following from Quintilian's remark on Roman satire. Hall boasts:

> I first adventure: follow me who list,
> And be the second English Satyrist.

But Hall knew the satirical poems of Geoffrey Chaucer and John Skelton, among other predecessors, and probably meant that he was the first to imitate systematically the formal satirists of Rome.

Influence of Horace and Juvenal. By their practice, the great Roman poets Horace and Juvenal set indelibly the lineaments of the genre known as the formal verse satire and, in so doing, exerted pervasive, if often indirect, influence on all subsequent literary satire. They gave laws to the form they established, but it must be said that the laws were very loose indeed. Consider, for example, style. In three of his Satires (I, iv; I, x; II, i) Horace discusses the tone appropriate to the satirist who out of a moral concern attacks the vice and folly he sees around him. As opposed to the harshness of Lucilius, Horace opts for mild mockery and playful wit as the means most effective for his ends. Although I portray examples of folly, he says, I am not a prosecutor and I do not like to give pain; if I laugh at the nonsense I see about me, I am not motivated by malice. The satirist's verse, he implies, should reflect this attitude: it should be easy and unpretentious, sharp when necessary, but flexible enough to vary from grave to gay. In short, the character of the satirist as projected by Horace is that of an urbane man of the world, concerned about folly, which he sees everywhere, but moved to laughter rather than rage.

Horace's concept of the satirist

Juvenal, over a century later, conceives the satirist's role differently. His most characteristic posture is that of the upright man who looks with horror on the corruptions of his time, his heart consumed with anger and frustration. Why does he write satire? Because tragedy and epic are irrelevant to his age. Viciousness and corruption so dominate Roman life that for an honest man it is difficult *not* to write satire. He looks about him, and his heart burns dry with rage; never has vice been more triumphant. How can he be silent (*Satires,* I)? Juvenal's declamatory manner, the amplification and luxuriousness of his invective, are wholly out of keeping with the stylistic prescriptions set by Horace. At the end of the scabrous sixth satire, a long, perfervid invective against women, Juvenal flaunts his innovation: in this poem, he says, satire has gone beyond the limits established by his predecessors; it has taken to itself the lofty tone of tragedy.

The results of Juvenal's innovation have been highly confusing for literary history. What *is* satire if the two poets universally acknowledged to be supreme masters of the form differ so completely in their work as to be almost incommensurable? The formulation of the English poet John Dryden has been widely accepted. Roman satire has two kinds, he says: comical satire and tragical satire, each with its own kind of legitimacy. These denominations have come to mark the boundaries of the satiric spectrum, whether reference is to poetry or prose or to some form of satiric expression in another medium. At the Horatian end of the spectrum, satire merges imperceptibly into comedy, which has an abiding interest in the follies of men but has not satire's reforming intent. The distinction between the two modes, rarely clear, is marked by the intensity with which folly is pursued: fops and fools and pedants appear in both, but only satire tries to mend men through them. And, although the great engine of both comedy and satire is irony, in satire, as the 20th-century critic Northrop Frye has said, irony is militant.

Comical satire and tragical satire

Boileau, Dryden, and Alexander Pope, writing in the 17th and 18th centuries—the modern age of satire—catch beautifully, when they like, the deft Horatian tone; however, satire's wit can also be sombre, deeply probing, and prophetic, as it explores the ranges of the Juvenalian end of the satiric spectrum, where satire merges with tragedy, melodrama, and nightmare. Pope's *Dunciad* ends with these lines:

> Lo! thy dread Empire, CHAOS! is restor'd;
> Light dies before thy uncreating word:

Thy hand, great Anarch! lets the curtain fall;
And Universal Darkness buries All.

It is the same darkness that falls on Book IV of Jonathan Swift's *Gulliver's Travels,* on some of Mark Twain's satire—*The Mysterious Stranger, To The Person Sitting in Darkness*—and on George Orwell's *1984.*

Structure of verse satire. Roman satire is hardly more determinate in its structure than in its style; the poems are so haphazardly organized, so randomly individual, that there seems little justification for speaking of them as a literary kind at all. Beneath the surface complexity of the poems, however, there exists, as one modern scholar has pointed out, a structural principle common to the satires of the Roman poets and their French and English followers. These poems have a bipartite structure: a thesis part, in which some vice or folly is examined critically from many different angles and points of view, and an antithesis part, in which an opposing virtue is recommended. The two parts are disproportionate in length and in importance, for satirists have always been more disposed to castigate wickedness than exhort to virtue.

Most verse satires are enclosed by a "frame." Just as a novel by the early-20th-century writer Joseph Conrad may be framed by a situation in which his narrator sits on a veranda in the tropics, telling his tale, stimulated into elaboration by the queries of his listeners, so the satire will be framed by a conflict of sorts between the satirist (or, more reasonably, his persona, a fictive counterpart, the "I" of the poem) and an adversary. Usually the adversary has a minor role, serving only to prod the speaker into extended comment on the issue (vice or folly) at hand; he may be sketchily defined, or he may be as effectively projected as Horace's Trebatius (*Satires,* II, i) or his awful bore (I, vi) or his slave Davus, who turns the tables on his master (II, vii). Similarly, the background against which the two talk may be barely suggested, or it may form an integral part of the poem, as in Horace's "Journey to Brundisium" (I, v) or in Juvenal's description of the valley of Egeria, where Umbricius unforgettably pictures the turbulence and decadence of Rome (*Satires,* III). In any event, the frame is usually there, providing a semidramatic situation in which vice and folly may reasonably be dissected.

Literary and rhetorical devices

The satirist has at his disposal an immense variety of literary and rhetorical devices: he may use beast fables, dramatic incidents, fictional experiences, imaginary voyages, character sketches, anecdotes, proverbs, homilies; he may employ invective, sarcasm, burlesque, irony, mockery, raillery, parody, exaggeration, understatement—wit in any of its forms—anything to make the object of attack abhorrent or ridiculous. Amid all this confusing variety, however, there is pressure toward order—internally, from the arraignment of vice and appeal to virtue, and externally, from the often shadowy dramatic situation that frames the poem.

The satiric spirit. Thus, although the formal verse satire of Rome is quantitatively a small body of work, it contains most of the elements later literary satirists employ. When satire is spoken of today, however, there is usually no sense of formal specification whatever; one has in mind a work imbued with the satiric spirit—a spirit that appears (whether as mockery, raillery, ridicule, or formalized invective) in the literature or folklore of all peoples, early and late, preliterate and civilized. According to Aristotle (*Poetics,* IV, 1448b–1449a), Greek Old Comedy developed out of ritualistic ridicule and invective—out of satiric utterances, that is, improvised and hurled at individuals by the leaders of the phallic songs. The function of these "iambic" utterances, it has been shown, was magical; they were thought to drive away evil influences so that the positive fertility magic of the phallus might be operative.

Power of ancient satire

This early connection of primitive "satire" with magic has a remarkably widespread history.

In the 7th century BC, the poet Archilochus, said to be the "first" Greek literary satirist, composed verses of such potency against his prospective father-in-law, Lycambes, that Lycambes and his daughter hanged themselves. In the next century the sculptors Bupalus and Athenis "knit their necks in halters," it is said, as a result of the "bitter rimes and biting libels" issued by the satirical poet Hipponax. Similar tales exist in other cultures. The chief function of the ancient Arabic poet was to compose satire (*hijāʾ*) against the tribal enemy. The satires were thought always to be fatal, and the poet led his people into battle, hurling his verses as he would hurl a spear. Old Irish literature is laced with accounts of the extraordinary power of the poets, whose satires brought disgrace and death to their victims:

. . . saith [King] Lugh to his poet, "what power can *you* wield in battle?"
"Not hard to say," quoth Carpre. . . ."I will satirize them, so that through the spell of my art they will not resist warriors."
("The Second Battle of Moytura," trans. by W. Stokes, *Revue Celtique,* XII [1891], 52–130.)

According to saga, when the Irish poet uttered a satire against his victim, three blisters would appear on the victim's cheek, and he would die of shame. One story will serve as illustration: after Deirdriu of the Sorrows came to her unhappy end, King Conchobar fell in love again—this time with the lovely Luaine. They were to be married; but, when the great poet Aithirne the Importunate and his two sons (also poets) saw Luaine, they were overcome with desire for her. They went to Luaine and asked her to sleep with them. She refused. The poets threatened to satirize her. And the story says:

The damsel refused to lie with them. So then they made three satires on her, which left three blotches on her cheeks, to wit, Shame and Blemish and Disgrace . . . . Thereafter the damsel died of shame. . . .
("The Wooing of Luaine . . . ," trans. by W. Stokes, *Revue Celtique,* XXIV [1903], 273–85.)

An eminent 20th-century authority on these matters adduces linguistic, thematic, and other evidence to show a functional relation between primitive "satire," such as that of Carpre and Aithirne, and the "real" satire of more sophisticated times. Today, among various preliterate peoples the power of personal satire and ridicule is appalling; among the Ashanti of West Africa, for example, ridicule is (or was recently) feared more than almost any other humanly inflicted punishment, and suicide is frequently resorted to as an escape from its terrors. Primitive satire such as that described above can hardly be spoken of in literary terms; its affiliations are rather with the magical incantation and the curse.

SATIRICAL MEDIA

Literature. When the satiric utterance breaks loose from its background in ritual and magic, as in ancient Greece (when it is free, that is, to develop in response to literary stimuli rather than the "practical" impulsions of magic), it is found embodied in an indefinite number of literary forms that profess to convey moral instruction by means of laughter, ridicule, mockery; the satiric spirit proliferates everywhere, adapting itself to whatever mode (verse or prose) seems congenial. Its targets range from one of Pope's dunces to the entire race of man, as in *Satyr Against Mankind* (1675), by John Wilmot, the earl of Rochester, from Erasmus' attack on corruptions in the church to Swift's excoriation of all civilized institutions in *Gulliver's Travels.* Its forms are as varied as its victims: from an anonymous medieval invective against social injustice to the superb wit of Chaucer and the laughter of Rabelais; from the burlesque of Luigi Pulci to the scurrilities of Pietro Aretino and the "black humour" of Lenny Bruce; from the flailings of John Marston and the mordancies of Francisco Gómez de Quevedo y Villegas to the bite of Jean de La Fontaine and the great dramatic structures of Ben Jonson and Molière; from an epigram of Martial to the fictions of Nikolay Gogol and of Günter Grass and the satirical utopias of Yevgeny Zamyatin, Aldous Huxley, and Orwell.

It is easy to see how the satiric spirit would combine readily with those forms of prose fiction that deal with the ugly realities of the world but that satire should find congenial a genre such as the fictional utopia seems odd. From the publication of Thomas More's eponymous *Utopia* (1516), however, satire has been an important ingredient of utopian fiction. More drew heavily on the satire of Horace, Juvenal, and Lucian in composing his

Use of satire in fictional utopias

great work. For example, like a poem by Horace, *Utopia* is framed by a dialogue between "Thomas More" (the historical man a character in his own fiction) and a seafaring philosopher named Raphael Hythloday. The two talk throughout a long and memorable day in a garden in Antwerp. "More's" function is to draw Hythloday out and to oppose him on certain issues, notably his defense of the communism he found in the land of Utopia. "More" is the adversary. Hythloday's role is to expound on the institutions of Utopia but also to expose the corruption of contemporary society. Thus he functions as a satirist. Here Hythloday explains why Englishmen, forced off their land to make way for sheep, become thieves:

> Forsooth . . . your sheep that were wont to be so meek and tame and so small eaters, now as I hear say, be become so great devourers and so wild, that they eat up and swallow down the very men themselves. They consume, destroy, and devour whole fields, houses, and cities. For look in what parts of the realm doth grow the finest and therefore dearest wool, there noblemen and gentlemen, yea and certain abbots, holy men no doubt, not contenting themselves with the yearly revenues and profits that were wont to grow to their forefathers and predecessors of their lands, nor being content that they live in rest and pleasure nothing profiting, yea, much annoying the weal-public, leave no ground for tillage. They enclose all into pastures; they throw down houses; they pluck down towns and leave nothing standing but only the church to be made a sheep-house.
>
> (More's *Utopia,* Everyman edition, 1951.)

Here are characteristic devices of the satirist, dazzlingly exploited: the beast fable compressed into the grotesque metaphor of the voracious sheep; the reality-destroying language that metamorphoses gentlemen and abbots into earthquakes and a church into a sheep barn; the irony coldly encompassing the passion of the scene. Few satirists of any time could improve on this.

Just as satire is a necessary element of the work that gave the literary form utopia its name, so the utopias of Lilliput, Brobdingnag, and Houyhnhnmland are essential to the satire of More's great follower Jonathan Swift. He sent Gulliver to different lands from those Hythloday discovered, but Gulliver found the same follies and the same vices, and he employed a good many of the same rhetorical techniques his predecessor had used to expose them. *Gulliver's Travels,* as one scholar points out, is a salute across the centuries to Thomas More. With this kind of precedent, it is not surprising that in the 20th century, when utopia turns against itself, as in Aldous Huxley's *Brave New World* (1932), the result is satire unrelieved.

Drama. The drama has provided a favourable environment for satire ever since it was cultivated by Aristophanes, working under the extraordinarily open political conditions of 5th-century Athens. In a whole series of plays—*The Clouds, The Frogs, Lysistrata,* and many others—Aristophanes lampoons the demagogue Cleon by name, violently attacks Athenian war policy, derides the audience of his plays for their gullible complacency, pokes fun at Socrates as representative of the new philosophical teaching, stages a brilliantly parodic poetic competition between the dramatists Aeschylus and Euripides in Hades, and in general lashes out at contemporary evils with an uninhibited and unrivalled inventiveness. But the theatre has rarely enjoyed the political freedom Aristophanes had—one reason, perhaps, that satire more often appears in drama episodically or in small doses than in the full-blown Aristophanic manner. In Elizabethan England, Ben Jonson wrote plays that he called "comicall satyres"—*Every Man Out of His Humour, Poetaster*—and there are substantial elements of satire in Shakespeare's plays—some in the comedies, but more impressively a dark and bitter satire in *Timon of Athens, Troilus and Cressida, Hamlet,* and *King Lear.* The 17th-century comedy of Molière sometimes deepens into satire, as with the exposure of religious hypocrisy in *Tartuffe* or the railing against social hypocrisy by Alceste in *The Misanthrope.* George Bernard Shaw considered himself a satirist. He once compared his country's morals to decayed teeth and himself to a dentist, obliged by his profession to give pain in the interests of better health. Yet, as inventive and witty as Shaw is, compared to the 20th-century German playwright Bertolt Brecht, whose anatomizing of social injustice cuts deep, Shaw is a gentle practitioner indeed.

Aristophanes' use of satire

Motion pictures and television. The movies have sometimes done better by satire than the theatre, and it is in the movies that an ancient doctrine having to do with principles of decorum in the use of satire and ridicule has been exploded. The English novelist Henry Fielding was reflecting centuries of tradition when, in the preface to *Joseph Andrews* (1742), he spoke of the inappropriateness of ridicule applied to black villainy or dire calamity. "What could exceed the absurdity of an Author, who should write *the Comedy of* Nero, *with the merry Incident of ripping up his Mother's Belly?*" Given this point of view, Hitler seems an unlikely target for satire; yet in *The Great Dictator* (1940) Charlie Chaplin managed a successful, if risky, burlesque. Chaplin has written, however, that, determined as he was to ridicule the Nazi notions of a superrace, if he had known of the horrors of the concentration camps, he could not have made the film. Stanley Kubrick's *Dr. Strangelove; or, How I Learned to Stop Worrying and Love the Bomb* (1964) denies all limitation; through some alchemy Kubrick created an immensely funny, savagely satirical film about the annihilation of the world. A combination of farce and nightmare, *Dr. Strangelove* satirizes military men, scientists, statesmen—the whole ethos of the technological age—in the most mordant terms; it shows the doomsday blast, yet leaves audiences laughing. "You can't fight in here," says the president of the United States as doom nears, "this is the War Room." The film's tone is less didactic than in most powerful satire—the mushroom cloud carries its own moral—yet satire's full force is there.

Television has not proved a notably receptive medium. *That Was the Week That Was,* a weekly satirical review started in England in 1962, had remarkable success for a time but succumbed to a variety of pressures, some of them political; when a version of the program was attempted in the United States, it was emasculated by restrictions imposed by sponsors fearful of offending customers and by program lawyers wary of libel suits. Jonathan Swift said that he wrote to vex the world rather than divert it; it is not an attitude calculated to sell soap.

Festivals. Yet satire does much more than vex, and even in Swift's work there is a kind of gaiety that is found in many nonliterary manifestations of the satirical spirit. Satire always accompanies certain festivals, for example, particularly saturnalian festivals. Many different cultures set aside a holiday period in which customary social restraints are abandoned, distinctions of rank and status are turned upside down, and institutions normally sacrosanct are subjected to ridicule, mockery, burlesque. The Romans had their Saturnalia, the Middle Ages its Feast of Fools; and in the 20th century many countries still have annual carnivals (Fasching in Austria, the Schnitzelbank in Basel, Switzerland, for example) at which, amid other kinds of abandon, an extraordinary freedom of satirical utterance is permitted. Even in Africa among the Ashanti, for whom ridicule has such terrors, there is a festival during which the sacred chief himself is satirized. "Wait until Friday," said the chief to the enquiring anthropologist, "when the people really begin to abuse me, and if you will come and do so too it will please me." Festivals such as these provide sanctioned release from social inhibition and repression, and, in these circumstances, satire directed at men in power or at taboo institutions acts as a safety valve for pent-up frustrations.

The gaiety of satire

Satire may often function this way. A story is told that the 16th-century pope Adrian VI was highly offended at satirical verses written against him and affixed to Pasquino's statue (a famous repository for lampoons in Rome), but he became a willing target once he realized that his enemies vented otherwise dangerous hostility in this relatively harmless manner. Similar mechanisms operate today when, at a nightclub or theatre, audiences listen to satirical attacks on political figures or on issues such as racial discrimination, identify with the satirist, laugh at his wit, and thereby discharge their own aggressive feelings. Satire of this order

is a far cry, of course, from that written by a Swift or Voltaire, whose work can be said to have a revolutionary effect.

Caricature and cartoon

Visual arts. The critique of satire may be conveyed even more potently in the visual arts than by way of the spoken or written word. In caricature and in what came to be known as the cartoon, artists since the Renaissance have left a wealth of startlingly vivid commentary on the men and events of their time. The names alone evoke their achievement: in England, William Hogarth, Thomas Rowlandson, Sir John Tenniel, Sir Max Beerbohm; in France, Charles Philipon (whose slow-motion metamorphosis of King Louis-Philippe into a *poire*—that is, "fathead," or "fool"—is classic) and Honoré Daumier; in Spain, Francisco Goya—out of Spain, Pablo Picasso; among recent political cartoonists, Sir David Low, Vicky (Victor Welsz), Herblock (Herbert Block), Conrad. The favourite medium of these men is the black-and-white print in which the satirical attack is pointed up by a brief verbal caption. The social impact of their art is incalculable. Dictators know this, and in times of social tension political cartoonists are among the first victims of the censor.

THE SATIRIST, THE LAW, AND SOCIETY

Indeed, the relations of satirists to the law have always been delicate and complex. Both Horace and Juvenal took extraordinary pains to avoid entanglements with authority—Juvenal ends his first satire with the self-protective announcement that he will write only of the dead. In England in 1599 the Archbishop of Canterbury and the Bishop of London issued an order prohibiting the printing of any satires whatever and requiring that the published satires of Hall, John Marston, Thomas Nashe, and others be burned. Today, the satirist attacks individuals only at the risk of severe financial loss to himself and his publisher. In totalitarian countries he risks imprisonment or death. Under extreme conditions satire against the reigning order is out of the question: so canonical is this rule that political analysts use the amount and character of satire permitted in the U.S.S.R. as a gauge of political pressure at any given time. Whereas the poet Osip Mandelstam was sent to a concentration camp and his death for composing a satirical poem on Stalin, today Isaac Deutscher can interpret as a favourable sign the increasing amount of Soviet satire against internal abuses.

Repressions of satire

One creative response the satirist makes to social and legal pressures is to try by rhetorical means to approach his target indirectly; that is, a prohibition of direct attack fosters the manoeuvres of indirection that will make the attack palatable: irony, innuendo, burlesque, parody, allegory, etc. It is a nice complication that the devices that render satire acceptable to society at the same time sharpen its point. "Abuse is not so dangerous," said Dr. Johnson, "when there is no vehicle of wit or delicacy, no subtle conveyance." The conveyances are born out of prohibition.

Anthony Cooper, 3rd earl of Shaftesbury, writing in the 18th century, recognized the "creative" significance of legal and other repressions on the writing of satire. "The greater the weight [of constraint] is, the bitterer will be the satire. The higher the slavery, the more exquisite the buffoonery." Shaftesbury's insight requires the qualification made above. Under a massively efficient tyranny, satire of the forms, institutions, or personalities of that tyranny is impossible. But, under the more relaxed authoritarianism of an easier going day, remarkable things could be done. Max Radin, a Polish-born American author, writes of how satirical journals in Germany before World War I, even in the face of a severe law, vied with each other to see how close they could come to caricatures of the Kaiser without actually producing them. "Satire which the censor understands," said the Austrian satirist Karl Kraus, "deserves to be banned." The 20th-century U.S. critic Kenneth Burke sums up this paradoxical aspect of satire's relation with the law by suggesting that the most inventive satire is produced when the satirist knowingly takes serious risks and is not sure whether he will be acclaimed or punished. The whole career of Voltaire is an excellent case in point. Bigots and tyrants may have turned pale at his name, as a famous hyperbole has it; but Voltaire's satire was sharpened and his life rendered painfully complicated as he sought to avoid the penalties of the law and the wrath of those he had angered. Men such as Voltaire and Kraus and the Russian Ye.I. Zamyatin attack evil in high places, pitting their wit and moral authority against cruder forms of power. In this engagement there is often something of the heroic.

Readers have an excellent opportunity to examine the satirist's claim to social approval by reason of the literary convention that decrees that he must justify his problematic art. Nearly all satirists write apologies, and nearly all the apologies project an image of the satirist as a plain, honest man, wishing harm to no worthy person but appalled at the evil he sees around him and forced by his conscience to write satire. Pope's claim is the most extravagant:

> Yes, I am proud; I must be proud to see
> Men not afraid of God, afraid of me:
> Safe from the Bar, the Pulpit, and the Throne,
> Yet touch'd and sham'd by *Ridicule* alone.
> O sacred Weapon! left for Truth's defence,
> Sole Dread of Folly, Vice, and Insolence!
> (*Epilogue to the Satires,* II, 208–13.)

Relationship between the satirist and his audience

After the great age of satire, which Pope brought to a close, such pretensions would have been wholly anachronistic. Ridicule depends on shared assumptions against which the deviant stands in naked relief. The satirist must have an audience that shares with him commitment to certain intellectual and moral standards that validate his attacks on aberration. The greatest satire has been written in periods when ethical and rational norms were sufficiently powerful to attract widespread assent yet not so powerful as to compel absolute conformity—those periods when the satirist could be of his society and yet apart, could exercise a double vision. Neoclassic writers had available to them as an implicit metaphor the towering standard of the classical past; for the 19th and 20th centuries no such metaphors have been available. It is odd, however, that, whereas the 19th century in general disliked and distrusted satire (there are obvious exceptions), our own age, bereft of unifying symbols, scorning traditional rituals, searching for beliefs, still finds satire a congenial mode in almost any medium. Although much of today's satire is self-serving and trivial, there are notable achievements. Joseph Heller's novel *Catch-22* (1961) once again employs farce as the agent of the most probing criticism: Who is sane, the book asks, in a world whose major energies are devoted to blowing itself up? Beneath a surface of hilariously grotesque fantasy, in which characters from Marx brothers comedy carry out lethal assignments, there is exposed a dehumanized world of hypocrisy, greed, and cant. Heller is a satirist in the great tradition. If he can no longer, like Pope, tell men with confidence what they should be for, he is splendid at showing them what they must be against. The reader laughs at the mad logic of *Catch-22,* and as he laughs he learns. This is the way satire has worked from the beginning. (R.C.E.)

Nonfictional prose

Defining nonfictional prose literature is an immensely challenging task. Nonfictional prose literature differs from bald statements of fact, such as those recorded in an old chronicle or inserted in a business letter or in an impersonal message of mere information. As used in a broad sense, the term nonfictional prose literature here designates writing intended to instruct (but not highly scientific and erudite writings in which no aesthetic concern is evinced), to impart wisdom or faith, and especially to please. Separate sections cover biographical literature and literary criticism.

NATURE

Nonfictional prose genres cover an almost infinite variety of themes, and they assume many shapes. In quantitative terms, if such could ever be valid in such nonmeasurable matters, they probably include more than half of all that has been written in countries having a literature of their

own. Nonfictional prose genres have flourished in nearly all countries with advanced literatures. The genres include political and polemical writings, biographical and autobiographical literature, religious writings, and philosophical, and moral or religious writings.

After the Renaissance, from the 16th century onward in Europe, a personal manner of writing grew in importance. The author strove for more or less disguised self-revelation and introspective analysis, often in the form of letters, private diaries, and confessions. Also of increasing importance were aphorisms after the style of the ancient Roman philosophers Seneca and Epictetus, imaginary dialogues, and historical narratives, and later, journalistic articles and extremely diverse essays. From the 19th century, writers in Romance and Slavic languages especially, and to a far lesser extent British and American writers, developed the attitude that a literature is most truly modern when it acquires a marked degree of self-awareness and obstinately reflects on its purpose and technique. Such writers were not content with imaginative creation alone: they also explained their work and defined their method in prefaces, reflections, essays, self-portraits, and critical articles. The 19th-century French poet Charles Baudelaire asserted that no great poet could ever quite resist the temptation to become also a critic: a critic of others and of himself. Indeed, most modern writers, in lands other than the United States, whether they be poets, novelists, or dramatists, have composed more nonfictional prose than poetry, fiction, or drama. In the instances of such monumental figures of 20th-century literature as the poets Ezra Pound, T.S. Eliot, and perhaps William Butler Yeats, or the novelists Thomas Mann and André Gide, that part of their output may well be considered by posterity to be equal in importance to their more imaginative writing.

It is virtually impossible to attempt a unitary characterization of nonfictional prose. The concern that any definition is a limitation, and perhaps an exclusion of the essential, is nowhere more apposite than to this inordinately vast and variegated literature. Ever since the ancient Greek and Roman philosophers devised literary genres, some critics have found it convenient to arrange literary production into kinds or to refer it to modes.

ELEMENTS

Obviously, a realm as boundless and diverse as nonfictional prose literature cannot be characterized as having any unity of intent, of technique, or of style. It can be defined, very loosely, only by what it is not. Many exceptions, in such a mass of writings, can always be brought up to contradict any rule or generalization. No prescriptive treatment is acceptable for the writing of essays, of aphorisms, of literary journalism, of polemical controversy, of travel literature, of memoirs and intimate diaries. No norms are recognized to determine whether a dialogue, a confession, a piece of religious or of scientific writing, is excellent, mediocre, or outright bad, and each author has to be relished, and appraised, chiefly in his own right. "The only technique," the English critic F.R. Leavis wrote in 1957, "is that which compels words to express an intensively personal way of feeling." Intensity is probably useful as a standard; yet it is a variable, and often elusive, quality, possessed by polemicists and by ardent essayists to a greater extent than by others who are equally great. "Loving, and taking the liberties of a lover" was Virginia Woolf's characterization of the 19th-century critic William Hazlitt's style: it instilled passion into his critical essays. But other equally significant English essayists of the same century, such as Charles Lamb or Walter Pater, or the French critic Hippolyte Taine, under an impassive mask, loved too, but differently. Still other nonfictional writers have been detached, seemingly aloof, or, like the 17th-century French epigrammatist La Rochefoucauld, sarcastic. Their intensity is of another sort.

Reality and imagination. Prose that is nonfictional is generally supposed to cling to reality more closely than that which invents stories, or frames imaginary plots. Calling it "realistic," however, would be a gross distortion. Since nonfictional prose does not stress inventiveness of themes and of characters independent of the author's self, it appears in the eyes of some moderns to be inferior to works of imagination. In the middle of the 20th century an immensely high evaluation was placed on the imagination, and the adjective "imaginative" became a grossly abused cliché. Many modern novels and plays, however, were woefully deficient in imaginative force, and the word may have been bandied about so much out of a desire for what was least possessed. Many readers are engrossed by travel books, by descriptions of exotic animal life, by essays on the psychology of other nations, by Rilke's notebooks or by Samuel Pepys's diary far more than by poetry or by novels that fail to impose any suspension of disbelief. There is much truth in Oscar Wilde's remark that "the highest criticism is more creative than creation and the primary aim of the critic is to see the object as in itself it really is not." A good deal of imagination has gone not only into criticism but also into the writing of history, of essays, of travel books, and even of the biographies or the confessions that purport to be true to life as it really happened, as it was really experienced.

Imagination in nonfictional prose

The imagination at work in nonfictional prose, however, would hardly deserve the august name of "primary imagination" reserved by the 19th-century English poet Samuel Taylor Coleridge to creators who come close to possessing semidivine powers. Rather, imagination is displayed in nonfictional prose in the fanciful invention of decorative details, in digressions practiced as an art and assuming a character of pleasant nonchalance, in establishing a familiar contact with the reader through wit and humour. The variety of themes that may be touched upon in that prose is almost infinite. The treatment of issues may be ponderously didactic and still belong within the literary domain. For centuries, in many nations, in Asiatic languages, in medieval Latin, in the writings of the humanists of the Renaissance, and in those of the Enlightenment, a considerable part of literature has been didactic. The concept of art for art's sake is a late and rather artificial development in the history of culture, and it did not reign supreme even in the few countries in which it was expounded in the 19th century. The ease with which digressions may be inserted in that type of prose affords nonfictional literature a freedom denied to writing falling within other genres. The drawback of such a nondescript literature lies in judging it against any standard of perfection, since perfection implies some conformity with implicit rules and the presence, however vague, of standards such as have been formulated for comedy, tragedy, the ode, the short story and even (in this case, more honoured in the breach than the observance) the novel. The compensating grace is that in much nonfictional literature that repudiates or ignores structure the reader is often delighted with an air of ease and of nonchalance and with that rarest of all virtues in the art of writing: naturalness.

Style. The writing of nonfictional prose should not entail the tension, the monotony, and the self-conscious craft of fiction writing. The search for *le mot juste* ("the precise word") so fanatically pursued by admirers of Flaubert and Maupassant is far less important in nonfictional prose than in the novel and the short story. The English author G.K. Chesterton (1874–1936), who was himself more successful in his rambling volumes of reflections and of religious apologetics than in his novels, defined literature as that rare, almost miraculous use of language "by which a man really says what he means." In essays, letters, reporting, and narratives of travels, the author's aim is often not to overpower his readers by giving them the impression that he knows exactly where he is leading them, as a dramatist or a detective-story writer does. Some rambling casualness, apparently irrelevant anecdotes, and suggestions of the conclusions that the author wishes his readers to infer are often more effective than extreme terseness.

The role of eloquence

There is also another manner of writing that is more attentive to the periodic cadences and elegance of prose, in the style of the ancient Roman orator Cicero. The 19th-century English essayist William Hazlitt praised the felicities of style and the refinements of the prose of the British statesman Edmund Burke (1729–97) as "that which went the nearest to the verge of poetry and yet never fell over." A number of English writers have been fond of

that harmonious, and rhetorical prose, the taste for which may well have been fostered not only by the familiarity with Cicero but also by the profound influence of the authorized version of the Bible (1611). Martin Luther's translation of the New Testament (1522) and of the Old Testament (1534) likewise molded much of German prose and German sensibility for centuries.

In the 20th century that type of prose lost favour with American and British readers, who ceased to cherish Latin orators and Biblical prose as their models. In German literature, however, in which harmonious balance and eloquence were more likely to be admired, and in other languages more directly derived from Latin, a musical style, akin to a prolonged poem in prose, was cultivated more assiduously, as exemplified in Italian in the writings of Gabriele D'Annunzio, in French in those by André Gide, and in German in *Die Aufzeichnungen des Malte Laurids Brigge (The Notebooks of Malte Laurids Brigge)* by the poet Rainer Maria Rilke. Such an elaborate style appears to be more easily tolerated by the readers in nonfictional writing, with its lack of cumulative continuity and, generally speaking, its more restricted size, than in novels such as Pater's *Marius the Epicurean* (1885) and occasionally in Thomas Mann's fiction, in which such a style tends to pall on the reader. Similarly, it is easier for the nonfictional prose writer to weave into his style faint suggestions of irony, archaisms, alliterations, and even interventions of the author that might prove catastrophic to credibility in fiction. Critics have argued that too close attention to style was harmful to the sweep necessary to fiction: they have contended that many of the greatest novelists, such as Dickens, Balzac, Dostoyevsky, and Zola at times "wrote" badly; assuredly, they treated language carelessly more than once. Essayists, historians, orators, and divines often affect a happy-go-lucky ease so as to put them on the same footing with the common reader, but they realize that language and style are vital. They must know what resources they can draw from vivid sensations, brilliant similes, balanced sentences, or sudden, epigrammatic, effects of surprise.

Author presence. The one feature common to most authors of nonfictional prose (a few staid historians and even fewer philosophers excepted) is the marked degree of the author's presence in all they write. That is to be expected in epistolary literature, and, although less inevitably, in the essay, the travel book, journalistic reporting, and polemical or hortatory prose. Although the 17th-century French religious philosopher Pascal hinted that "the ego is hateful," the author's presence is still strongly felt. This presence endows their works with a personal and haunting force that challenges, converts, or repels, but hardly ever leaves the reader indifferent. Saint Paul's epistles owe their impact—perhaps second to none in the history of the Western World—to the self that vehemently expresses itself in them, showing no concern whatever for the niceties of Attic prose. In the treatises, discourses, and philosophical argumentation of the great writers of the Enlightenment, such as Voltaire, Diderot, and Rousseau, they frequently resort to the first person singular, which results in a vivid concreteness in the treatment of ideas. To think the abstract concretely, a precept reminiscent of the 18th-century philosophers, was also the aim of the 20th-century philosophers Jean-Paul Sartre and Maurice Merleau-Ponty when they naturalized Existentialist thought in France. The growth of personal literature in its myriad shapes is one of the striking features of modern literary evolution.

The very fact that the writer of nonfictional prose does not seek an imaginary projection to impart his vision, his anguish, and his delights to readers also underlines the nature of his intention. A school of critics has vigorously attacked "the intentional fallacy," which leads biographers and some literary historians to ask what an artist intended before evaluating the completed work of art. But in a work of apologetics or of homiletics, in a work of history or of sociology, in a critical or even in a desultory and discursive essay, and certainly in aphorisms or maxims or both, the intention of the author remains omnipresent. This intention may be disguised under the mask of a parable, under the interlocutors of a philosophical dialogue, or under the admonitions of a prophet, but the reader is never oblivious of the thinker's intent. The reader has a sophisticated enjoyment of one who shares the creator's intent and travels familiarly along with him. He respects and enjoys in those authors the exercise of an intelligence flexible enough to accept even the irrational as such.

APPROACHES

In terms of approach, that is, the attitude of the writer as it can be inferred from the writing, the distinguishing features of nonfictional prose writings are the degree of presence of the ego and of the use of a subjective, familiar tone. Such devices are also used, of course, by authors of fiction, but to a lesser extent. Similarly, the basic modes of writing—the descriptive, the narrative, the expository, and the argumentative—are found in both nonfictional literature and in fiction, but in different degrees.

The descriptive mode. In nonfictional prose, essayists, moralists, naturalists, and others regularly evoked nature scenes. The most sumptuous masters of prose composed landscapes as elaborately as landscape painters. The French writer and statesman Chateaubriand (1768–1848), for example, who was not outstandingly successful in inventing plots or in creating characters independent from his own self, was a master of description; his writings influenced the French Romantic poets, who set the impassive splendour of outward nature in contrast to the inner anguish of mortals. The 19th-century English art critic John Ruskin had a more precise gift of observation, as revealed in his descriptions of Alpine mountains and of the humblest flowers or mosses, but his ornate and sonorous prose was the climax of a high-flown manner of writing that later read like the majestic relic of another era. American nonfictional writers of the same period such as Ralph Waldo Emerson and Henry Thoreau scrupulously described the lessons of organization, of unity, and of moral beauty to be deciphered from the vicissitudes of nature. Russian essayists vied with novelists in their minute yet rapturous descriptions of the thaw releasing the torrents of spring or the implacable force of the long Northern winters. Writers more inclined to the observation of social life, in satirical sketches of the mechanically polite and artificial habitués of salons, helped the novel of social life come into existence in several Western countries.

Narrative. The narrative element is less conspicuous in writing that does not purport to relate a story than in fictional works, but there is a role for narrative in letters, diaries, autobiographies, and historical writing. Most often, an incident is graphically related by a witness, as in letters or memoirs; an anecdote may serve to illustrate a moral advice in an essay; or an entertaining encounter may be inserted into an essay or a travel sketch. Digression here represents the utmost in art; it provides a relief from the persistent attention required when the author is pursuing his purpose more seriously. Similarly, such writing provides a pleasant contrast to the rigid structure of the majority of novels since the late 19th century. In historical writing, however, simplicity and clarity of narrative are required, though it may be interspersed with speeches, with portraits, or with moral and polemical allusions. In other forms of nonfictional prose, the meandering fancy of the author may well produce an impression of freedom and of truth to life unattainable by the more carefully wrought novel. Many writers have confessed to feeling relieved when they ceased to create novels and shifted to impromptu sketches or desultory essays. The surrealist essayists of the 20th century poured their scorn on detective fiction as the most fiercely logical form of writing. In contrast, the author of essays or other nonfictional prose may blend dreams and facts, ventures into the illogical, and delightful eccentricities.

Expository and argumentative modes. The rules of old-fashioned rhetoric apply better to expository and argumentative prose than to the other modes. These rules were first set down in ancient Greece by teachers who elicited them from the smooth eloquence of Socrates, the impassioned and balanced reasoning of Demosthenes, and others. The ancient Romans went further still in codifying

figures of speech, stylistic devices, and even the gestures of the orator. Such treatises played a significant part in the education of the Renaissance Humanists, of the classical and Augustan prose writers of 17th-century England and France, of the leaders of the French Revolution in the 18th century, and even in 19th-century historians and statesmen such as Guizot in France and Macaulay and Gladstone in Britain. But the sophisticated oratory of such 18th-century British orators as Richard Brinsley Sheridan, Edmund Burke, and Charles Fox or, more recently, that of Winston Churchill, hardly seems attuned to audiences in the age of television.

It has been suggested by students of German history that Adolf Hitler, in his vituperative speeches at Nuremberg in the 1930s, fascinated the Germans because they had been unaccustomed, unlike other Western nations, to eloquence in their leaders. If a large part of a population is illiterate, such as the Cubans under Fidel Castro, unending flows of eloquence may constitute a convenient means of educating the masses. Elsewhere, a more familiar and casual type of address from political leaders tends to be preferred in an era of mass media. The gift of a superior orator has been facetiously defined as that of saying as little as possible in as many words as possible. Like sermons, many types of formal address such as lectures, political speeches, and legal pleadings appear to be doomed as documents of literary value, as Burke's or Lincoln's orations and addresses were when they were learned by heart by the younger generations and helped mold the style and contribute to the moral education of men.

THE ESSAY

In modern literatures, the category of nonfictional prose that probably ranks as the most important both in the quantity and in the quality of its practitioners is the essay.

Modern origins. Before the word itself was coined in the 16th century by Montaigne and Bacon, what came to be called an essay was called a treatise, and its attempt to treat a serious theme with consistency deprived it of the seductive charm relished in the later examples of that form of literature. In this sense, the word "essay" would hardly fit the didactic tone of Aristotle's *Rhetoric* or his *Metaphysics*. There were, however, ancient masters of an early form of the essay, such as Cicero discoursing on the pleasantness of old age or on the art of "divination"; Seneca, on anger or clemency; and Plutarch, more superficially and casually, on the passing of oracles. The relentless desire to analyze one's own contradictions, especially among Christians, who, like Saint Paul, were aware of their duality and of "doing the evil which they would not," also contributed to the emergence of the essay. But Christian writing tended to be highly didactic, as may be seen in the work of Saint Augustine of the 5th century, or of the 12th-century theologian Abélard, or even in the Latin writings on "the solitary life" or on "the scorn of the world" by the 14th-century Italian poet Petrarch. Not until the Renaissance, with its increasing assertion of the self, was the flexible and deliberately nonchalant and versatile form of the essay perfected by Montaigne.

Montaigne's pre-eminence as an essayist

Montaigne, who established the term essay, left his mark on almost every essayist who came after him in continental Europe, and perhaps even more in English-speaking countries. Emerson made him one of his six *Representative Men* along with others of the stature of Plato, Shakespeare, and Goethe. Hazlitt lauded Montaigne's qualities as precisely those that "we consider in great measure English," and another English romantic writer, Leigh Hunt, saw him as "the first man who had the courage to say as an author what he felt as a man." And the 20th-century poet T.S. Eliot declared him to be the most important writer to study for an insight into the literature of France. With Montaigne, the essay achieved for the first time what it can achieve better than any other form of writing, except perhaps the epistolary one: a means of self-discovery. It gave the writer a way of reaching the secret springs of his behaviour, of seizing the man and the author at once in his contradictions, in his profound disunity, and in his mobility. The essay was symbolic of man's new attitude toward himself, revelling in change, and hence in growth, and forsaking his age-old dream of achieving an underlying steadfastness that might make him invulnerable and similar to the gods. Now he set out to accept himself whole, with his body and his physical and behavioural pecularities, and thereby repudiate medieval asceticism. He would portray his foibles and unworthiness, hoping to rise above his own mediocrity, or, at the other extreme, he would exalt himself in the hope that he might become the man he depicted. Montaigne in his essays pursued an ethical purpose, but with no pompousness or rhetoric. He offered an ideal that was adopted by his successors for 200 years: perfecting man as a tolerant, undogmatic, urbane social being. But, unlike medieval Christian writers, he would not sacrifice to others the most dearly cherished part of himself. To others he would lend himself, but his personality and his freedom were his own, and his primary duty was to become a wiser human being.

No essayist after Montaigne touched on so many varied aspects of life with such an informal, felicitous, and brilliant style. The later writers who most nearly recall the charm of Montaigne include, in England, Robert Burton, though his whimsicality is more erudite, Sir Thomas Browne, and Laurence Sterne, and in France, with more self-consciousness and pose, André Gide and Jean Cocteau.

Uses of the essay. In the age that followed Montaigne's, at the beginning of the 17th century, social manners, the cultivation of politeness, and the training of an accomplished gentleman became the theme of many essayists. This theme was first exploited by the Italian Baldassare Castiglione in his *Il cortegiano* (1528; *The Courtier*). The influence of the essay and of genres allied to it, such as maxims, portraits, and sketches, proved second to none in molding the behaviour of the cultured classes, first in Italy, then in France, and, through French influence, in most of Europe in the 17th century. Among those who pursued this theme was the 17th-century Spanish Jesuit Baltasar Gracián in his essays on the art of worldly wisdom.

With the advent of a keener political awareness with the age of Enlightenment, in the 18th century, the essay became all-important as the vehicle for a criticism of society and of religion. Because of its flexibility, its brevity, and its potential both for ambiguity and for allusions to current events and conditions, it was an ideal tool for philosophical reformers. *The Federalist Papers* in America and the tracts of the French Revolutionaries, are among the countless examples of attempts during this period to improve the condition of man through the essay.

The advantage of this form of writing was that it was not required to conform to any unity of tone or to similar strictures assigned to other genres since it was for a long time not even considered a genre. After ponderous apologies for traditional faith failed to repulse the onslaught of deism and atheism, traditionalists of the 18th and 19th centuries, such as Burke and Coleridge, abandoned unwieldy dogmatic demonstrations in favour of the short, provocative essay. In the defense of the past, it served as the most potent means of educating the masses. French Catholics, German pietists, and a number of individual English and American authors confided to the essay their dismay at what they saw as modern vulgarity and a breakdown of the coherence of the Western tradition. Essays such as Paul Elmer More's long series of *Shelburne Essays* (published between 1904 and 1935), T.S. Eliot's *After Strange Gods* (1934) and *Notes Towards the Definition of Culture* (1948), and others that attempted to reinterpret and redefine culture, established the genre as the most fitting to express the genteel tradition at odds with the democracy of the new world.

Journalistic essays

The proliferation of magazines in the United States, and the public's impatience with painstaking demonstrations and polemics, helped establish the essay just as firmly as a receptacle for robust, humorous common sense, unpretentiously expressed, as in the writings of Oliver Wendell Holmes (1809–94). Creative writers resorted to it to admonish their compatriots when they seemed too selfishly unconcerned by the tragedies of the world. Archibald MacLeish, for instance, did so in *A Time to Speak* (1941). Lewis Mumford, Allen Tate, and other literary and social critics became crusaders for moral and spiritual reform;

others seized upon the essay for scathingly ironical and destructive criticism of their culture: for example, James Gibbons Huneker (1860–1921), an admirer of iconoclasts and of egoists, as he called them, proposed European examples to Americans he deemed to be too complacent and lethargic; and, more vociferously still, H.L. Mencken (1880–1956), a self-appointed foe of prejudices, substituted his own for those he trounced in his contemporaries.

In other new countries, or in cultures acquiring an awareness of their own ambitious identity, the essay became semipolitical, earnestly nationalistic, and often polemical, playful, or bitter. Such essays sometimes succeeded in shaking the elite out of its passivity. In Uruguay, for example, José Enrique Rodó (1872–1917), in an analogy to the characters in Shakespeare's *Tempest,* compared what should be the authentic South American to the spirit Ariel, in a work thus entitled, in contrast to the bestial Caliban, representing the materialism of North America. In Canada Olivar Asselin (1874–1937) used the essay to advocate the development of a genuine French-Canadian literature. Among the older cultures of Europe, Salvatore Quasimodo (1901–68), the Italian poet and Nobel laureate, appended critical and hortatory essays to some of his volumes of verse, such as *Il falso e vero verde* (1956; "The False and True Green"). Other European heirs to this tradition of the essay include Stefan Zweig and Hugo von Hofmannsthal in Austria and Thomas Mann and Bertolt Brecht in Germany; their sprightly and incisive essays on the arts can be traced to the 19th century German philosopher Arthur Schopenhauer.

Essays for entertainment

One of the functions of literature is to please and to entertain; and the essay, as it grew into the biggest literary domain of all, did not lose the art of providing escape. Essayists have written with grace on children, on women, on love, on sports, as in Robert Louis Stevenson's collection *Virginibus Puerisque* (1881), or Willa Cather's pleasant reflections in *Not Under Forty* (1936). Ernest Renan (1823–92), one of the most accomplished French masters of the essay, found relief from his philosophical and historical studies in his half-ironical considerations on love, and Anatole France (1844–1924), his disciple, and hosts of others have alternated playful essays with others of high seriousness. Sports, games, and other forms of relaxation have not been so often or so felicitously treated. Izaak Walton's *The Compleat Angler* (1953), however, enjoys the status of a minor classic, and the best of the modern Dutch essayists, Johan Huizinga (1872–1945), has reflected with acuteness on *Homo ludens,* or man at play. A Frenchman, Jean Prévost (1901–44), who was to die as a hero of the Resistance to the German occupation of France during World War II, opened his career as an essayist with precise and arresting analyses of the *Plaisirs des sports* (1925). But there are surprisingly few very significant works, except in chapters of novels or in short stories, on the joys of hunting, bullfighting, swimming, or even, since Anthelme Brillat-Savarin's overpraised essay, *Physiologie du goût* (1825; "The Physiology of Taste") on gourmet enjoyment of the table.

Serious speculations, on the other hand, have tended to overburden the modern essay, especially in German and in French, and to weigh it with philosophy almost as pedantic as that of academic treatises, though not as rigorous. The several volumes of Jean-Paul Sartre's *Situations,* published from 1947 on, constitute the most weighty and, in the first two volumes in particular, the most original body of essay writing of the middle of the 20th century. Albert Camus' *Mythe de Sisyphe* (1942; *Myth of Sisyphus*) and his subsequent *Homme révolté* (1951; *The Rebel*) consist of grave, but inconsistent and often unconvincing, essays loosely linked together. Émile Chartier (1868–1951), under the pseudonym Alain, exercised a lasting influence over the young through the disjointed, urbane, and occasionally provoking reflections scattered through volume after volume of his essays, entitled *Propos.*

Political essays

Apart from philosophical speculation, which most readers prefer in limited quantities, the favourite theme of many modern essays has been speculation on the character of nations. It is indeed difficult to generalize on the national temper of a nation or on the characteristics of a given culture. The authors who have done it—Emerson in his essay on *English Traits* (1856), Hippolyte Taine in his studies of the English people, Alexis de Tocqueville in his *Democracy in America* (1835, 1840)—blended undeniable conclusions with controversial assertions. Rather than systematic studies, desultory essays that weave anecdotes, intuitions, and personal remarks, ever open to challenge, have proved more effective in attempting to delineate cultures. In the 20th century, the masters of this form of writing have been among the most able in the art of essay writing: Salvador de Madariaga in Spanish, Hermann Keyserling in German, and Elie Faure in French. Some nations are much more prone than others to self-scrutiny. Several of the finest Spanish essayists were vexed by questions of what it meant to be a Spaniard, especially after the end of the 19th century when Spain was compelled to put an end to its empire. Angel Ganivet in his essay on *Idearium español* (1897; *Spain: An Interpretation*), Ortega y Gasset in *España invertebrada* (1922; *Invertebrate Spain*), and Miguel de Unamuno in almost every one of his prose essays dealt with this subject. A Spanish-born essayist, George Santayana (1863–1952), was one of the most accomplished masters of written English prose; because of his cosmopolitan culture and the subtlety of his insights, he was one of the most percipient analysts of the English and of the American character.

Laments on the decline of the essay in the 20th century have been numerous since the 1940s, when articles in most journals tended to become shorter and to strive for more immediate effect. As a result, the general reader grew accustomed to being attacked rather than seduced. Still, the 20th century could boast of the critical essays of Virginia Woolf in England, of Edmund Wilson in America, and of Albert Thibaudet and Charles du Bos in France, all of whom maintained the high standards of excellence set by their predecessors of the previous century. It is regrettable that, in the language in which the best modern essays have been written, English, the term "essay" should also have acquired the connotation of a schoolboy's attempts at elementary composition. For the essay requires vast and varied information, yet without pedantry or excessive specialization. It must give the impression of having been composed spontaneously, with relish and zest. It should communicate an experience or depict a personality with an air of dilettantism, and of love of composition, and it should make accessible to the reader knowledge and reflection and the delight of watching a fine mind at work. The essayist should possess the virtues that one of the most influential English essayists, Matthew Arnold, praised in *Culture and Anarchy* (1869): "a passion . . . to divest knowledge of all that was harsh, uncouth, difficult, abstract, professional, exclusive; to humanize it."

HISTORY

Among the ancient Greeks and Romans, history was the branch of literature in which the most expert and the most enduring prose was written. It only recovered its supreme rank in nonfictional prose in the 18th century. Earlier, however, at the beginning of the 16th century, in Florence, Italy, Niccolò Machiavelli and Francesco Guicciardini prepared the way for history to become great literature by marrying it to the nascent science of politics and by enlarging its scope to include elements of the philosophy of history. In the 18th century Voltaire, tersely and corrosively, and Edward Gibbon, with more dignity, established history again as one of the great literary arts. In the 19th century, their lessons were taken to heart, as writers and readers realized that, in Thomas Carlyle's words, "every nation's true Bible is its history." In some nations, historians, together with epic and political poets, instilled into the people a will to recover the national consciousness that had been stifled or obliterated. Macaulay's ambition, to see history replace the latest novel on a lady's dressing table, was endorsed as an eminently reasonable and beneficent ambition by scholars throughout the 19th century. After an eclipse during the first half of the 20th century, when erudition and distrust of elaborate style prevailed, the poetry of history has again been praised by the most scrupulous practitioners of that discipline.

Poetry, in that context, does not mean fiction or unfaithfulness to facts, or a mere prettification, which would be tantamount to falsification; rather, it is the recognition that, as G.M. Trevelyan, Regius professor at Cambridge University, proclaimed "The appeal of history to us all is in the last analysis poetic." Few historians today would wholly agree with the once sacrosanct formula of Leopold von Ranke (1795–1886) that their task is to record the past as it really took place. They well know that, for modern history, facts are so plentiful and so very diverse that they are only meaningful insofar as the historian selects from them, places them in a certain order, and interprets them. Since World War II, as history drew increasingly on sociology, anthropology, political and philosophical speculation, and psychoanalysis, the conviction that objectivity could be maintained by a scholar dealing with the past came to be questioned and in large measure renounced. The Italian philosopher Benedetto Croce's laconic warning that all history is contemporary history (*i.e.,* bound to the historian's time and place, hence likely to be replaced by another one after a generation) has come to be generally accepted. Nietzsche, who had sharply questioned the historical methods of his German countrymen in the 1870s, stressed the need to relate history to the present and to present it in a living and beautiful form, if it is to serve the forces of life. "You can only explain the past," he said, "by what is highest in the present."

Interpretive history

In Germany, Italy, Spain, England, America, and most of all in France, where the vogue of sheer, and often indigestible, erudition was never wholeheartedly adopted, more literary talent may have gone into historical writing than into the novel or the short story. Many reasons account for the brilliance, and the impact, of this branch of nonfictional prose. Modern man has a powerful interest in origins—of civilization, of Christianity, of the world initiated by the Renaissance, or the French Revolution, or the rise of the masses. History invites an explanation of what is in terms of its genesis, not statically but in the process of becoming. The breadth of men's curiosity has expanded significantly since the 18th century, when belief in the absolutes of religious faith tended to be supplanted by greater concern for the relative world in which men live, move, and exist. A primary factor in the increasing importance of history is the bewilderment concerning the revolutions that occurred in or threatened so many countries in the latter part of the 20th century. As fiction, philosophy, and the exact sciences failed to provide a plausible explanation, many anguished readers turned to the record of brutal change in earlier periods. The historians who addressed themselves to those immense subjects, with their myriad ramifications, often composed monumental works of a synthetical character, such as those of Arnold Toynbee or Henri Pirenne, but they also cultivated the essay. Sometimes these essays appeared as short and pregnant volumes of reflections, such as Isaiah Berlin's *Historical Inevitability* (1954), sometimes in collections of articles that first appeared in magazines.

DOCTRINAL, PHILOSOPHICAL, AND RELIGIOUS PROSE

The question of how much of doctrinal writing, dealing with faith, ethics, and philosophy, can be called literature can only be answered subjectively by each reader, judging each case on its own merits. There have been philosophers who felt in no way flattered to be included among what they considered unthinking men of letters. The prejudice lingers in some quarters that profundity and clarity are mutually exclusive and that philosophy and social sciences therefore are beyond the reach of the layman. On the other hand, many writers, while often profound and fastidiously rigorous in their thought, such as Paul Valéry, have vehemently objected to being called philosophers. Nonetheless, a vast number of philosophical works owe their influence and perhaps their greatness to their literary merits.

In periods when philosophical speculation became very abstruse, as in Germany in the 19th century, men of letters often acted as intermediaries between the highly esoteric thinkers and the public. Much of the impact of the erudite 19th-century German philosopher Georg Wilhelm Friederich Hegel was due to the more easily approachable writings of those who took issue with him, such as the Existentialist thinker Søren Kierkegaard, or to those who reinterpreted him, such as Karl Marx. Similarly, the thoughts of 20th-century German phenomenologist Edmund Husserl achieved wider circulation by receiving more literary expression in the writings of Jean-Paul Sartre. In modern Europe, the men of letters of Germany were long the most deeply imbued with abstract philosophy. After World War II, however, French writers appeared to take on a zest for abstract speculation, for turgid prose, and for the coining of abstruse terms. Much of French literature in the years after the war has been characterized as "literature as philosophy."

A very few philosophers have reached greatness by evolving a coherent, comprehensive system, ambitiously claiming to account for the world and man. Such harmonious constructions by the greatest philosophers, such as Descartes and Spinoza, might be compared to epic poems in sometimes embracing more than there actually appears to be between heaven and earth. These philosophical systems were conceived by powerful imaginative thinkers whose creative abilities were not primarily of an aesthetic order. The ability and the ambition to produce such systems has appeared in very few countries or cultures. The Slavic, the Spanish, and Spanish-American cultures have been richer in thinkers than in philosophers; that is, in men who reflected on the problems of their own country, who attempted to evolve a philosophy from history, or who applied a broad view to moral or political questions, rather than in men who constructed abstract philosophical systems.

More and more in the 20th century, the sciences that are called in some countries "social" and in others "humane" have replaced the all-encompassing philosophical systems of past ages. In Spain, Miguel de Unamuno (1864–1936) and José Ortega y Gasset (1883–1955) marked the thought and the sensibility of Spanish-speaking peoples far more than systematic philosophers might have done. Their writing, which disdains impeccable logic, is no less thought-provoking for being instinct with passion and with arresting literary effects.

Russian philosophical essays

In Russia, the doctrinal writers whose thought was most influential and often most profound were also those whose prose was most brilliant. They generally centred their speculations on two Russian preoccupations: the revival of Christian thought and charity in the Orthodox faith; and the relationship of Russia to Western Europe, branded by the Slavophiles as alien and degenerate. The consistency of ancient Greek and later Western thinkers, from Aristotle through Descartes, was of scant concern to them, but in the vitality of their style, some of these Russian theorists were masters, whose turbulent, paradoxical ideas were taken to heart by novelists, poets, and statesmen. Among these masters, Aleksandr Herzen (1812–70) combined romantic ardour and positivism, formulating a passionately Russian type of socialism; he left his mark in autobiography, political letters, fiction, and chiefly philosophy of history in *From the Other Shore* (1851). Nikolay Danilevsky (1822–85), a scientist who turned to philosophy, attempted to convince his compatriots that the manifest destiny of their country was to offer a purer and fresher ideology in lieu of that of the decadent West. V.V. Rozanov (1856–1919) was an apocalyptic prophet preaching an unusual interpretation of Christian religion; a number of his intuitions and passionate assertions are found in the novel *The Possessed* (1871–72), by Fyodor Dostoyevsky, whose own nonfictional prose is of considerable quality and conviction. The strangest and most contradictory, but also the most brilliant prose writer, among those thinkers who were torn between East and West, between a jealous Orthodox faith and the attraction of Catholic Rome, was Vladimir Solovyov (1853–1900). He blended the most personal type of visionary mysticism with an incisive humour in a manner reminiscent of Kierkegaard. His philosophical essay-dialogue-treatise, *Three Conversations on War, Progress and the End of Human History* (1900), is representative of the nonfictional Russian prose that, while not widely known outside Russia, is as revealing as the Russian novel

of the permanent contradictions and aspirations of the Slavic character.

The role of nonfictional prose in the American literature of ideas is significant, as can be seen in several of Emerson's philosophical essays and addresses; in Walt Whitman's *Democratic Vistas* (1871); in William James's pleasantly written essays on religious experience and on sundry psychological and ethical topics; in George Santayana's dexterous and seductive developments on beauty, on nature, on poets, on the genteel tradition, all envisaged with ironical sympathy. Irving Babbitt (1865–1933), Thorstein Veblen (1857–1929), and Lewis Mumford are among the many American writers who, in the 20th century, maintained the tradition of writing on abstract or moral themes with clarity and elegant simplicity. Earlier, Thomas Jefferson and Benjamin Franklin had expressed their lay philosophy in a manner they wished to be widely accessible.

In France the tradition *haute vulgarisation*—"higher vulgarization" or popularization—never died and was seldom slighted by the specialists. There, and to a slightly lesser extent in Britain, much of the most valuable writing in prose was an elucidation of the view of life underlying the creations of eminent men in many fields. Such doctrinal writing, expounding innermost convictions and sometimes representing a diversion from more intensive pursuits, constitutes a by no means negligible portion of the writings of the philosopher Bertrand Russell, of the poet William Butler Yeats, and others. The novelist or the poet may well use nonfictional prose to purge his own anger, to give vent to his vituperation against his confrères, and to relieve his imagination of all the ideological burden that might otherwise encumber it. D.H. Lawrence preserved the purity of his storyteller's art by expressing elsewhere his animadversions against Thomas Hardy or Sigmund Freud. Albert Camus stripped his fiction and short stories of the ideological musings found in his philosophical volumes. Marcel Proust succeeded in incorporating many abstract discussions of the value of art, love, and friendship in his very original and loose type of fiction; but his great work, *À la recherche du temps perdu* might well have gained even more from the excision of those dissertations and the writing of more volumes like his *Chroniques* (1927) or *Contre Sainte-Beuve* (1954). The masters of nonfictional prose in French in the 20th century have been those thinkers who were also superb stylists and who deemed it a function of philosophy to understand the aesthetic phenomenon: Henri Bergson (1859–1941), Paul Valéry (1871–1945), and Gaston Bachelard (1884–1962). No more poetical advocate of reverie has arisen in the 20th century than *La Poétique de la rêverie* (1960; *The Poetics of Reverie*) and the posthumous collection of essays, *Le Droit de rêver* (1970; "*The Right to Dream*"), by Bachelard, who was also a philosopher of science. A major influence on him, as on several earlier poets endowed with profound intellect, such as Baudelaire and Valéry, was Edgar Allan Poe, the impact of whose essays on poetics, on cosmology, and on dreams and reveries has been immense and beneficent. More than a century after his death, many of Poe's American compatriots have conceded that the storyteller and the poet in Poe counted for less, as his European admirers had divined, than the writer of critical and doctrinal prose rich in dazzling intuitions.

Religious literature

Although lectures, articles, and other prosaic admonitions have tended to take their place, sermons, funeral orations, allegories, and the visions of eternal punishment brandished by theologians constitute some of the most unforgettable prose. This form of nonfictional prose literature dates from before the Christian Era; Jewish thought and style were molded by commentaries on the Old Testament and compilations of the wisdom of the sages. Later, and more nearly literary, works of this nature include Sebastian Brant's didactic, poetical, and satirical *Narrenschiff* (1494; *Ship of Fools*), and the mystic writings of Jakob Böhme (1575–1624) in Germany, the moving sermons of Jón Vídalín (1666–1720) in Iceland. In England, Richard Baxter (1615–91) and John Bunyan (1628–88) were among the most eloquent of the 17th-century Puritans who composed doctrinal works of literary merit; along with the epic poet John Milton (1608–74), whose prose works hardly count for less than his poetry, they exercised a powerful influence on the English language through their doctrinal prose. Their contemporary, the Anglican Jeremy Taylor (1613–67), wrote the most sustained and dignified prose of an age that, on the continent, would be called Baroque. A little later, in northern Europe, the Norwegian Ludvig Holberg (1684–1754), who spent most of his life in Denmark and became best known as a comic writer, also advised his contemporaries how to live morally in his *Ethical Thoughts* and other didactic treatises. The Swede Emanuel Swedenborg (1688–1772), less gifted as a writer but far more original in his blend of mysticism and science, outshone all previous Scandinavians in impressing the imagination of other Europeans. No less influential, Søren Kierkegaard (1813–55), because of his stimulating ambiguities, his bold treatment of traditional theology and philosophy, and his extraordinary ability to write vivid, biting, and provoking prose, was, a century after his death, one of the most potent forces in the literature and thought of Western civilization.

Many 20th-century readers experience a feeling of remoteness in this kind of doctrinal writing, which stems in part from a lack of vital interest in the beliefs it embodies and from a coolness toward religious dogmatism or fanaticism. Intolerance has shifted from religion to the domain of politics. But the contemporary estrangement from that rich literary heritage is due also to a distrust of high-flown eloquence. Cotton Mather's *Essays to do good* (1710) has few readers in present-day New England, despite that region's Puritan tradition, and Jonathan Edwards (1703–58), a writer of great spiritual warmth and imaginative style who was the first of the great prose writers of America, is admired today chiefly by specialists.

A less sonorous style, one that does not ring so monotonously ornate to the reader's ears, is now preferred. In Spain, Antonio de Guevara (*c.* 1480–1545), a preacher who was at his best in his familiar and satirical moments, and St. Teresa of Avila (1515–82), in her records of her mystical ecstasies, have withstood the changing tides of taste. The French also succeeded in maintaining their appreciation of their two greatest religious writers, Pascal and Bossuet, at the very top of the nonfictional prose writers; both are still revered and occasionally imitated. Pascal took over traditional theology and treated it as literature; his unfinished *Pensées* have exercised far more influence than the rationalism of the greatest French philosophers on the sensibilities of the French. Bossuet's orations reveal the magnificent but refrigerating decorum that seems inseparable from eulogies of the dead—a genre that precludes full sincerity and cultivates tremulous emotion to a dangerous degree. Bossuet's sermons and treatises, however, include masterpieces of simple, terse, direct oratory, which show him as the majestic defender of the unity of faith, of absolutism, and of tradition. His was the last significant endeavour in the 17th century to arrest the flow of relativism and of rebellious individualism, which had engulfed Western civilization with the Renaissance, the Reformation, and Humanism. The two most brilliant writers of religious prose in France in the 20th century were Pierre Teilhard de Chardin (1881–1955), a poetical writer with a luxury of images, and Simone Weil (1909–43), more terse and restrained; they steered a middle course between dogmatism and humility in luring the lay reader to their ardent expressions of conviction.

POLITICAL, POLEMICAL, AND SCIENTIFIC PROSE

In the 20th century, political, economic, and social thought has attempted to reach scientific precision through the use of quantitative data, processing machines, and mathematical formulas. Through such means, other disciplines eventually were elevated to the status of sciences. Literature lost a great deal as a result of this scientific urge, and political and economic thought may have lost even more; for example, the ability to be understood, and perhaps applied, by men of affairs and leaders of nations. The result has been that momentous decisions may be made independent of political theory, which is more often called upon to explain them afterward. Albert Einstein

remarked that politics is much more baffling and difficult than physics, and that consequences of errors in politics are likely to make far more difference to the world than the miscalculations of science. Politics is often defined as the art of the possible; it is also an art of improvisation, since the fleeting occasions must be grasped when they appear, and risks must be taken without a full array of scientific data. Like military action, however, political action can be studied in historical writings and in the literary testimonials of men who ran the affairs of their country. Thucydides, Cicero, Caesar, Milton, Burke, Napoleon, and Jefferson were such men of action who were also endowed with uncommon literary gifts. In varying degrees, Benjamin Disraeli, Winston Churchill, Woodrow Wilson, Clemenceau, Lenin, and de Gaulle owed some of their insight and effectiveness to their literary efforts.

Authors, however, are by no means infallible in dealing with the unpredictable course of political life. Interpreting and channelling public opinion proved insuperably difficult, for example, to Alphonse de Lamartine in the revolutionary period of 1848–49 in France, to the bookish Aleksandr Kerensky during the 1917 revolution in Russia, and to a number of brilliant writers who attempted to guide the Spanish Republic in the 1930s. Crowds often can be moved more readily by vapid, repetitious, or inflammatory speeches than by profound or wise counsel. Abraham Lincoln's Gettysburg Address, Churchill's speeches during Britain's finest hour early in World War II, and de Gaulle's lofty eloquence regarding the crises of three decades in France were admired less when they were delivered than afterward. As they are collected, studied, and engraved in the mental make-up of millions of future citizens, such speeches have an effectiveness second to no other form of nonfictional prose. Novels may exercise immense influence through the acute social criticism they embody, but their impact upon the sensibility and the behaviour of their readers is probably less than that of political prose.

Spanish political essays

Although the Spanish language cannot boast of any political thinker comparable to Plato, Machiavelli, or Rousseau, it may boast a large number of fine writers on political topics. Generally, these writers reveal a restrained and terse style, like the poets of Spain, the Latin country least addicted to inflation of language. Garcilaso de la Vega (1539–1616), the son of an Inca mother, wrote with courage and talent of the Peruvians and other cultures of the New World cruelly wrecked by their Catholic conquerors. The Argentinian Domingo Faustino Sarmiento (1811–88) fought in battle and with his pen against his country's dictator and left a masterpiece of social insight, written with rare effectiveness, *Facundo* (1845). Miguel Ángel Asturias (1899–1974), from Guatemala, scathingly depicted the evils of dictatorship in Central America. Like many others in South America, where versatility is not uncommon, Francisco de Miranda (1750–1816) of Venezuela was both a political writer and a statesman.

Italy, after Machiavelli, failed to produce political writers of very great eminence, even during its liberation and unification in the 19th century. The universal thinker Benedetto Croce (1866–1952), however, had the courage to publish, during the Fascist era, the most impassioned defense of liberty in volumes such as *La Storia come pensiero e come azione* (1938; *History as the Story of Liberty*). Another Italian, but from another political direction, Antonio Gramsci (1891–1937), one of the most intelligent exponents of Communism in western Europe, was aware of the vital significance of literary form to spread political ideas. He bitterly deplored the lack, in his country, of a popular literature that reflected the morality and sentiment of the people.

In France political speculation was more comprehensive: few political theoricians have proved as influential as the philosophers of the Enlightenment, especially Montesquieu and Rousseau. It was the good fortune of the French that during their Revolution at the end of the 18th century and throughout the 19th century, its keenest political minds were also writers of admirable prose. Tocqueville's observations became a sacred text for many a student of America and of pre-Revolutionary France. Since the French seldom give ideas serious consideration unless they are well expressed, however, it was a misfortune that most political speculation after the Napoleonic age was written by gifted, often brilliant, conservatives, such as Joseph de Maistre, Auguste Comte, Frédéric Le Play, Renan, Taine, and Charles Maurras. Those advocating a socialistic view, such as Jean Jaurès and the more elegant and genteel Léon Blum, failed to express their theories in classic prose. The level of political comment in the magazines and newspapers in France is consistently high, but the writers tend to be either too clear-sighted or too arrogant to grant their statesmen a chance to act. "Fair play" is an untranslatable phrase in French, and politics in France, unlike some other countries, is never regarded as a game or sport. Rather, it is a passionate affair of the heart and intellect, conducted in a mood of intransigence. The English essayist Walter Bagehot (1826–77), observing the French at the time of the 1851 Coup d'Etat, commented wryly that "the most essential quality for a free people, whose liberty is to be progressive, permanent and on a large scale, is much stupidity .... Stupidity is nature's favorite resource for preserving steadiness of conduct and consistency of opinion."

English and American political works, from the 17th century on, excel all others; they constitute the richest form of nonfictional prose in the English language. John Milton's *Areopagitica* (1644) and his other political pamphlets are monuments of political prose that survive to this day as classics. Edmund Burke's *Letter to a Noble Lord* (1796) was praised a century and a half after its composition as the greatest piece of invective in the English language. William Godwin's *Political Justice* (1793) does not compare in the majesty of its prose to those supreme models, but it did inflame Shelley and other men of letters of the time. Walter Bagehot wrote equally well on literature, politics, and economics, and *The Economist,* which he founded, was the best written weekly of its kind in any language. John Stuart Mill and Thomas Carlyle also helped to maintain the tradition of political and social thought expressed as literature through the 19th century.

The decline of polemics

Polemical prose significantly declined in the modern era. Few moderns express the rage for invective seen in the verse of satirists such as the ancient Roman Juvenal, or Alexander Pope in 17th-century England, or even in the writings of Christian disputants such as Martin Luther. Voltaire rejoiced in flaying not only his enemies but also some, such as Montesquieu and Rousseau, who were fundamentally in agreement with him in the fight against the religion of his age. Literary polemics of a high order were employed against the cultural imperialism of the French in Gotthold Lessing's *Hamburgische Dramaturgie* (1767–69; *Hamburg Dramaturgy*). Besides these examples, the polemics of more recent periods seems tame, or else gross and venomous. Later practitioners of the literature of insult include Émile Zola, particularly in his celebrated article on the Dreyfus affair, *J' Accuse* (1898). Later writers, however, often overreach themselves; their rhetoric sounds vapid and their epigrams strained.

The rift between the two cultures, scientific and humanistic, is probably not as pronounced or final as it has been alleged to be. Around the time the division was enunciated, in mid 20th-century, it was possible to point to a number of eminent scientists who were also masters of prose writing—Henri Poincaré, Jean Rostand, and Gaston Bachelard in France; Bertrand Russell and Alfred North Whitehead in England; and René Dubos and Robert Oppenheimer in the U.S. The peril for scientists who undertake to write for laymen appears to lie in a temptation to resort to florid language and to multiply pretentious metaphors and elaborate cadences in their prose. Some scientists who wrote on astronomy, on anthropology, and on geology have not altogether escaped that pitfall: Sir James Jeans, Loren Eiseley, Sir James Frazer, Teilhard de Chardin. The marriage of the "two cultures" in one mind, which was no less concerned with scientific truth than with beauty of form, was found frequently in older times; Aristotle, Hippocrates, Galileo, Newton, and Goethe all showed strong interest in both. The popularization of science reached a level of a lucid and elegant art with the

writings of Bernard de Fontenelle (1657–1757) in French, Francesco Algarotti (1712–64) in Italian, and later, with a masterpiece of scientific rigour expressed in flexible and precise prose, *Introduction à l'étude de la médecine expérimentale,* by the physiologist Claude Bernard (1813–78).

OTHER FORMS

Reportage. Journalism often takes on a polemical cast in countries in which libel laws are not stringent. Polemical journalism flourished in continental Europe when a journalist's insults could be avenged only in a duel; one of the great journalists of this heroic era of the press in France, Armand Carrel, died in such a duel with another journalist in 1836. Most journalistic literature, however, deserves none of the ill-repute that is associated with its more polemical expressions. Rather, it is a remarkably elastic form, as adaptable to sarcasm and the puncturing of illusions as to reflection, subtle persuation, and infectious geniality. Among the eminent writers who explored its possibilities in the 18th century, Joseph Addison and Richard Steele offered models of polished English prose in the journals *The Tatler* and *The Spectator,* and Jonathan Swift and Oliver Goldsmith also used it effectively in England. In France Voltaire, the novelist Abbé Prévost, and the dramatist Pierre-Carlet de Marivaux all found effective use for the form. By the 19th century, most eminent men of letters attempted to broaden their audiences by means of articles and essays in the press, and in the 20th century, the influence of journalism pervaded the most important works of some authors. Some of the works of G.B. Shaw and H.G. Wells, for example, were reminiscent of journalism in the manner in which they sought topical controversy and challenged social and political prejudices. Many of the finest essays of Virginia Woolf, John Middleton Murry, and Aldous Huxley represented British literary journalism at its most intelligent level. In America, the more heterogeneous public to which authors must address themselves and, later, the competition of the audiovisual media, were not propitious to the flowering of literary journalism of that type. In a more ephemeral genre, that of political reflections couched in clear, pungent style, Walter Lippmann composed models of commentaries on politics and ethics.

Ego in journalistic style

The more self-centred and passionate writers seldom succeeded in journalistic prose as well as those who could forget their ego and adapt their style to a public that wanted to be entertained, moved, or convinced, perhaps, but whose attention span extended no further than the 15 minutes of a train ride or of a hurried breakfast. In France, Proust dreamt for years of appearing as a journalist on the first column of the journal *Le Figaro.* But he and his contemporaries Gide, Claudel, and Valéry, and, later, the imperious and nervous André Malraux, did not conform to the limitations of the newspaper article. On the other hand, Colette, Paul Morand, and François Mauriac proved conspicuously successful in writing the brief, gripping, taut article dear to readers of many of the better continental dailies and weeklies.

The insidious appeal of journalistic writing to thinkers, novelists, and poets is similar to the siren charm of conversation for the author who enjoys talking brilliantly at dinner parties. As Oscar Wilde ruefully remarked, conversationalists and journalists, intent on reporting on the ephemeral, pour whatever genius is theirs into their lives, and only their talent into their works.

Aphorisms and sketches. Authors of maxims and aphorisms, on the contrary, strive for the brevity of inscriptions on medals and public buildings and for a diamond-like resistance to the devastation of time upon diffuse and padded writing. This form is periodically revived. In modern letters, in the latter half of the 20th century, a condensed and enigmatic sort of prose was preferred to poetry by several poets, who invested their sensations, their illuminations, or their reflections with the mystery and éclat of aphorisms. Among the French, who have always favoured the maxim for philosophical, psychological, and ethical advice, a great poet, René Char, came to be more and more fascinated by that epigrammatic form, harking back to the ancient Greek philosopher whom he admired most, Heracleitus. Char found in the aphorism a means of "pulverizing language" and of allowing the isolated words or groups of words, freed from rhetoric and from the exigencies of clarity, to emerge like rocks from a sunken archipelago. Other French prose writers, including Camus, Char's warmest admirer, and Malraux, likewise scattered through their prose works striking aphorisms that summed up the sense of a situation or the experience of a lifetime. French novels, from the 18th century through the 20th, reflect the influence of the unforgettable maxims coined by the 17th-century moralists Pascal, La Rochefoucauld, and La Bruyère. The novelist could never long resist the seduction of brevity, the challenge of condensing wisdom into a neat, usually bitter, formula, which usually suggested to the reader not to expect overmuch from life and to take revenge upon its little ironies by denouncing it in advance.

Maxims and other pointed and epigrammatic phrases of the sort the ancient Romans called *sententiae* can become too sophisticated or can too obviously strive for effect. This form of expression reached its point of perfection, balancing profundity and solidity of content with pointedness of form, with the moralists of the 17th and 18th centuries in France, whom Nietzsche ranked above all other writers. They included Pascal and La Rochefoucauld and, later, Sébastien Chamfort (1740/41–94), a satirical pessimist often quoted by Schopenhauer and Joseph Joubert (1754–1824). This form, even more than poetry, represents the most economical means of communicating long experience and for imparting moral advice. In a very few words, or at most a few lines, an aphorism may enclose enough matter for the plot of a novel. It may trounce the prejudices of snobbery more vigorously than a long, meandering novel of manners. The greatest of the 19th-century poets, Goethe, Novalis, Leopardi, Vigny, and Baudelaire, as well as painters such as Delacroix, Cézanne, Degas, and later Braque, cherished the epigrammatic, incisive form of expression. One of the advantages of the aphorism or *pensée* is that it can easily produce an impression of depth when it may be only a commonplace pungently expressed. Another is that it allows several approaches to a subject by the skilled prose writer. If he is of a fiery temperament, prone to enthusiasms and lashing out in wrath against what he deems to be false, he can, like Nietzsche, embrace contradictions and sponsor opposed attitudes. If Epictetus, Pascal, and Nietzsche had expressed their reflections consistently and systematically, their works would probably be forgotten. Nonetheless, as Pascal shrewdly remarked, the aphoristic prose style is, of all the manners of writing, the one that engraves itself most lastingly in the memories of men.

That form, in verse and in prose, probably constitutes the most widespread form of literature. It is found in many nations that long lived without fiction, epics, or even popular poetry. It is found in ancient sayings that interlard the speeches of the 20th-century leaders both of the U.S.S.R. and of China; in the book of Proverbs of the Bible; in the Qur'ān; in the Afrikaans language of South Africa in the 20th-century writings of J. Langenhoven. Proverbs, maxims, riddles, and even conundrums make up a large part of black African folklore. The animal tales of these people also provide lessons in the form of aphorisms that are neither as platitudinous nor as didactic as Aesop's fables.

Portraits and character sketches

Portraits and sketches are a form of literature that thrives in cultures in which the court, the salon, or the café plays an important role. The few examples left by the ancient Greeks, such as by Theophrastes, pale beside the vivid portraits of real individuals drawn by the ancient Roman historian Tacitus and by the impassioned orator Cicero. In the Classical age of 17th-century France, the character sketch was cultivated in the salons and reached its summit with La Bruyère. That form of writing, however, suffered from an air of artificiality and of virtuosity. It lacked the ebullience and the imagination in suggesting telltale traits that characterize the portraits of the duc de Saint-Simon (1675–1755). Collections of sketches and characters, however, tend to strike the reader as condescending

and ungenerous insofar as the writer exempts himself of the foibles he ridicules in others.

Humorous essays

The humorous article or essay, on the other hand, is a blend of sympathy and gentle pity with irony, a form of criticism that gently mocks not only others but the mocker himself. Humour strikes deep roots in the sensibility of a people, and each nation tends to feel that its own brand of humour is the only authentic one. Its varieties of humorous writing are endless, and few rules can ever be formulated on them. Humorous literature on the highest literary level includes that of Cervantes in Spain, of Sterne, Lamb, and Thackeray in England, of Jean Paul in Germany, and of Rabelais, Montaigne, and Voltaire in France. Romantic authors have, as a rule, been too self-centred and too passionate to acquire the distance from their own selves that is essential to humour. In the 20th century, some of the most original examples of what has been called the "inner-directed smile" are in the works of the Argentine Jorge Luis Borges and by one of the writers he admires most, the English essayist G.K. Chesterton (1874–1936). In both writers, and in other virtuosos of the intellectual fantasy, there is a persistent refusal to regard themselves as being great, though greatness seems to be within their reach. The humorist will not take himself seriously. Chesterton hides the depth of his religious convictions, while Borges facetiously presents his prodigious erudition and indulges in overelaborate and flowery prose. Borges likes to put on and take off masks, to play with labyrinths and mirrors, but always with a smile. By sketching what appear to be fanciful portraits rather than overtly fictional stories, he creates a half-imaginary character whose presence haunts us in all his writings—that of the author himself.

Dialogues. The dialogue form has long been used as a vehicle for the expression of ideas. It is especially cherished by authors eager to eschew the forbidding tone of formality that often accompanies the expression of serious thought. The writer of a dialogue does not directly address his public, but instead revels in the multiple facets of ideas. By playing this dialectical game he can appear to present contrary views as their respective proponents might and then expose the errors of those he opposes, leading the readers to accept his own conclusions. The advantages of the dialogue are clear: ideas that might have remained abstruse and abstract become concrete and alive. They assume dramatic force. A constant element in the dialogue is irony; etymologically, the term derives from a form of interrogation in which the answer is known beforehand by the questioner. The earliest models of the genre, by the ancient Greeks Plato and Lucian, have never been excelled. Sophistry is another element of the dialogue. In Plato and in the dialogues of Pascal's *Provinciales* (1656–57; "Provincial Letters"), the protagonist plays with the naiveté of his opponents, who always end by surrendering. The writer of a dialogue cannot affect the same casual and self-indulgent attitude as the author of a personal essay since the characters and their statements must be plausible. Nor can he pursue an argument consistently, as he might in a critical, historical, or philosophical essay. Something must persist in the dialogue of the spontaneity and the versatility of an actual conversation among witty and thoughtful people.

There was much seriousness and occasionally some pedantry in early dialogues in several literatures. The dialogues of Bardesanes (154–222) in Syriac, rendered into English as *On Fate,* are on the subject of the laws of the country. A hundred years earlier, Lucian, who was also Syrian, proved himself a master of flowing and ironical Greek prose in his satirical dialogues. The Italian Renaissance writer Pietro Aretino (1492–1556) proved himself the equal of Lucian in verve in his *Dialogues* using the same mold and the same title as Lucian. Others who used the dialogue form included Castiglione and Pietro Bembo (1470–1547) in Italy; and in Spain Juan Luis Vives (1492–1540), León Hebreo (1460–*c.* 1521), and Juan de Valdés (*c.* 1500–41), who treated questions of faith and of languages in dialogues. The genre flourished in the 18th century: Lessing, Diderot, and the Irish philosopher George Berkeley. Diderot's works largely consist of sprightly, rambling, and provocative discussions between the various aspects of his own remarkable mentality. Bold conjectures, determined onslaughts on prejudices, insights into physiology and biology, and erotic fantasies all enter into his dialogues. In the 19th century a number of complex literary personalities, who were capable of accepting the most diverse, and even conflicting points of view, such as Renan and Valéry, had a predilection for the dialogue. Among the devices used by authors of dialogue—many of whom lacked the sustained inventiveness required by fiction—was to attribute their words to the illustrious dead. The French prelate Fénelon, for example, composed *Dialogues des morts* (1700–18), and so did many others, including the most felicitous master of that prose form, the English poet Walter Savage Landor, in his *Imaginary Conversations* (1824) and *Pentameron* (1837).

Travel and epistolary literature. The literature of travel has declined in quality in the age when travel has become most common—the present. In this nonfictional prose form, the traveller himself has always counted for more than the places he visited, and in the past, he tended to be an adventurer or a connoisseur of art, of landscapes, or of strange customs who was also, occasionally, a writer of merit. The few travel books by ancient Greek geographers, such as Strabo and Pausanias of the 1st and 2nd centuries AD, are valuable as a storehouse of remarks on ancient people, places, and creeds. Travel writing of some literary significance appears in the late-13th-century writings of Marco Polo. Works of a similar vein appeared in the 17th century in the observations of Persia two French Huguenots, Jean-Baptiste Tavernier and Jean Chardin, whose writings were lauded by Goethe. Many books of documentary value were later written by English gentlemen on their grand tour of the Continent. The 18th-century Italian egotist Casanova and his more reliable and sharper compatriot Giuseppe Baretti (1719–89) also produced significant travel writings.

The form comprises many of the finest writings in prose during the Romantic age. Not only were the Romantics more alive to picturesqueness and quaintness but also they were in love with nature. They were eager to study local colours and climates and to depict them in the settings for their imaginative stories. Also, travel gave the Romantic writer the illusion of flight from his wearied self. The leisurely record of Goethe's journey to Italy in 1786–88 counts more readers than most of his novels. *Pismo russkogu puteshestvennika* (1791–92; Eng. trans., *Letters of a Russian Traveler, 1789–1790,* 1957) by Nikolay Karamzin is one of the earliest documents in the development of Russian Romanticism. Ivan Goncharov (1812–91), the Russian novelist who stubbornly limited his fiction to his own geographical province, recorded in *Frigate Pallas* his experience of a tour around the world. Nowhere else in the whole range of literature is there anything comparable to *Peterburg* (1913–14), by a virtuoso of poetic style, Andrey Bely; it is a travel fantasy within a city that is both real and transfigured into a myth. Neither James Joyce's Dublin nor Balzac's Paris is as vividly recreated as the former Russian capital in Bely's book. Other travel writers of note include the multinational Lafcadio Hearn (1850–1904), who interpreted Japan with sensitivity and insight. Earlier, two other Westerners wrote on Asia, the English historian Alexander W. Kinglake (1809–91), in *Eothen* (1844), and, more incisively, the French diplomat Joseph-Arthur, comte de Gobineau (1816–82); both blended a sense of the picturesqueness of the East with shrewdness in the interpretation of the people. One of the most thoughtful and, in spite of the author's excessive self-assurance, most profound books on Asia is *Das Reisetagebuch eines Philosophen* (1919; *Travel Diary of a Philosopher*), by the German thinker Hermann Keyserling (1880–1946). With an insatiable interest in countries, Keyserling also interpreted the soul of South America and, less perceptively, analyzed the whole spectrum of European nations. Among the thousands of travel books on Italy, there are a few masterpieces of rapturous or humorous prose: in English, the writings of D.H. Lawrence on Sardinia, on Etruscan Italy, and on the Italian character are more lucid and less strained than other of his prose cogitations. Venice, "man's most beautiful artifact," as Bernard Beren-

son called it, inspired Rousseau, Chateaubriand, Maurice Barrès, Anatole France, and hundreds of other Frenchmen to write some of their finest pages of prose. After World War I, there was a distinct yearning for new possibilities of salvation among war-ridden Europeans, dimly descried in Asia, in Russia, or in America, and travel literature assumed a metaphysical and semireligious significance. The mood of the writers who expressed this urge was somewhat Byronic; they were expert at poetizing the flight from their own selves. Blaise Cendrars (1887–1961) in his novel *Emmène-moi au bout du monde* (1956; "Take Me Away to the End of the World"), epitomizes the urge to seek adventures and a rediscovery of oneself through strange travels. The very theme of travel, of the protagonist being but a traveller on this earth, has been, from Homer's *Odyssey* onward, one of the most laden with magical, and symbolical, associations in literature. Countless authors have played moving and delicate variations on it.

The letter as a genre

Of all the branches of nonfictional prose, none is less amenable to critical definition and categorization than letter writing. The instructions of the ancient grammarians, which were repeated a thousand times afterward in manuals purporting to teach how to write a letter, can be reduced to a few very general platitudes: be natural and appear spontaneous but not garrulous and verbose; avoid dryness and declamatory pomp; appear neither unconcerned nor effusive; express emotion without lapsing into sentimentality; avoid pedantry on the one hand and banter and levity on the other. Letters vary too much in content, however, for generalizations to be valid to all types. What is moving in a love letter might sound indiscreet in a letter of friendship; an analysis of the self may fascinate some readers, while others prefer anecdotes and scandal. La Bruyère, at the end of the 17th century, remarked that women succeed better than men in the epistolary form. It has also been claimed that a feminine sensibility can be seen in the letters of the most highly acclaimed male masters of this form, such as Voltaire, Mirabeau, Keats, and Baudelaire. Advice to practitioners of the art of letter writing usually can be expressed in the often-quoted line in Shakespeare's *Hamlet:* "To thine own self be true." The English biographer Lytton Strachey (1880–1932), a copious and versatile letter writer himself, wrote: "No good letter was ever written to convey information, or to please its recipient: it may achieve both those results incidentally; but its fundamental purpose is to express the personality of the writer." There are, however, numerous and even contradictory ways of expressing that personality.

Although critics have issued endless disquisitions on the craft of fiction and other genres, they have generally remained silent on the epistolary genre, though it has sometimes been the form of prose that outlives all others. Ever since the expression of the writer's personality became one of the implicit purposes of writing in the 18th century, the letters of such eminent authors as Diderot, Rousseau, Byron, and Flaubert have probably offered at least as much delight as any of their other writings. Impressive monuments of scholarship have been erected on the presentation of the complete letters of Thackeray, George Eliot, Swinburne, and Henry James. The literatures of France and England are notably richer in letter writing of the highest order than are the literatures of the United States and Germany. Contrary to many pessimistic predictions regarding the effect on letter writing of modern means of communication, such as the telephone, together with an apparently increasing penchant for haste, some of the richest, most revealing, and most thoughtful letters of all times have been written in the 20th century; those of the English writers Katherine Mansfield and D.H. Lawrence are paramount among them.

Personal literature. The cult of the ego (that is, a preoccupation with self-analysis) is a late development in the history of literature. There were, to be sure, men in ancient times who were absorbed in their own selves, but there is almost no autobiographical literature from ancient Greece and, in spite of Cicero and Pliny the Younger, there is little from ancient Rome. The confession, made as humble as possible and often declamatory in the exposition of the convert's repented sins, was an outgrowth of Christianity; masters of confessional literature were Saint Augustine, Petrarch, and the English Puritans. Autobiographical writing took a different form in the 18th century in the work of men who would have agreed with Goethe that personality is the most precious possession. After the publication of Rousseau's *Confessions* in France in 1781, the passion for looking into one's heart (and other organs as well) spread to other literatures of western Europe. Many a novelist thereafter kept a precise record of his cogitations, anxieties, and harrowing moments of inability to create. Poets and painters, including Delacroix, Constable, and Braque, have often done the same. There is only a very tenuous separation between fiction of this sort from nonfiction; the introspective novel in the first person singular has much in common with a diary, or a volume of personal reminiscences. In his long novel *À la recherche du temps perdu,* Proust revealed himself in three ways—as the author, as the narrator, and as the characters who are projections of his own self. An autobiography once was ordinarily written toward the end of a life, as a fond recollection or an impassioned justification of a lifetime's deeds. More and more, it has come to be written also by men and women in their prime. The names of writers whose autobiographical writings have become classics is legion. Henry Adams (1838–1918) owes his place in American letters chiefly to his book on his education; in 20th-century English letters, Osbert and Sacheverell Sitwell, Leonard Woolf, and Stephen Spender may similarly survive in literature through autobiographical works. André Gide, always uncertain of his novelist's vocation, felt more at ease laying bare the secret of his life in autobiographies and journals.

Although imaginative fiction has probably suffered from excesses of introspection and of analyses of the author's own artistic pangs, knowledge of man's inner life has been enriched by such confessions. The most profound truths on human nature, however, have been expressed not in the form of autobiography but in its transposition into fiction. Readers generally have found more truth in literature created from the possibilities of life than from the personal record of the one life that the author has lived.

In conclusion, the variety of nonfictional prose is prodigious. It can be written on almost any conceivable subject. Almost any style may be used, from casual digressions or sumptuous and sonorous sentences to sharp maxims and elliptical statements. But nonfictional prose seldom gives the reader a sense of its being inevitable, as does the best poetry or fiction. Nonfictional prose seldom can answer positively the question that Rilke and D.H. Lawrence suggest that any potential writer should ask: Would I die if I were prevented from writing? (H.M.P.)

Biographical literature

One of the oldest forms of literary expression, biographical literature seeks to recreate in words the life of a human being, that of the writer himself or of another person, drawing upon the resources, memory and all available evidences—written, oral, pictorial.

ASPECTS

Historical. Biography is sometimes regarded as a branch of history, and earlier biographical writings—such as the 15th-century *Mémoires* of the French councellor of state, Philippe de Commynes, or George Cavendish's 16th-century life of Thomas Cardinal Wolsey—have often been treated as historical material rather than as literary works in their own right. Some entries in ancient Chinese chronicles included biographical sketches; imbedded in the Roman historian Tacitus' *Annals* is the most famous biography of the emperor Tiberius; conversely, Sir Winston Churchill's magnificent life of his ancestor John Churchill, first duke of Marlborough, can be read as a history (written from a special point of view) of Britain and much of Europe during the War of the Spanish Succession (1701–14). Yet there is general recognition today that history and biography are quite distinct forms of literature. History usually deals in generalizations about a period of time (for example, the Renaissance), about a group of people

in time (the English colonies in North America), about an institution (monasticism during the Middle Ages). Biography focusses upon a single human being and deals in the particulars of his life.

Both biography and history, however, are concerned with the past, and it is in the hunting down, evaluating, and selection of sources that they are akin. In this sense biography can be regarded as a craft rather than an art: techniques of research and general rules for testing evidence can be learned by anyone and thus need involve comparatively little of that personal commitment associated with art.

A biographer in pursuit of an individual long dead is usually hampered by a lack of sources: it is often impossible to check or verify what written evidence there is; there are no witnesses to cross-examine. No method has yet been developed by which to overcome such problems. Each life, however, presents its own opportunities as well as specific difficulties to the biographer: the ingenuity with which he handles gaps in the record—by providing information, for example, about the age that casts light upon the subject—has much to do with the quality of his resulting work. James Boswell knew comparatively little about Dr. Johnson's earlier years; it is one of the greatnesses of his *Life of Samuel Johnson LL.D.* (1791) that he succeeded, without inventing matter or deceiving the reader, in giving the sense of a life progressively unfolding. Another masterpiece of reconstruction in the face of little evidence is A.J.A. Symons' biography of the English author and eccentric Frederick William Rolfe, *The Quest for Corvo* (1934). A further difficulty is the unreliability of most collections of papers, letters, and other memorabilia edited before the 20th century. Not only did editors feel free to omit and transpose materials, but sometimes the authors of documents revised their personal writings for the benefit of posterity, often falsifying the record and presenting their biographers with a difficult situation when the originals were no longer extant.

Boswell's great life of Dr. Johnson

The biographer writing the life of a person recently dead is often faced with the opposite problem: an abundance of living witnesses and a plethora of materials, which include the subject's papers and letters, sometimes reports of telephone conversations and conferences transcribed from tape, as well as the record of interviews granted the biographer by his subject's friends and associates. Frank Friedel, for example, in creating a biography of the United States president Franklin D. Roosevelt (1882–1945), has had to wrestle with something like 40 tons of paper. But finally, when writing the life of any man, whether long or recently dead, the biographer's chief responsibility is vigorously to test the authenticity of his materials by whatever rules and techniques are open to him.

Psychological. Assembling a string of facts in chronological order does not constitute the life of a person, it only gives an outline of events. The biographer therefore seeks to elicit from his materials the motives for his subject's actions and to discover the shape of his personality. The biographer who has known his subject in life enjoys the advantage of his own direct impressions, often fortified by what the subject has himself revealed in conversations, and of his having lived in the same era (thus avoiding the pitfalls in depicting distant centuries). But on the debit side, such a biographer's view is coloured by the emotional factor almost inevitably present in a living association. Conversely, the biographer who knows his subject only from written evidence, and perhaps from the report of witnesses, lacks the insight generated by a personal relationship but can generally command a greater objectivity in his effort to probe his subject's inner life.

Biographers of the 20th century have had at their disposal the psychological theories and practice of Sigmund Freud and of his followers and rivals. The extent to which these new biographical tools for the unlocking of personality have been employed and the results of their use have varied greatly. On the one hand, some biographers have deployed upon their pages the apparatus of psychological revelation—analysis of behaviour symbols, interpretation based on the Oedipus complex, detection of Jungian archetypal patterns of behaviour, and the like. Other biographers, usually the authors of scholarly large-scale lives, have continued to ignore the psychological method; while still others, though avoiding explicit psychological analysis and terminology, have nonetheless presented aspects of their subjects' behaviours in such a way as to suggest psychological interpretations. In general, the movement, since World War I, has been toward a discreet use of the psychological method, from Katherine Anthony's *Margaret Fuller* (1920) and Joseph Wood Krutch's study of Edgar Allan Poe (1926), which enthusiastically embrace such techniques, through Erik Erikson's *Young Man Luther* (1958) and *Gandhi's Truth on the Origins of Militant Nonviolence* (1969), where they are adroitly and sagaciously used by a biographer who is himself a psychiatrist, to Leon Edel's vast biography of Henry James (5 vol., 1953–72), where they are used with sophistication by a man of letters. The science of psychology has also begun to affect the biographer's very approach to his subject: a number of 20th-century authors seek to explore their own involvement with the person they are writing about before embarking upon the life itself.

Psychological interpretation of behaviour

Ethical. The biographer, particularly the biographer of a contemporary, is often confronted with an ethical problem: how much of the truth, as he has been able to ascertain it, should be printed? Since the inception of biographical criticism in the later 18th century, this somewhat arid—because unanswerable—question has dominated both literary and popular discussion of biographical literature. Upon the publication of the *Life of Samuel Johnson,* James Boswell was bitterly accused of slandering his celebrated subject. More than a century and a half later, Lord Moran's *Winston Churchill: The Struggle for Survival, 1940–1965* (1966), in which Lord Moran used the Boswellian techniques of reproducing conversations from his immediate notes and jottings, was attacked in much the same terms (though the question was complicated by Lord Moran's confidential position as Churchill's physician). In the United States, William Manchester's *Death of a President* (1967), on John F. Kennedy, created an even greater stir in the popular press. There the issue is usually presented as "the public's right to know"; but for the biographer it is a problem of his obligation to preserve historical truth as measured against the personal anguish he may inflict on others in doing so. Since no standard of "biographical morality" has ever been agreed upon—Boswell, Lord Moran, and Manchester have all, for example, had eloquent defenders—the individual biographer must steer his own course. That course in the 20th century is sometimes complicated by the refusal of the custodians of the papers of important persons, particularly national political figures, to provide access to all the documents.

The public's right to know

Aesthetic. Biography, while related to history in its search for facts and its responsibility to truth, is truly a branch of literature because it seeks to elicit from facts, by selection and design, the illusion of a life actually being lived. Within the bounds of given data, the biographer seeks to transform plain information into illumination. If he invents or suppresses material in order to create an effect, he fails truth; if he is content to recount facts, he fails art. This tension, between the requirements of authenticity and the necessity for an imaginative ordering of materials to achieve lifelikeness, is perhaps best exemplified in the biographical problem of time. On the one hand, the biographer seeks to portray the unfolding of a life with all its cross-currents of interests, changing emotional states, events; yet in order to avoid reproducing the confusion and clutter of actual daily existence, he must interrupt the flow of diurnal time and group his materials so as to reveal traits of personality, grand themes of experience, and the actions and attitudes leading to moments of high decision. His achievement as a biographical artist will be measured, in great part, by his ability to suggest the sweep of chronology and yet to highlight the major patterns of behaviour that give a life its shape and meaning.

KINDS

Biographies are difficult to classify. It is easily recognizable that there are many kinds of lifewriting, but one kind can easily shade into another; no standard basis for classifi-

cation has yet been developed. A fundamental division offers, however, a useful preliminary view: biographies written from personal knowledge of the subject and those written from research.

Firsthand knowledge. The biography that results from what might be called a vital relationship between the biographer and his subject often represents a conjunction of two main biographical forces: a desire on the part of the writer to preserve "the earthly pilgrimage of a man," as the 19th-century historian Thomas Carlyle calls it (*Critical and Miscellaneous Essays,* 1838), and an awareness that he has the special qualifications, because of direct observation and access to personal papers, to undertake such a task. This kind of biography is, in one form or another, to be found in most of the cultures that preserve any kind of written biographical tradition, and it is commonly to be found in all ages from the earliest literatures to the present. In its first manifestations, it was often produced by, or based upon the recollections of, the disciples of a religious figure—such as the biographical fragments concerning Buddha, portions of the Old Testament, and the Christian gospels. It is sometimes called "source biography" because it preserves original materials, the testimony of the biographer, and often intimate papers of the subject (which have proved invaluable for later biographers and historians—as exemplified by Einhard's 9th-century *Vita Karoli imperatoris* ["Life of Charlemagne"] or Thomas Moore's *Letters and Journals of Lord Byron* [1830]). Biography based on a living relationship has produced a wealth of masterpieces: Tacitus' life of his father-in-law in the *Agricola,* William Roper's life of his father-in-law Sir Thomas More (1626), John Gibson Lockhart's biography (1837–38) of his father-in-law Sir Walter Scott, Johann Peter Eckermann's *Conversations with Goethe* (1836; trans. 1839), and Ernest Jones's *Life and Work of Sigmund Freud* (1953–57). Indeed, what is generally acknowledged as the greatest biography ever written belongs to this class: James Boswell's *Life of Samuel Johnson.*

The relationship between writer and subject

Research. Biographies that are the result of research rather than firsthand knowledge present a rather bewildering array of forms. First, however, there should be mentioned two special kinds of biographical activity.

Reference collections. Since the late 18th century, the Western world—and, in the 20th century, the rest of the world as well—has produced increasing numbers of compilations of biographical facts concerning both the living and the dead. These collections stand apart from literature. Many nations have multivolume biographical dictionaries such as the *Dictionary of National Biography* in Britain and the *Dictionary of American Biography* in the United States; general encyclopaedias contain extensive information about figures of world importance; classified collections such as *Lives of the Lord Chancellors* (Britain) and biographical manuals devoted to scholars, scientists, and other groups are available in growing numbers; information about living persons is gathered into such national collections as *Who's Who?* (Britain), *Chi è?* (Italy), and *Who's Who in America?*

Character sketches. The short life, however, is a genuine current in the mainstream of biographical literature and is represented in many ages and cultures. Excluding early quasi-biographical materials about religious or political figures, the short biography first appeared in China at about the end of the 2nd century BC, and two centuries later it was a fully developed literary form in the Roman Empire. The *Shih-chi* ("Historical Records"), by Ssu-ma Ch'ien (145?–*c.* 85 BC), include lively biographical sketches, very short and anecdotal with plentiful dialogue, grouped by character-occupation types such as "maligned statesmen," "rash generals," "assassins," a method that became established tradition with the *Han shu* (*History of the Former Han Dynasty*), by Ssu-ma Ch'ien's successor and imitator, Pan Ku (AD 32–92). Toward the end of the first century AD, in the Mediterranean world, Plutarch's *Lives of the Noble Grecians and Romans,* which are contrasting pairs of biographies, one Greek and one Roman, appeared; there followed within a brief span of years the *Lives of the Caesars,* by the Roman emperor Hadrian's librarian Suetonius. These works established a quite subtle mingling of character sketch with chronological narrative that has ever since been the dominant mark of this genre. Plutarch, from an ethical standpoint emphasizing the political virtues of man as governor, and Suetonius, from the promptings of sheer biographical curiosity, develop their subjects with telling details of speech and action; and though Plutarch, generally considered to be the superior artist, has greatly influenced other arts than biographical literature—witness Shakespeare's Roman plays, which are based on his *Lives*—Suetonius created in the *Life of Nero* one of the supreme examples of the form. Islāmic literature, from the 10th century, produced short "typed" biographies based on occupation—saints, scholars, and the like—or on arbitrarily chosen personal characteristics. The series of brief biographies has continued to the present day with such representative collections as, in the Renaissance, Giorgio Vasari's *Lives of the Most Eminent Italian Painters, Sculptors, and Architects,* Thomas Fuller's *History of the Worthies of England* in the 17th century, Samuel Johnson's *Lives of the English Poets* in the 18th, and, in more recent times, the "psychographs" of the American Gamaliel Bradford (*Damaged Souls,* 1923), Lytton Strachey's *Eminent Victorians* (1918) and the "profiles" that have become a hallmark of the weekly magazine *The New Yorker.*

The first short biography

Further classification of biographies compiled by research can be achieved by regarding the comparative objectivity of approach. For convenience, six categories, blending one into the other in infinite gradations and stretching from the most objective to the most subjective, can be employed.

Informative biography. This, the first category, is the most objective and is sometimes called "accumulative" biography. The author of such a work, avoiding all forms of interpretation except selection—for selection, even in the most comprehensive accumulation, is inevitable—seeks to unfold a life by presenting, usually in chronological order, the paper remains, the evidences, relating to that life. This biographer takes no risks but, in turn, seldom wins much critical acclaim: his work is likely to become a prime source for biographers who follow him. During the 19th century, the *Life of Milton: Narrated in Connection with the Political, Ecclesiastical, and Literary History of his Time* (7 vol., 1859–94), by David Masson, and *Abraham Lincoln: A History* (10 vol., 1890), by John G. Nicolay and John Hay, offer representative samples. In the 20th century such works as Edward Nehls's, *D.H. Lawrence: A Composite Biography* (1957–59) and David Alec Wilson's collection of the life records of Thomas Carlyle (1923–29), in six volumes, continue the traditions of this kind of life writing.

Inevitability of selection in biography

Critical biography. This second category, scholarly and critical, unlike the first, does offer a genuine presentation of a life. These works are very carefully researched; sources and "justifications" (as the French call them) are scrupulously set forth in notes, appendixes, bibliographies; inference and conjecture, when used, are duly labelled as such; no fictional devices or manipulations of material are permitted, and the life is generally developed in straight chronological order. Yet such biography, though not taking great risks, does employ the arts of selection and arrangement. The densest of these works, completely dominated by fact, have small appeal except to the specialist. Those written with the greatest skill and insight are in the first rank of modern life writing. In these scholarly biographies—the "life and times" or the minutely detailed life—the author is able to deploy an enormous weight of matter and yet convey the sense of a personality in action, as exemplified in Leslie Marchand's *Byron* (1957), with some 1,200 pages of text and 300 pages of notes, Dumas Malone's *Jefferson and his Time* (4 vol., 1948–70), Churchill's *Marlborough* (1933–38), Douglas S. Freeman's *George Washington* (1948–57). The critical biography aims at evaluating the works as well as unfolding the life of its subject, either by interweaving the life in its consideration of the works or else by devoting separate chapters to the works. Critical biography has had its share of failures: except in skillful hands, criticism clumsily intrudes upon the continuity of a life, or the works of the subject are made to yield doubtful interpretations of character, particularly

The problems of critical biography

in the case of literary figures. It has to its credit, however, such fine biographies as Arthur S. Link, *Wilson* (5 vol., 1947–65); Richard Ellmann, *James Joyce* (1959); Ernest Jones, *The Life and Works of Sigmund Freud;* Douglas S. Freeman, *Lee* (1934–35); and Edgar Johnson, *Charles Dickens* (1952).

"Standard" biography. This third, and central, category of biography, balanced between the objective and the subjective, represents the mainstream of biographical literature, the practice of biography as an art. From antiquity until the present—within the limits of the psychological awareness of the particular age and the availability of materials—this kind of biographical literature has had as its objective what Sir Edmund Gosse called "the faithful portrait of a soul in its adventures through life." It seeks to transform, by literary methods that do not distort or falsify, the truthful record of fact into the truthful effect of a life being lived. Such biography ranges in style and method from George Cavendish's 16th-century life of Cardinal Wolsey, Roger North's late-17th-century lives of his three brothers, and Boswell's life of Johnson to modern works like Lord David Cecil's *Melbourne,* Garrett Mattingly's *Catherine of Aragon,* Andrew Turnbull's *Scott Fitzgerald,* and Leon Edel's *Henry James.*

Interpretative biography. This fourth category of life writing is subjective and has no standard identity. At its best it is represented by the earlier works of Catherine Drinker Bowen, particularly her lives of Tchaikovsky, *"Beloved Friend"* (1937), and Oliver Wendell Holmes, *Yankee from Olympus* (1944). She molds her sources into a vivid narrative, worked up into dramatic scenes that always have some warranty of documentation—the dialogue, for example, is sometimes devised from the indirect discourse of letter or diary. She does not invent materials; but she quite freely manipulates them—that is to say, interprets them—according to the promptings of insight, derived from arduous research, and with the aim of unfolding her subject's life as vividly as possible. (Mrs. Bowen, much more conservative in her later works, clearly demonstrates the essential distance between the third and fourth categories: her distinguished life of Sir Edward Coke, *The Lion and the Throne* [1957], foregoes manipulation and the "recreation" of dialogue and limits interpretation to the artful deployment of biographical resources.) Very many interpretative biographies stop just short of fictionalizing in the freedom with which they exploit materials. The works of Frank Harris (*Oscar Wilde,* 1916) and Hesketh Pearson (*Tom Paine, Friend of Mankind,* 1937; *Beerbohm Tree,* 1956) demonstrate this kind of biographical latitude.

Fictionalized biography. The books in this fifth category belong to biographical literature only by courtesy. Materials are freely invented, scenes and conversations are imagined; unlike the previous category, this class often depends almost entirely upon secondary sources and cursory research. Its authors, well represented on the paperback shelves, have created a hybrid form designed to mate the appeal of the novel with a vague claim to authenticity. This form is exemplified by writers such as Irving Stone, in his *Lust for Life* (on van Gogh) and *The Agony and the Ecstasy* (on Michelangelo). Whereas the compiler of biographical information (the first category) risks no involvement, the fictionalizer admits no limit to it.

Cursory research in fictionalized biography

Fiction presented as biography. The sixth and final category is outright fiction, the novel written as biography or autobiography. It has enjoyed brilliant successes. Such works do not masquerade as lives; rather, they imaginatively take the place of biography where perhaps there can be no genuine life writing for lack of materials. Among the most highly regarded examples of this genre are, in the guise of autobiography, Robert Graves's books on the Roman emperor Claudius, *I, Claudius* and *Claudius the God and His Wife Messalina;* Mary Renault's *The King Must Die* on the legendary hero Theseus; and Marguerite Yourcenar's *Memoirs of Hadrian.* The diary form of autobiography was amusingly used by George and Weedon Grossmith to tell the trials and tribulations of their fictional character, Charles Poster, in *The Diary of a Nobody* (1892). In the form of biography this category includes Graves's *Count Belisarius* and Hope Muntz's *Golden Warrior* (on Harold II, vanquished at the Battle of Hastings, 1066). Some novels-as-biography, using fictional names, are designed to evoke rather than re-create an actual life, such as W. Somerset Maugham's *Moon and Sixpence* (Gauguin) and *Cakes and Ale* (Thomas Hardy) and Robert Penn Warren's *All the King's Men* (Huey Long).

"Special-purpose" biography. In addition to these six main categories, there exists a large class of works that might be denominated "special-purpose" biography. In these works the art of biography has become the servant of other interests. They include potboilers (written as propaganda or as a scandalous exposé) and "as-told-to" narratives (often popular in newspapers) designed to publicize a celebrity. This category includes also "campaign biographies" aimed at forwarding the cause of a political candidate (Nathaniel Hawthorne's *Life of Franklin Pierce* [1852] being an early example); the weighty commemorative volume, not infrequently commissioned by the widow (which, particularly in Victorian times, has usually enshrouded the subject in monotonous eulogy); and pious works that are properly called hagiography, or lives of holy men, written to edify the reader.

Informal autobiography. Autobiography, like biography, manifests a wide variety of forms, beginning with the intimate writings made during a life that were not intended (or apparently not intended) for publication.

Letters, diaries, and journals. Broadly speaking, the order of this category represents a scale of increasingly self-conscious revelation. Collected letters, especially in carefully edited modern editions such as W.S. Lewis' of the correspondences of the 18th-century man of letters Horace Walpole (34 vol., 1937–65), can offer a rewarding though not always predictable experience: some eminent people commit little of themselves to paper, while other lesser figures pungently re-create themselves and their world. The 15th-century *Paston Letters* constitute an invaluable chronicle of the web of daily life woven by a tough and vigorous English family among the East Anglian gentry during the Wars of the Roses; the composer Mozart and the poet Byron, in quite different ways, are among the most revealing of letter writers. Diarists have made great names for themselves out of what seems a humble branch of literature. To mention only two, in the 20th century the young Jewish girl Anne Frank created such an impact by her recording of narrow but intense experience that her words were translated to stage and screen; while a comparatively minor figure of 17th-century England, Samuel Pepys—he was secretary to the navy—has immortalized himself in a diary that exemplifies the chief qualifications for this kind of writing—candour, zest, and an unself-conscious enjoyment of self. The somewhat more formal journal is likewise represented by a variety of masterpieces, from the notebooks, which reveal the teeming, ardent brain of Leonardo da Vinci, and William Wordsworth's sister Dorothy's sensitive recording of experience in her *Journals* (1897), to French foreign minister Armand de Caulaincourt's recounting of his flight from Russia with Napoleon (translated as *With Napoleon in Russia,* 1935) and the *Journals* of the brothers Goncourt, which present a confidential history of the literary life of mid-19th-century Paris.

Memoirs and reminiscences. These are autobiographies that usually emphasize *what* is remembered rather than *who* is remembering; the author, instead of recounting his life, deals with those experiences of his life, people, and events that he considers most significant. (The extreme contrast to memoirs is the spiritual autobiography, so concentrated on the life of the soul that the author's outward life and its events remains a blur. The artless *res gestae,* a chronology of events, occupies the middle ground.)

The nature of memoirs

In the 15th century, Philippe de Commynes, modestly effacing himself except to authenticate a scene by his presence, presents in his *Mémoires* a life of Louis XI, master of statecraft, as witnessed by one of the most sagacious counsellors of the age. The memoirs of Giacomo Casanova boast of an 18th-century rake's adventures; those of Hector Berlioz explore with great brilliance the trials of a great composer, the reaches of an extraordinary personality, and the musical life of Europe in the first part

of the 19th century. The memoir form is eminently represented in modern times by Sir Osbert Sitwell's polished volumes, presenting a tapestry of recollections that, as has been observed, "tells us little about what it feels like to be in Sir Osbert's skin"—a phrase perfectly illustrating the difference between memoirs and formal autobiography.

Formal autobiography. This category offers a special kind of biographical truth: a life, reshaped by recollection, with all of recollection's conscious and unconscious omissions and distortions. The novelist Graham Greene says that, for this reason, an autobiography is only "a sort of life" and uses the phrase as the title for his own autobiography (1971). Any such work is a true picture of what, at one moment in a life, the subject wished—or is impelled—to reveal of that life. An event recorded in the autobiographer's youthful journal is likely to be somewhat different from that same event recollected in later years. Memory being plastic, the autobiographer regenerates his materials as he uses them. The advantage of possessing unique and private information, accessible to no researching biographer, is counterbalanced by the difficulty of establishing a stance that is neither overmodest nor aggressively self-assertive. The historian Edward Gibbon declares, ". . . I must be conscious that no one is so well qualified as myself to describe the service of my thoughts and actions." The 17th-century English poet Abraham Cowley provides a rejoinder: "It is a hard and nice subject for a man to write of himself; it grates his own heart to say anything of disparagement and the reader's ears to hear anything of praise from him."

There are but few and scattered examples of autobiographical literature in antiquity and the Middle Ages. In the 2nd century BC the Chinese classical historian Ssu-ma Ch'ien included a brief account of himself in the *Shihchi,* "Historical Records." It is stretching a point to include, from the 1st century BC, the letters of Cicero (or, in the early Christian era, the letters of St. Paul); and Julius Caesar's *Commentaries* tell little about Caesar, though they present a masterly picture of the conquest of Gaul and the operations of the Roman military machine at its most efficient. The *Confessions* of St. Augustine, of the 5th century AD, belong to a special category of autobiography discussed below; the 14th-century *Letter to Posterity* of the Italian poet Petrarch is but a brief excursion in the field.

The first autobiography

Speaking generally, then, it can be said that autobiography begins with the Renaissance in the 15th century; and, surprisingly enough, the first example was written not in Italy but in England by a woman entirely untouched by the "new learning" or literature. In her old age Margery Kempe, the sobbing mystic, or hysteric, of Lynn in Norfolk, dictated an account of her bustling, far-faring life, which, however concerned with religious experience, racily reveals her somewhat abrasive personality and the impact she made upon her fellows. This is done in a series of scenes, mainly developed by dialogue. Though calling herself, in abject humility, "the creature," Margery knew, and has effectively transmitted the proof, that she was a remarkable person.

The first full-scale formal autobiography was written a generation later by a celebrated Humanist publicist of the age, Enea Silvio Piccolomini, after he was elevated to the papacy, in 1458, as Pius II—the result of an election that he recounts with astonishing frankness spiced with malice. In the first book of his autobiography—misleadingly named *Commentarii,* in evident imitation of Caesar—Pius II traces his career up to becoming pope; the succeeding 11 books (and a fragment of a 12th, which breaks off a few months before his death in 1464) present a panorama of the age, with its cruel and cultivated Italian tyrants, cynical *condottieri* (professional soldiers), recalcitrant kings, the politics and personalities behind the doors of the Vatican, and the urbane but exuberant character of the Pope himself. Pius II exploits the plasticity of biographical art by creating opportunities—especially when writing of himself as the connoisseur of natural beauties and antiquities—for effective autobiographical narration. His "Commentaries" show the art of formal autobiography in full bloom in its beginnings; they rank as one of its half dozen greatest exemplars.

The neglected autobiography of the Italian physician and astrologer Gironimo Cardano, a work of great charm, and the celebrated adventures of the goldsmith and sculptor Benvenuto Cellini in Italy of the 16th century; the uninhibited autobiography of the English historian and diplomat Lord Herbert of Cherbury, in the early 17th; and Colley Cibber's *Apology for The Life of Colley Cibber, Comedian* in the early 18th—these are representative examples of biographical literature from the Renaissance to the Age of Enlightenment. The latter period itself produced three works that are especially notable for their very different reflections of the spirit of the times as well as of the personalities of their authors: the urbane autobiography of Edward Gibbon, the great historian; the plainspoken, vigorous success story of an American who possessed all the talents, Benjamin Franklin; and the somewhat morbid introspection of a revolutionary Swiss-French political and social theorist, the *Confessions* of J.-J. Rousseau—the latter leading to two autobiographical explorations in poetry during the Romantic Movement (flourished 1798–1837) in England, Wordsworth's *Prelude* and Byron's *Childe Harold,* cantos III and IV. Significantly, it is at the end of the 18th century that the word autobiography apparently first appears in print, in *The Monthly Review,* 1797.

Autobiography in the Age of Enlightenment

Specialized forms. These might roughly be grouped under four heads: thematic, religious, intellectual, and fictionalized. The first grouping includes books with such diverse purposes as Adolf Hitler's *Mein Kampf* (1924), *The Americanization of Edward Bok* (1920), and Richard Wright's *Native Son* (1940). Religious autobiography claims a number of great works, ranging from the *Confessions* of St. Augustine and Peter Abelard's *Historia Calamitatum* (*The Story of My Misfortunes*) in the Middle Ages to the autobiographical chapters of Thomas Carlyle's *Sartor Resartus* ("The Everlasting No," "Centre of Indifference," "The Everlasting Yea") and Cardinal John Newman's beautifully wrought *Apologia* in the 19th century. That century and the early 20th saw the creation of several intellectual autobiographies. The *Autobiography* of the philosopher John S. Mill, severely analytical, concentrates upon "an education which was unusual and remarkable." It is paralleled, across the Atlantic, in the bleak but astringent quest of *The Education of Henry Adams* (printed privately 1906; published 1918). Edmund Gosse's sensitive study of the difficult relationship between himself and his Victorian father, *Father and Son* (1907), and George Moore's quasi-novelized crusade in favour of Irish art, *Hail and Farewell* (1911–14), illustrate the variations of intellectual autobiography. Finally, somewhat analogous to the novel as biography (for example, Graves's *I, Claudius*) is the autobiography thinly disguised as, or transformed into, the novel. This group includes such works as Samuel Butler's *Way of All Flesh* (1903), James Joyce's *Portrait of the Artist as a Young Man* (1916), George Santayana's *Last Puritan* (1935), and the gargantuan novels of Thomas Wolfe (*Look Homeward, Angel* [1929], *Of Time and the River* [1935]).

HISTORICAL DEVELOPMENT

Western literature. *Antiquity.* In the Western world, biographical literature can be said to begin in the 5th century BC with the poet Ion of Chios, who wrote brief sketches of such famous contemporaries as Pericles and Sophocles. It continued throughout the classical period for a thousand years, until the dissolution of the Roman Empire in the 5th century AD. Broadly speaking, the first half of this period exhibits a considerable amount of biographical activity, of which much has been lost; such fragments as remain of the rest—largely funeral elegies and rhetorical exercises depicting ideal types of character or behaviour—suggest that from a literary point of view the loss is not grievous. (An exception is the life of the Roman art patron Pomponius Atticus, written in the 1st century BC by Cornelius Nepos.) Biographical works of the last centuries in the classical period, characterized by numerous sycophantic accounts of emperors, share the declining energies of the other literary arts. But although there are few genuine examples of life writing, in the modern sense of the term, those few are masterpieces. The two greatest teachers

of the classical Mediterranean world, Socrates and Jesus Christ, both prompted the creation of magnificent biographies written by their followers. To what extent Plato's life of Socrates keeps to strict biographical truth cannot now be ascertained (though the account of Socrates given by Plato's contemporary the soldier Xenophon, in his *Memorabilia,* suggests a reasonable faithfulness) and he does not offer a full-scale biography. Yet in his two consummate biographical dialogues—*The Apology* (recounting the trial and condemnation of Socrates) and the *Phaedo* (a portrayal of Socrates' last hours and death)—he brilliantly recreates the response of an extraordinary character to the crisis of existence. Some 400 years later there came into being four lives of Jesus, the profound religious significance of which has inevitably obscured their originality—their homely detail, anecdotes, and dialogue that, though didactic in purpose, also evoke a time and a personality. The same century, the first of the Christian era, gave birth to the three first truly "professional" biographers—Plutarch and Suetonius (discussed above) and the historian Tacitus, whose finely wrought biography of his father-in-law, *Agricola,* concentrating on the administration rather than the man, has something of the monumental quality of Roman architecture. The revolution in thought and attitude brought about by the growth of Christianity is signalled in a specialized autobiography, the *Confessions* of St. Augustine; but the biographical opportunity suggested by Christian emphasis on the individual soul was, oddly, not to be realized. If the blood of the martyrs fertilized the seed of the new faith, it did not promote the art of biography. The demands of the church and the spiritual needs of men, in a twilight world of superstition and violence, transformed biography into hagiography. There followed a thousand years of saints' lives: the art of biography forced to serve ends other than its own.

Plato's accounts of Socrates

Middle Ages. This was a period of biographical darkness, an age dominated by the priest and the knight. The priest shaped biography into an exemplum of otherworldliness, while the knight found escape from daily brutishness in allegory, chivalric romances, and broad satire (the fabliaux). Nevertheless, glimmerings can be seen. A few of the saints' lives, like Eadmer's *Life of Anselm,* contain anecdotal materials that give some human flavour to their subjects; the 13th-century French nobleman Jean, sire de Joinville's life of St. Louis (Louis IX of France), *Mémoires,* offers some lively scenes. The three most interesting biographical manifestations came early. Bishop Gregory of Tours' *History of the Franks* depicts artlessly but vividly, from firsthand observation, the lives and personalities of the four grandsons of Clovis and their fierce queens in Merovingian Gaul of the 6th century. Bede's *Ecclesiastical History of the English People,* of the 8th century, though lacking the immediacy and exuberance—and the violent protagonists—of Gregory, presents some valuable portraits, like those of "the little dark man," Paulinus, who converted the King of Northumbria to Christianity.

Einhard's *Life of Charlemagne*

Most remarkable, however, a self-consciously wrought work of biography came into being in the 9th century: this was *The Life of Charlemagne,* written by a cleric at his court named Einhard. He is aware of his biographical obligations and sets forth his point of view and his motives:

> I have been careful not to omit any facts that could come to my knowledge, but at the same time not to offend by a prolix style those minds that despise everything modern .... No man can write with more accuracy than I of events that took place about me, and of facts concerning which I had personal knowledge....

He composes the work in order to ensure that Charlemagne's life is not "wrapped in the darkness of oblivion" and out of gratitude for "the care that King Charles bestowed upon me in my childhood, and my constant friendship with himself and his children." Though Einhard's biography, by modern standards, lacks sustained development, it skillfully reveals the chief patterns of Charlemagne's character—his constancy of aims, powers of persuasion, passion for education. Einhard's work is far closer to modern biography than the rudimentary poetry and drama of his age are to their modern counterparts.

Renaissance. Like the other arts, biography stirs into fresh life with the Renaissance in the 15th century. Its most significant examples were autobiographical, as has already been mentioned. Biography was chiefly limited to uninspired panegyrics of Italian princes by their court Humanists, such as Simonetta's life of the great *condottiere,* Francesco Sforza, duke of Milan.

During the first part of the 16th century in England, now stimulated by the "new learning" of Erasmus, John Colet, Thomas More, and others, there were written three works that can be regarded as the initiators of modern biography: More's *History of Richard III,* William Roper's *Mirrour of Vertue in Worldly Greatness; or, the life of Syr Thomas More,* and George Cavendish's *Life of Cardinal Wolsey.* The *History of Richard III* (written about 1513 in both an English and a Latin version) unfortunately remains unfinished; and it cannot meet the strict standards of biographical truth since, under the influence of classical historians, a third of the book consists of dialogue that is not recorded from life. However, it is a brilliant work, exuberant of wit and irony, that not only constitutes a biographical landmark but is also the first piece of modern English prose. With relish, More thus sketches Richard's character:

> He was close and secret, a deep dissembler, lowly of countenance, arrogant of heart, outwardly companionable where he inwardly hated, not hesitating to kiss whom he thought to kill.

Worked up into dramatic scenes, this biography, as reproduced in the *Chronicles* of Edward Hall and Raphael Holinshed, later provided both source and inspiration for Shakespeare's rousing melodramatic tragedy, *Richard III.* The lives written by Roper and Cavendish display interesting links, though the two men were not acquainted: they deal with successive first ministers destroyed by that brutal master of politics, Henry VIII; they are written from first hand observation of their subjects by, respectively, a son-in-law and a household officer; and they exemplify, though never preach, a typically Renaissance theme: *Indignatio principis mors est*—"the Prince's anger is death." Roper's work is shorter, more intimate, and simpler; in a series of moving moments it unfolds the struggle within Sir Thomas More between his duty to conscience and his duty to his king. Cavendish offers a more artful and richly developed narrative, beautifully balanced between splendid scenes of Wolsey's glory and vanity and ironically contrasting scenes of disgrace, abasement, and painfully achieved self-knowledge.

Biography in the Renaissance

The remaining period of the Renaissance, however, is disappointingly barren. In Russia, where medieval saints' lives had also been produced, there appears a modest biographical manifestation in the *Stepennaya Kniga* ("Book of Degrees," 1563), a collection of brief lives of princes and prelates. Somewhat similarly, in France, which was torn by religious strife, Pierre Brantôme wrote his *Lives of Famous Ladies* and *Lives of Famous Men.* The Elizabethan Age in England, for all its magnificent flowering of the drama, poetry, and prose, did not give birth to a single biography worthy of the name. Sir Fulke Greville's account of Sir Philip Sidney (1652) is marred by tedious moralizing; Francis Bacon's accomplished life of the first Tudor monarch, *The Historie of the Raigne of King Henry the Seventh* (1622), turns out to be mainly a history of the reign. But Sir Walter Raleigh suggests an explanation for this lack of biographical expression in the introduction to his *History of the World* (1614): "Whosoever, in writing a modern history, shall follow truth too near the heels, it may haply strike out his teeth"—as Sir John Hayward could testify, having been imprisoned in the Tower of London because his account (1599) of Richard II's deposition, two centuries earlier, had aroused Queen Elizabeth's anger.

17th and 18th centuries. In the 17th century the word biography was first employed to create a separate identity for this type of writing. That century and the first half of the 18th presents a busy and sometimes bizarre biographical landscape. It was an era of experimentation and preparation rather than of successful achievement. In the New World, the American Colonies began to develop a scattered biographical activity, none of it of lasting impor-

tance. France offers the celebrated *Letters* of the Marquise de Sévigné to her daughter, an intimate history of the Age of Louis XIV; numerous memoirs, such as those of Louis de Rouvroy, duc de Saint-Simon, and the acerbic ones of the Cardinal de Retz (1717); and the philosopher and critic Pierre Bayle's *Dictionnaire historique et critique* (1697), which was followed by specialized biographical collections and reference works. England saw an outpouring, beginning in the earlier 17th century, of Theophrastan "characters" (imaginary types imitated from the work of Theophrastus, a follower of Aristotle), journals, diaries, the disorganized but vivid jottings of John Aubrey (later published in 1898 as *Brief Lives*); and in the earlier 18th century there were printed all manner of sensational exposés, biographical sketches of famous criminals, and the like. In this era women appear for the first time as biographers. Lady Fanshawe wrote a life of her ambassador-husband (1829); Lucy Hutchinson, one of her Puritan warrior-husbands (written after 1664, published 1806); and Margaret Cavendish, duchess of Newcastle, produced a warm, bustling life—still good reading today—of her duke, an amiable mediocrity (*The Life of the thrice Noble Prince William Cavendishe, Duke Marquess, and Earl of Newcastle,* 1667). This age likewise witnessed the first approach to a professional biographer, the noted lover of angling, Izaak Walton, whose five lives (of the poets John Donne [1640] and George Herbert [1670], the diplomat Sir Henry Wotton [1651], and the ecclesiastics Richard Hooker [1665] and Robert Sanderson [1678]) tend to endow their diverse subjects with something of Walton's own genteel whimsicality but nonetheless create skillful biographical portraits. The masterpieces of the age are unquestionably Roger North's biographies (not published until 1742, 1744) of his three brothers: Francis, the lord chief justice, "my best brother"; the lively merchant-adventurer Sir Dudley, his favourite; and the neurotic scholar John. Also the author of an autobiography, Roger North likewise produced, as a preface to his life of Francis, the first extensive critical essay on biography, which anticipates some of the ideas of Samuel Johnson and James Boswell.

The first "professional" biographer

The last half of the 18th century witnessed the remarkable conjunction of these two remarkable men, from which sprang what is generally agreed to be the world's supreme biography, Boswell's *Life of Samuel Johnson LL.D.* (1791). Dr. Johnson, literary dictator of his age, critic and lexicographer who turned his hand to many kinds of literature, himself created the first English professional biographies in *The Lives of the English Poets.* In essays and in conversation, Johnson set forth principles for biographical composition: the writer must tell the truth—"the business of the biographer is often to . . . display the minute details of daily life," for it is these details that recreate a living character; and men need not be of exalted fame to provide worthy subjects.

For more than one reason the somewhat disreputable and incredibly diligent Scots lawyer James Boswell can be called the unique genius of biographical literature, bestriding both autobiography and biography. Early in his acquaintance with Johnson he was advised by the Doctor "to keep a journal of my [Boswell's] life, full and unreserved." Boswell followed this advice to the letter. His gigantic journals offer an unrivalled self-revelation of a fascinatingly checkered character and career—whether as a young rake in London or thrusting himself upon the aged Rousseau or making his way to Voltaire's seclusion at Ferney in Switzerland with the aim of converting that celebrated skeptic to Christianity. Boswell actively helped to stage the life of Johnson that he knew he was going to write—drawing out Johnson in conversation, setting up scenes he thought likely to yield rich returns—and thus, at moments, he achieved something like the novelist's power over his materials, being himself an active part of what he was to re-create. Finally, though he invented no new biographical techniques, in his *Life of Samuel Johnson* he interwove with consummate skill Johnson's letters and personal papers, Johnson's conversation as assiduously recorded by the biographer, material drawn from interviews with large numbers of people who knew Johnson, and his own observation of Johnson's behaviour, to elicit the living texture of a life and a personality. Boswell makes good his promise that Johnson "will be seen as he really was . . . ." The influence of Boswell's work penetrated throughout the world and, despite the development of new attitudes in biographical literature, has persisted to this day as a pervasive force. Perhaps equally important to life writers has been the inspiration provided by the recognition accorded Boswell's *Life* as a major work of literary art. Since World War II there have often been years, in the United States, when the annual bibliographies reveal that more books or articles were published about Johnson and Boswell than about all the rest of biographical literature together.

Boswell's influence on biography

19th century. The *Life of Johnson* may be regarded as a representative psychological expression of the Age of Enlightenment, and it certainly epitomizes several typical characteristics of that age: devotion to urban life, confidence in common sense, emphasis on man as a social being. Yet in its extravagant pursuit of the life of one individual, in its laying bare the eccentricities and suggesting the inner turmoil of personality, it may be thought of as part of that revolution in self-awareness, ideas, aspirations, exemplified in Rousseau's *Confessions,* the French Revolution, the philosophical writings of the German philosopher Immanuel Kant, the political tracts of Thomas Paine, and the works of such early Romantic poets as Robert Burns, William Blake, Wordsworth—a revolution that in its concern with the individual psyche and the freedom of man seemed to augur well for biographical literature. This promise, however, was not fulfilled in the 19th century.

That new nation, the United States of America, despite the stimulus of a robust and optimistic society, flamboyant personalities on the frontier, a generous share of genius, and the writing of lives by eminent authors such as Washington Irving and Henry James, produced no biographies of real importance. One professional biographer, James Parton, published competent, well-researched narratives, such as his lives of Aaron Burr and Andrew Jackson, but they brought him thin rewards and are today outmoded. In France, biography was turned inward, to romantic introspection, a trend introduced by Étienne Pivert de Senancour's *Obermann* (1804). It was followed by autobiographies thinly disguised as novels such as Benjamin Constant's *Adolphe* (1816), *La Vie de Henri Brulard* of Stendhal (Marie Henri Beyle), and similar works by Alphonse de Lamartine and Alfred de Musset, in which the emotional malaise of the hero is subjected to painstaking analysis. In Great Britain the 19th century opened promisingly with an outburst of biographical–autobiographical production, much of which came from prominent figures of the Romantic Movement, including Samuel Taylor Coleridge, Robert Southey, William Hazlitt, and Thomas De Quincey. Thomas Moore's *Letters and Journals of Lord Byron* (1830), John Gibson Lockhart's elaborate life (1837–38) of his father-in-law, Sir Walter Scott, and, later, Elizabeth Cleghorn Gaskells' *Life of Charlotte Bronte* (1857), James Anthony Froude's study of Carlyle (2 vol. 1882; 2 vol. 1884), John Forster's *Life of Charles Dickens* (1872–74) all followed, to some degree, what may loosely be called the Boswell formula. Yet most of these major works are marred by evasions and omissions of truth—though Lockhart and Froude, for example, were attacked as conscienceless despoilers of the dead—and, before the middle of the century, biography was becoming stifled. As the 20th century biographer and critic Sir Harold Nicolson wrote in *The Development of English Biography* (1927), "Then came earnestness, and with earnestness hagiography descended on us with its sullen cloud . . . ." Insistence on respectability, at the expense of candour, had led Carlyle to observe acridly, "How delicate, how decent is English biography, bless its mealy mouth!" and to pillory its productions as "vacuum-biographies."

Biography during the Victorian age

20th century. The period of modern biography was ushered in, generally speaking, by World War I. All the arts were in ferment, and biographical literature shared in the movement, partly as a reaction against 19th-century conventions, partly as a response to advances in psychology, and partly as a search for new means of expression. This

revolution, unlike that at the end of the 18th century, was eventually destined to enlarge and enhance the stature of biography. The chief developments of modern life writing may be conveniently classified under five heads: (1) an increase in the numbers and general competence of biographies throughout the Western world; (2) the influence on biographical literature of the counterforces of science and fictional writing; (3) the decline of formal autobiography and of biographies springing from a personal relationship; (4) the range and variety of biographical expression; and (5) the steady, though moderate, growth of a literature of biographical criticism. Only the first three of these developments need much elaboration.

Little has been said about biography since the Renaissance in Germany, Spain, Italy, Scandinavia, and the Slavic countries because, as in the case of Russia, there had been comparatively little biographical literature and because biographical trends, particularly since the end of the 18th century, generally followed those of Britain and France. Russian literary genius in prose is best exemplified during both the 19th and 20th centuries in the novel. In the 19th century, however, Leo Tolstoy's numerous autobiographical writings, such as *Childhood* and *Boyhood,* and Sergey Aksakov's *Years of Childhood* and *A Russian Schoolboy,* and in the 20th century, Maksim Gorky's autobiographical trilogy (*Childhood; In the World;* and *My Universities,* 1913–23) represent, in specialized form, a limited biographical activity. The close control of literature exercised by the 20th-century Communist governments of eastern Europe has created a wintry climate for biography. The rest of Europe, outside the iron curtain, has manifested in varying degrees the fresh biographical energies and practices illustrated in British–American life writing: biography is now, as never before, an international art that shares a more or less common viewpoint.

The second characteristic of modern biography, its being subject to the opposing pressures of science and fictional writing, has a dark as well as a bright side. Twentieth-century fiction, boldly and restlessly experimental, has, on the one hand, influenced the biographer to aim at literary excellence, to employ devices of fiction suitable for biographical ends; but, on the other, fiction has also probably encouraged the production of popular pseudo-biography, hybrids of fact and fancy, as well as of more subtle distortions of the art form. Science has exerted two quite different kinds of pressure: the prestige of the traditional sciences, in their emphasis on exactitude and rigorous method, has undoubtedly contributed to a greater diligence in biographical research and an uncompromising scrutiny of evidences; but science's vast accumulating of facts—sometimes breeding the worship of fact for its own sake—has helped to create an atmosphere in which today's massive, note-ridden and fact-encumbered lives proliferate and has probably contributed indirectly to a reluctance in the scholarly community to take the risks inevitable in true biographical composition.

Influence of fiction on biography

The particular science of psychology, as earlier pointed out, has conferred great benefits upon the responsible practitioners of biography. It has also accounted in large part, it would appear, for the third characteristic of modern biography: the decline of formal autobiography and of the grand tradition of biography resulting from a personal relationship. For psychology has rendered the self more exposed but also more elusive, more fascinatingly complex and, in the darker reaches, somewhat unpalatable. Since honesty would force the autobiographer into a self-examination both formidable to undertake and uncomfortable to publish, instead he generally turns his attention to outward experiences and writes memoirs and reminiscences—though France offers something of an exception in the journals of such writers as André Gide (1947–51), Paul Valéry (1957), François Mauriac (1934–50), Julien Green (1938–58). Similarly, psychology, in revealing the fallacies of memory, the distorting power of an emotional relationship, the deceits of observation, has probably discouraged biography written by a friend of its subject. Moreover, so many personal papers are today preserved that a life-long friend of the subject scarcely has time to complete his biography.

The impact of Lytton Strachey

After World War I, the work of Lytton Strachey played a somewhat similar role to that of Boswell in heading a "revolution" in biography. *Eminent Victorians* and *Queen Victoria* (1921), followed by *Elizabeth and Essex* (1928), with their artful selection, lacquered style, and pervasive irony, exerted an almost intoxicating influence in the 1920s and '30s. Writers seeking to capitalize on Strachey's popularity and ape Strachey's manner, without possessing Strachey's talents, produced a spate of "debunking" biographies zestfully exposing the clay feet of famous historical figures. By World War II, however, this kind of biography had been discredited; Strachey's adroit detachment and literary skill were recognized to be his true value, not his dangerously interpretative method; and, since that time, biography has steadied into an established, if highly varied, form of literature.

Other literatures. Biography as an independent art form, with its concentration upon the individual life and its curiosity about the individual personality, is essentially a creation of Western man. In the Orient, for all its long literary heritage, and in Islām, too, biographical literature does not show the development, nor assume the importance, of Western life writing. In China, until comparatively recently, biography had been an appendage, or by-product, of historical writing and scholarly preoccupation with the art of government, in the continuing tradition of the "Historical Records" of Ssu-ma Ch'ien and Pan Ku. In India it has been the enduring concern for spiritual values and for contemplation or mystical modes of existence that have exerted the deepest influence on literature from the first millennium BC to the present, and this has not provided a milieu suitable to biographical composition. Generally speaking, the literary history of Japan, too, offers only fragmentary or limited examples of life writing.

It was not until the beginning of the 20th century in China that biography began to appear as an independent form (and this was evidently the result of western influence), when Liang Ch'i-ch'ao (1873–1929) wrote a number of lives, including one of Confucius, and was followed by Hu Shih (1891–1962), who, like his predecessor, worked to promote biographical composition as an art form. Except for China after the establishment of the Communist state in 1949, biography in the Orient—notably in India and Japan—has shared, to a limited extent, the developments in biographical literature demonstrated in the rest of the world.

BIOGRAPHICAL LITERATURE TODAY

In the United States, Great Britain, and the rest of the Western world generally, biography today enjoys a moderate popular and critical esteem. In the year 1929, at the height of the biographical "boom," there were published in the United States 667 new biographies; in 1962 exactly the same number appeared, the population in the meantime having increased by something like 50 percent. On the average, in the English-speaking world, biographical titles account for approximately 5 percent of the annual output of books. Yet they have won their share of literary prizes and for their authors a considerable degree of literary eminence; if few universally acclaimed masterpieces are being produced, it is probably true that the art of biography is seeing a higher general level of achievement than ever before. The recreation of a life is also now being attempted in other media than that of prose. Biographical drama has of course been staged from before the time of Shakespeare; it continues to be popular, whether translated from narrative to the theatre (as the *Diary of Anne Frank*) or written specifically for the stage, like Jean Anouilh's *Becket* and Robert Bolt's study of Sir Thomas More, *A Man for All Seasons* (which nonetheless owes a great deal to William Roper). The cinema often follows with its versions of such plays; it likewise produces original biographical films, generally with indifferent success. Television, too, offers historical "recreations" of various sorts, and with varying degrees of responsibility, but has achieved only a few notable examples of biographical illumination, for the conflict between gripping visual presentation and the often undramatic, but important, biographical truth is difficult to resolve. Biography, indeed, seems less innovative, less

rewarding of experiment, and less adaptable to new media, than does fiction or perhaps even history. Words are no longer the only way to tell a story and perhaps in time will not be regarded as the chief way; but so far they seem the best way of unfolding the full course of a life and exploring the quirks and crannies of a personality. Anchored in the truth of fact, though seeking the truth of interpretation, biography tends to be more stable than other literary arts; and its future would appear to be a predictably steady evolution of its present trends. (P.M.K.)

Literary criticism

Construed loosely, literary criticism is the reasoned consideration of literary works and issues. It applies, as a term, to any argumentation about literature, whether or not specific works are analyzed. Plato's cautions against the risky consequences of poetic inspiration in general in his *Republic* are thus often taken as the earliest important example of literary criticism. More strictly construed, the term covers only what has been called "practical criticism," the interpretation of meaning and the judgment of quality. Criticism in this narrow sense can be distinguished not only from aesthetics (the philosophy of artistic value) but also from other matters that may concern the student of literature: biographical questions, bibliography, historical knowledge, sources and influences, and problems of method. Thus, especially in academic studies, "criticism" is often considered to be separate from "scholarship." In practice, however, this distinction often proves artificial, and even the most single-minded concentration on a text may be informed by outside knowledge, while many notable works of criticism combine discussion of texts with broad arguments about the nature of literature and the principles of assessing it. Criticism will here be taken to cover all phases of literary understanding, though the emphasis will be on the evaluation of literary works and of their authors' places in literary history. A longer and more theoretical discussion of the role of criticism will be found in the article ARTS, CRITICISM OF THE. One particular aspect of literary criticism is covered in the article HISTORY, THE STUDY OF: *Textual criticism.* The following article deals with criticism largely in the context of Western literature. Its role in other literatures, along with the history of the literatures themselves, is treated in articles on the arts of various peoples, such as SOUTH ASIAN ARTS; and AFRICAN ARTS.

Criticism and scholarship

FUNCTIONS

The functions of literary criticism vary widely, ranging from the reviewing of books as they are published to systematic theoretical discussion. Though reviews may sometimes determine whether a given book will be widely sold, many works succeed commercially despite negative reviews, and many classic works, including Herman Melville's *Moby Dick* (1851), have acquired appreciative publics long after being unfavourably reviewed and at first neglected. One of criticism's principal functions is to express the shifts in sensibility that make such revaluations possible. The minimal condition for such a new appraisal is, of course, that the original text survive. The literary critic is sometimes cast in the role of scholarly detective, unearthing, authenticating, and editing unknown manuscripts. Thus, even rarefied scholarly skills may be put to criticism's most elementary use, the bringing of literary works to a public's attention.

The variety of criticism's functions is reflected in the range of publications in which it appears. Criticism in the daily press rarely displays sustained acts of analysis and may sometimes do little more than summarize a publisher's claims for a book's interest. Weekly and biweekly magazines serve to introduce new books but are often more discriminating in their judgments, and some of these magazines, such as *The* (London) *Times Literary Supplement* and *The New York Review of Books,* are far from indulgent toward popular works. Sustained criticism can also be found in monthlies and quarterlies with a broad circulation, in "little magazines" for specialized audiences, and in scholarly journals and books.

Because critics often try to be lawgivers, declaring which works deserve respect and presuming to say what they are "really" about, criticism is a perennial target of resentment. Misguided or malicious critics can discourage an author who has been feeling his way toward a new mode that offends received taste. Pedantic critics can obstruct a serious engagement with literature by deflecting attention toward inessential matters. As the contemporary French philosopher-critic Jean-Paul Sartre observed, the critic may announce that French thought is a perpetual colloquy between Pascal and Montaigne not in order to make those thinkers more alive but to make thinkers of his own time more dead. Criticism can antagonize authors even when it performs its function well. Authors who regard literature as needing no advocates or investigators are less than grateful when told that their works possess unintended meaning or are imitative or incomplete.

Criticism as a target of criticism

What such authors may tend to forget is that their works, once published, belong to them only in a legal sense; their true owner is the public, which will appropriate them for its own concerns regardless of the critic. The critic's responsibility is not to the author's self-esteem but to the public and to his own standards of judgment, which are usually more exacting than the public's. Justification for his role rests on the premise that literary works are not in fact self-explanatory. A critic is socially useful to the extent that society wants, and receives, a fuller understanding of literature than it could have achieved without him. In filling this appetite the critic whets it further, helping to create a public that cares about artistic quality. Without sensing the presence of such a public an author may either prostitute his talent or squander it in sterile acts of defiance. In this sense the critic is not a parasite but, potentially, someone who is responsible in part for the existence of good writing in his own time and afterward.

Although some critics believe that literature should be discussed in isolation from other matters, criticism usually seems to be openly or covertly involved with social and political debate. Since literature itself is often partisan, is always rooted to some degree in local circumstances, and has a way of calling forth affirmations of ultimate values, it is not surprising that the finest critics have never paid much attention to the alleged boundaries between criticism and other types of discourse. Especially in modern Europe, literary criticism has occupied a central place in debate about cultural and political issues. Sartre's own *What Is Literature?* (1947) is typical in its wide-ranging attempt to prescribe the literary intellectual's ideal relation to the development of his society and to literature as a manifestation of human freedom. Similarly, some prominent American critics, including Alfred Kazin, Lionel Trilling, Kenneth Burke, Philip Rahv, and Irving Howe, began as political radicals in the 1930s and sharpened their concern for literature on the dilemmas and disillusionments of that era. Trilling's influential *The Liberal Imagination* (1950) is simultaneously a collection of literary essays and an attempt to reconcile the claims of politics and art.

Such a reconciliation is bound to be tentative and problematic if the critic believes, as Trilling does, that literature possesses an independent value and a deeper faithfulness to reality than is contained in any political formula. In Marxist states, however, literature has usually been considered a means to social ends, and criticism has been cast in forthrightly partisan terms. Dialectical materialism does not necessarily turn the critic into a mere guardian of party doctrine, but it does forbid him to treat literature as a cause in itself, apart from the working class's needs as interpreted by the party. Where this utilitarian view prevails, the function of criticism is taken to be continuous with that of the state itself, namely, furtherance of the social revolution. The critic's main obligation is not to his texts but to the masses of people whose consciousness must be advanced in the designated direction. In periods of severe orthodoxy the practice of literary criticism has not always been distinguishable from that of censorship.

HISTORICAL DEVELOPMENT

Antiquity. Although almost all of the criticism ever written dates from the 20th century, questions first posed

by Plato and Aristotle are still of prime concern, and every critic who has attempted to justify the social value of literature has had to come to terms with the opposing argument made by Plato in *The Republic.* The poet as a man and poetry as a form of statement both seemed untrustworthy to Plato, who depicted the physical world as an imperfect copy of transcendent ideas and poetry as a mere copy of the copy. Thus, literature could only mislead the seeker of truth. Plato credited the poet with divine inspiration, but this, too, was cause for worry; a man possessed by such madness would subvert the interests of a rational polity. Poets were therefore to be banished from the hypothetical republic.

In his *Poetics*—still the most respected of all discussions of literature—Aristotle countered Plato's indictment by stressing what is normal and useful about literary art. The tragic poet is not so much divinely inspired as he is motivated by a universal human need to imitate, and what he imitates is not something like a bed (Plato's example) but a noble action. Such imitation presumably has a civilizing value for those who empathize with it. Tragedy does arouse emotions of pity and terror in its audience, but these emotions are purged in the process (*katharsis*). In this fashion Aristotle succeeded in portraying literature as satisfying and regulating human passions instead of inflaming them.

Although Plato and Aristotle are regarded as antagonists, the narrowness of their disagreement is noteworthy. Both maintain that poetry is mimetic, both treat the arousing of emotion in the perceiver, and both feel that poetry takes its justification, if any, from its service to the state. It was obvious to both men that poets wielded great power over others. Unlike many modern critics who have tried to show that poetry is more than a pastime, Aristotle had to offer reassurance that it was not socially explosive.

Aristotle's practical contribution to criticism, as opposed to his ethical defense of literature, lies in his inductive treatment of the elements and kinds of poetry. Poetic modes are identified according to their means of imitation, the actions they imitate, the manner of imitation, and its effects. These distinctions assist the critic in judging each mode according to its proper ends instead of regarding beauty as a fixed entity. The ends of tragedy, as Aristotle conceived them, are best served by the harmonious disposition of six elements: plot, character, diction, thought, spectacle, and song. Thanks to Aristotle's insight into universal aspects of audience psychology, many of his dicta have proved to be adaptable to genres developed long after his time.

Later Greek and Roman criticism offers no parallel to Aristotle's originality. Much ancient criticism, such as that of Cicero, Horace, and Quintilian in Rome, was absorbed in technical rules of exegesis and advice to aspiring rhetoricians. Horace's verse epistle *The Art of Poetry* is an urbane amplification of Aristotle's emphasis on the decorum or internal propriety of each genre, now including lyric, pastoral, satire, elegy, and epigram, as well as Aristotle's epic, tragedy, and comedy. This work was later to be prized by Neoclassicists of the 17th century not only for its rules but also for its humour, common sense, and appeal to educated taste. *On the Sublime,* by the Roman-Greek known as "Longinus," was to become influential in the 18th century but for a contrary reason: when decorum began to lose its sway encouragement could be found in Longinus for arousing elevated and ecstatic feeling in the reader. Horace and Longinus developed, respectively, the rhetorical and the affective sides of Aristotle's thought, but Longinus effectively reversed the Aristotelian concern with regulation of the passions.

Horatian and Longinian tendencies

Medieval period. In the Christian Middle Ages criticism suffered from the loss of nearly all the ancient critical texts and from an antipagan distrust of the literary imagination. Such Church Fathers as Tertullian, Augustine, and Jerome renewed, in churchly guise, the Platonic argument against poetry. But both the ancient gods and the surviving classics reasserted their fascination, entering medieval culture in theologically allegorized form. Encyclopaedists and textual commentators explained the supposed Christian content of pre-Christian works and the Old Testament. Although there was no lack of rhetoricians to dictate the correct use of literary figures, no attempt was made to derive critical principles from emergent genres such as the fabliau and the chivalric romance. Criticism was in fact inhibited by the very coherence of the theologically explained universe. When nature is conceived as endlessly and purposefully symbolic of revealed truth, specifically literary problems of form and meaning are bound to be neglected. Even such an original vernacular poet of the 14th century as Dante appears to have expected his *Divine Comedy* to be interpreted according to the rules of scriptural exegesis.

The Renaissance. Renaissance criticism grew directly from the recovery of classic texts and notably from Giorgio Valla's translation of Aristotle's *Poetics* into Latin in 1498. By 1549 the *Poetics* had been rendered into Italian as well. From this period until the later part of the 18th century Aristotle was once again the most imposing presence behind literary theory. Critics looked to ancient poems and plays for insight into the permanent laws of art. The most influential of Renaissance critics was probably Lodovico Castelvetro, whose 1570 commentary on Aristotle's *Poetics* encouraged the writing of tightly structured plays by extending and codifying Aristotle's idea of the dramatic unities. It is difficult today to appreciate that this obeisance to antique models had a liberating effect; one must recall that imitation of the ancients entailed rejecting scriptural allegory and asserting the individual author's ambition to create works that would be unashamedly great and beautiful. Classicism, individualism, and national pride joined forces against literary asceticism. Thus, a group of 16th-century French writers known as the Pléiade—notably Pierre de Ronsard and Joachim du Bellay—were simultaneously classicists, poetic innovators, and advocates of a purified vernacular tongue.

Classicism, individualism, and national pride as literary forces

The ideas of the Italian and French Renaissance were transmitted to England by Roger Ascham, George Gascoigne, Sir Philip Sidney, and others. Gascoigne's "Certayne notes of Instruction" (1575), the first English manual of versification, had a considerable effect on poetic practice in the Elizabethan Age. Sidney's *Defence of Poesie* (1595) vigorously argued the poet's superiority to the philosopher and the historian on the grounds that his imagination is chained neither to lifeless abstractions nor to dull actualities. The poet "doth not only show the way, but giveth so sweet a prospect into the way, as will entice any man to enter into it." While still honouring the traditional conception of poetry's role as bestowing pleasure and instruction, Sidney's essay presages the Romantic claim that the poetic mind is a law unto itself.

Neoclassicism and its decline. The Renaissance in general could be regarded as a neoclassical period, in that ancient works were considered the surest models for modern greatness. Neoclassicism, however, usually connotes narrower attitudes that are at once literary and social: a worldly-wise tempering of enthusiasm, a fondness for proved ways, a gentlemanly sense of propriety and balance. Criticism of the 17th and 18th centuries, particularly in France, was dominated by these Horatian norms. French critics such as Pierre Corneille and Nicolas Boileau urged a strict orthodoxy regarding the dramatic unities and the requirements of each distinct genre, as if to disregard them were to lapse into barbarity. The poet was not to imagine that his genius exempted him from the established laws of craftsmanship.

Neoclassicism had a lesser impact in England, partly because English Puritanism had kept alive some of the original Christian hostility to secular art, partly because English authors were on the whole closer to plebeian taste than were the court-oriented French, and partly because of the difficult example of Shakespeare, who magnificently broke all of the rules. Not even the relatively severe classicist Ben Jonson could bring himself to deny Shakespeare's greatness, and the theme of Shakespearean genius triumphing over formal imperfections is echoed by major British critics from John Dryden and Alexander Pope through Samuel Johnson. The science of Newton and the psychology of Locke also worked subtle changes on neoclassical themes. Pope's *Essay on Criticism* (1711) is a Horatian compendium of maxims, but Pope feels obliged

to defend the poetic rules as "Nature methodiz'd"—a portent of quite different literary inferences from Nature. Dr. Johnson, too, though he respected precedent, was above all a champion of moral sentiment and "mediocrity," the appeal to generally shared traits. His preference for forthright sincerity left him impatient with such intricate conventions as those of the pastoral elegy.

Progress of a cult of "feeling"

The decline of Neoclassicism is hardly surprising; literary theory had developed very little during two centuries of artistic, political, and scientific ferment. The 18th century's important new genre, the novel, drew most of its readers from a bourgeoisie that had little use for aristocratic dicta. A Longinian cult of "feeling" gradually made headway, in various European countries, against Neoclassical canons of proportion and moderation. Emphasis shifted from concern for meeting fixed criteria to the subjective state of the reader and then of the author himself. The spirit of nationalism entered criticism as a concern for the origins and growth of one's own native literature and as an esteem for such non-Aristotelian factors as "the spirit of the age." Historical consciousness produced by turns theories of literary progress and primitivistic theories affirming, as one critic put it, that "barbarous" times are the most favourable to the poetic spirit. The new recognition of strangeness and strong feeling as literary virtues yielded various fashions of taste for misty sublimity, graveyard sentiments, medievalism, Norse epics (and forgeries), Oriental tales, and the verse of plowboys. Perhaps the most eminent foes of Neoclassicism before the 19th century were Denis Diderot in France and, in Germany, Gotthold Lessing, Johann von Herder, Johann Wolfgang von Goethe, and Friedrich Schiller.

Romanticism. Romanticism, an amorphous movement that began in Germany and England at the turn of the 19th century, and somewhat later in France, Italy, and the United States, found spokesmen as diverse as Goethe and August and Friedrich von Schlegel in Germany, William Wordsworth and Samuel Taylor Coleridge in England, Madame de Staël and Victor Hugo in France, Alessandro Manzoni in Italy, and Ralph Waldo Emerson and Edgar Allan Poe in the United States. Romantics tended to regard the writing of poetry as a transcendentally important activity, closely related to the creative perception of meaning in the world. The poet was credited with the godlike power that Plato had feared in him; Transcendental philosophy was, indeed, a derivative of Plato's metaphysical Idealism. In the typical view of Percy Bysshe Shelley, poetry "strips the veil of familiarity from the world, and lays bare the naked and sleeping beauty, which is the spirit of its forms."

Wordsworth's preface to *Lyrical Ballads* (1800), with its definition of poetry as the spontaneous overflow of powerful feelings and its attack on Neoclassical diction, is regarded as the opening statement of English Romanticism. In England, however, only Coleridge in his *Biographia Literaria* (1817) embraced the whole complex of Romantic doctrines emanating from Germany; the British empiricist tradition was too firmly rooted to be totally washed aside by the new metaphysics. Most of those who were later called Romantics did share an emphasis on individual passion and inspiration, a taste for symbolism and historical awareness, and a conception of art works as internally whole structures in which feelings are dialectically merged with their contraries. Romantic criticism coincided with the emergence of aesthetics as a separate branch of philosophy, and both signalled a weakening in ethical demands upon literature. The lasting achievement of Romantic theory is its recognition that artistic creations are justified, not by their promotion of virtue, but by their own coherence and intensity.

The late 19th century. The Romantic movement had been spurred not only by German philosophy but also by the universalistic and utopian hopes that accompanied the French Revolution. Some of those hopes were thwarted by political reaction, while others were blunted by industrial capitalism and the accession to power of the class that had demanded general liberty. Advocates of the literary imagination now began to think of themselves as enemies or gadflies of the newly entrenched bourgeoisie. In some hands the idea of creative freedom dwindled to a bohemianism pitting "art for its own sake" against commerce and respectability. Aestheticism characterized both the Symbolist criticism of Charles Baudelaire in France and the self-conscious decadence of Algernon Swinburne, Walter Pater, and Oscar Wilde in England. At an opposite extreme, realistic and naturalistic views of literature as an exact record of social truth were developed by Vissarion Belinsky in Russia, Gustave Flaubert and Émile Zola in France, and William Dean Howells in the United States. Zola's program, however, was no less anti-bourgeois than that of the Symbolists; he wanted novels to document conditions so as to expose their injustice. Post-Romantic disillusion was epitomized in Britain in the criticism of Matthew Arnold, who thought of critical taste as a substitute for religion and for the unsatisfactory values embodied in every social class.

Toward the end of the 19th century, especially in Germany, England, and the United States, literary study became an academic discipline "at the doctoral level." Philology, linguistics, folklore study, and the textual principles that had been devised for biblical criticism provided curricular guidelines, while academic taste mirrored the prevailing impressionistic concern for the quality of the author's spirit. Several intellectual currents joined to make possible the writing of systematic and ambitious literary histories. Primitivism and Medievalism had awakened interest in neglected early texts; scientific Positivism encouraged a scrupulous regard for facts; and the German idea that each country's literature had sprung from a unique national consciousness provided a conceptual framework. The French critic Hippolyte Taine's *History of English Literature* (published in French, 1863–69) reflected the prevailing determinism of scientific thought; for him a work could be explained in terms of the race, milieu, and moment that produced it. For other critics of comparable stature, such as Charles Sainte-Beuve in France, Benedetto Croce in Italy, and George Saintsbury in England, historical learning only threw into relief the expressive uniqueness of each artistic temperament.

Merging intellectual currents

THE 20TH CENTURY

The ideal of objective research has continued to guide Anglo-American literary scholarship and criticism and has prompted work of unprecedented accuracy. Bibliographic procedures have been revolutionized; historical scholars, biographers, and historians of theory have placed criticism on a sounder basis of factuality. Important contributions to literary understanding have meanwhile been drawn from anthropology, linguistics, philosophy, and psychoanalysis. Impressionistic method has given way to systematic inquiry from which gratuitous assumptions are, if possible, excluded. Yet demands for a more ethically committed criticism have repeatedly been made, from the New Humanism of Paul Elmer More and Irving Babbitt in the United States in the 1920s, through the moralizing criticism of the Cambridge don F.R. Leavis and of the American poet Yvor Winters, to the most recent demands for "relevance."

No sharp line can be drawn between academic criticism and criticism produced by authors and men of letters. Many of the latter are now associated with universities, and the main shift of academic emphasis, from impressionism to formalism, originated outside the academy in the writings of Ezra Pound, T.S. Eliot, and T.E. Hulme, largely in London around 1910. Only subsequently did such academics as I.A. Richards and William Empson in England and John Crowe Ransom and Cleanth Brooks in the United States adapt the New Criticism to reform of the literary curriculum—in the 1940s. New Criticism has been the methodological counterpart to the strain of modernist literature characterized by allusive difficulty, paradox, and indifference or outright hostility to the democratic ethos. In certain respects the hegemony of New Criticism has been political as well as literary; and anti-Romantic insistence on irony, convention, and aesthetic distance has been accompanied by scorn for all revolutionary hopes. In Hulme conservatism and classicism were explicitly linked. Romanticism struck him as "spilt religion," a dangerous

exaggeration of human freedom. In reality, however, New Criticism owed much to Romantic theory, especially to Coleridge's idea of organic form, and some of its notable practitioners have been left of centre in their social thought.

Multiplicity and factionalism

The totality of Western criticism in the 20th century defies summary except in terms of its restless multiplicity and factionalism. Schools of literary practice, such as Imagism, Futurism, Dadaism, and Surrealism, have found no want of defenders and explicators. Ideological groupings, psychological dogmas, and philosophical trends have generated polemics and analysis, and literary materials have been taken as primary data by sociologists and historians. Literary creators themselves have continued to write illuminating commentary on their own principles and aims. In poetry, Paul Valéry, Ezra Pound, Wallace Stevens; in the theatre, George Bernard Shaw, Antonin Artaud, Bertolt Brecht; and in fiction, Marcel Proust, D.H. Lawrence, and Thomas Mann have contributed to criticism in the act of justifying their art.

Most of the issues debated in 20th-century criticism appear to be strictly empirical, even technical, in nature. By what means can the most precise and complete knowledge of a literary work be arrived at? Should its social and biographical context be studied or only the words themselves as an aesthetic structure? Should the author's avowed intention be trusted, or merely taken into account, or disregarded as irrelevant? How is conscious irony to be distinguished from mere ambivalence, or allusiveness from allegory? Which among many approaches—linguistic, generic, formal, sociological, psychoanalytic, and so forth—is best adapted to making full sense of a text? Would a synthesis of all these methods yield a total theory of literature? Such questions presuppose that literature is valuable and that objective knowledge of its workings is a desirable end. These assumptions are, indeed, so deeply buried in most critical discourse that they customarily remain hidden from critics themselves, who imagine that they are merely solving problems of intrinsic interest.

What separates modern criticism from earlier work is its catholicity of scope and method, its borrowing of procedures from the social scienes, and its unprecedented attention to detail. As literature's place in society has become more problematic and peripheral, and as humanistic education has grown into a virtual industry with a large group of professionals serving as one another's judges, criticism has evolved into a complex discipline, increasingly refined in its procedures but often lacking a sense of contact with the general social will. Major modern critics, to be sure, have not allowed their "close reading" to distract them from certain perennial questions about poetic truth, the nature of literary satisfaction, and literature's social utility, but even these matters have sometimes been cast in "value-free" empirical terms.

Science as the outstanding influence

Recourse to scientific authority and method, then, is the outstanding trait of 20th-century criticism. The sociology of Marx, Max Weber, and Karl Mannheim, the mythological investigations of Sir James George Frazer and his followers, Edmund Husserl's phenomenology, Claude Levi-Strauss's anthropological structuralism, and the psychological models proposed by Sigmund Freud and C.G. Jung have all found their way into criticism. The result has been not simply an abundance of technical terms and rules, but a widespread belief that literature's governing principles can be located outside literature. Jungian "archetypal" criticism, for example, regularly identifies literary power with the presence of certain themes that are alleged to inhabit the myths and beliefs of all cultures, while psychoanalytic exegetes interpret poems in exactly the manner that Freud interpreted dreams. Such procedures may encourage the critic, wisely or unwisely, to discount traditional boundaries between genres, national literatures, and levels of culture; the critical enterprise begins to seem continuous with a general study of man. The impetus toward universalism can be discerned even in those critics who are most skeptical of it, the so-called historical relativists who attempt to reconstruct each epoch's outlook and to understand works as they appeared to their first readers. Historical relativism does undermine cross-cultural notions of beauty, but it reduces the record of any given period to data from which inferences can be systematically drawn. Here, too, in other words, uniform methodology tends to replace the intuitive connoisseurship that formerly typified the critic's sense of his role.

The debate over poetic truth may illustrate how modern discussion is beholden to extraliterary knowledge. Critics have never ceased disputing whether literature depicts the world correctly, incorrectly, or not at all, and the dispute has often had more to do with the support or condemnation of specific authors than with ascertainable facts about mimesis. Today it may be almost impossible to take a stand regarding poetic truth without also coming to terms with positivism as a total epistemology. The spectacular achievements of physical science have (with logic questioned by some) downgraded intuition and placed a premium on concrete, testable statements very different from those found in poems. Some of the most influential modern critics, notably I.A. Richards in his early works, have accepted this value order and have confined themselves to behavioristic study of how literature stimulates the reader's feelings. A work of literature, for them, is no longer something that captures an external or internal reality, but is merely a locus for psychological operations; it can only be judged as eliciting or failing to elicit a desired response.

Literature as a form of knowing

Other critics, however, have renewed the Shelleyan and Coleridgean contention that literary experience involves a complex and profound form of knowing. In order to do so they have had to challenge Positivism in general. Such a challenge cannot be convincingly mounted within the province of criticism itself and must depend rather on the authority of antipositivist epistemologists such as Alfred North Whitehead, Ernst Cassirer, and Michael Polanyi. If it is now respectable to maintain, with Wallace Stevens and others, that the world is known through imaginative apprehensions of the sort that poetry celebrates and employs, this is attributable to developments far outside the normal competence of critics.

The pervasive influence of science is most apparent in modern criticism's passion for total explanation of the texts it brings under its microscope. Even formalist schools, which take for granted an author's freedom to shape his work according to the demands of art, treat individual lines of verse with a dogged minuteness that was previously unknown, hoping thereby to demonstrate the "organic" coherence of the poem. The spirit of explanation is also apparent in those schools that argue from the circumstances surrounding a work's origin to the work itself, leaving an implication that the former have caused the latter. The determinism is rarely as explicit or relentless as it was in Taine's scheme of race, milieu, and moment, but this may reflect the fact that causality in general is now handled with more sophistication than in Taine's day.

Whether criticism will continue to aim at empirical exactitude or will turn in some new direction cannot be readily predicted, for the empiricist ideal and its sanctuary, the university, are not themselves secure from attack. The history of criticism is one of oscillation between periods of relative advance, when the imaginative freedom of great writers prompts critics to extend their former conceptions, and periods when stringent moral and formal prescriptions are laid upon literature. In times of social upheaval criticism may more or less deliberately abandon the ideal of disinterested knowledge and be mobilized for a practical end. Revolutionary movements provide obvious instances of such redirection, whether or not they identify their pragmatic goals with the cause of science. It should be evident that the future of criticism depends on factors that lie outside criticism itself as a rationally evolving discipline. When a whole society shifts its attitudes toward pleasure, unorthodox behaviour, or the meaning of existence, criticism must follow along.

Literature as a source of order

As Matthew Arnold foresaw, the waning of religious certainty has encouraged critics to invest their faith in literature, taking it as the one remaining source of value and order. This development has stimulated critical activity, yet, paradoxically, it may also be responsible in part

for a growing impatience with criticism. What Arnold could not have anticipated is that the faith of some moderns would be apocalyptic and Dionysian rather than a sober and attenuated derivative of Victorian Christianity. Thought in the 20th century has yielded a strong undercurrent of anarchism which celebrates libidinous energy and self-expression at the expense of all social constraint, including that of literary form. In the critical writings of D.H. Lawrence, for example, fiction is cherished as an instrument of unconscious revelation and liberation. A widespread insistence upon prophetic and ecstatic power in literature seems at present to be undermining the complex, irony-minded formalism that has dominated modern discourse. As literary scholarship has acquired an ever-larger arsenal of weapons for attacking problems of meaning, it has met with increasing resentment from people who wish to be nourished by whatever is elemental and mysterious in literary experience.

An awareness of critical history suggests that the development is not altogether new, for criticism stands now approximately where it did in the later 18th century, when the Longinian spirit of expressiveness contested the sway of Boileau and Pope. To the extent that modern textual analysis has become what Hulme predicted, "a classical revival," it may not be welcomed by those who want a direct and intense rapport with literature. What is resisted now is not Neoclassical decorum but impersonal methodology, which is thought to deaden commitment. Such resistance may prove beneficial if it reminds critics that rationalized procedures are indeed no substitute for engagement. Excellent work continues to be written, not because a definitive method or synthesis of methods has been found, but on the contrary because the best critics still understand that criticism is an exercise of private sympathy, discrimination, and moral and cultural reflection.

(F.C.C.)

CHILDREN'S LITERATURE

Children's literature first clearly emerged as a distinct and independent form of literature in the second half of the 18th century, before which it had been at best only in an embryonic stage. During the 20th century, however, its growth has been so luxuriant as to make defensible its claim to be regarded with the respect—though perhaps not the solemnity—that is due any other recognized branch of literature.

DEFINITION OF TERMS

"Children." All potential or actual young literates, from the instant they can with joy leaf through a picture book or listen to a story read aloud, to the age of perhaps 14 or 15, may be called children. Thus "children" includes "young people." Two considerations blur the definition. Today's young teenager is an anomaly: his environment pushes him toward a precocious maturity. Thus, though he may read children's books, he also, and increasingly, reads adult books. Second, the child survives in many adults. As a result, some children's books (*e.g.,* Lewis Carroll's *Alice in Wonderland,* A.A. Milne's *Winnie-the-Pooh,* and, at one time, Munro Leaf's *Story of Ferdinand*) are also read widely by adults.

"Literature." In the term children's literature, the more important word is literature. For the most part, the adjective imaginative is to be felt as preceding it. It comprises that vast, expanding territory recognizably staked out for a junior audience, which does not mean that it is not also intended for seniors. Adults admittedly make up part of its population: children's books are written, selected for publication, sold, bought, reviewed, and often read aloud by grown-ups. Sometimes they seem also to be written with adults in mind, as for example the popular French *Astérix* series of comics parodying history. Nevertheless, by and large there *is* a sovereign republic of children's literature. To it may be added five colonies or dependencies: first, "appropriated" adult books satisfying two conditions—they must generally be read by children and they must have sharply affected the course of children's literature (Daniel Defoe's *Robinson Crusoe,* Jonathan Swift's *Gulliver's Travels,* the collection of folktales by the brothers Jacob and Wilhelm Grimm, the folk-verse anthology *Des Knaben Wunderhorn* ["The Boy's Magic Horn"], edited by Achim von Arnim and Clemens Brentano, and William Blake's *Songs of Innocence*); second, books the audiences of which seem not to have been clearly conceived by their creators (or their creators may have ignored, as irrelevant, such a consideration) but that are now fixed stars in the child's literary firmament (Mark Twain's *Adventures of Huckleberry Finn,* and Charles Perrault's fairy tales; third, picture books and easy-to-read stories commonly subsumed under the label of literature but qualifying as such only by relaxed standards (though Beatrix Potter and several other writers do nonetheless qualify); fourth, first quality children's versions of adult classics (Walter de la Mare's *Stories from the Bible,* perhaps Howard Pyle's retellings of the Robin Hood ballads and tales; finally, the domain of once oral "folk" material that children have kept alive—folktales and fairy tales; fables, sayings, riddles, charms, tongue twisters; folksongs, lullabies, hymns, carols, and other simple poetry; rhymes of the street, the playground, the nursery; and, supremely, *Mother Goose* and nonsense verse.

Categories often considered children's literature

Five categories that are often considered children's literature are excluded from this section. The broadest of the excluded categories is that of unblushingly commercial and harmlessly transient writing, including comic books, much of which, though it may please young readers, and often for good reasons, is for the purposes of this article notable only for its sociohistorical, rather than literary, importance. Second, all books of systematic instruction are barred except those sparse examples (*e.g.,* the work of John Amos Comenius) that illuminate the history of the subject. Third, excluded from discussion is much high literature that was not originally intended for children: from the past, Jean de La Fontaine's *Fables,* James Fenimore Cooper's Leatherstocking tales, Sir Walter Scott's *Ivanhoe,* Charlotte Brontë's *Jane Eyre,* Alexandre Dumas' *Three Musketeers,* Rudyard Kipling's *Kim;* from the modern period, Marjorie Kinnan Rawlings' *Yearling,* J.D. Salinger's *Catcher in the Rye, The Diary of Anne Frank,* Thor Heyerdahl's *Kon-Tiki,* Enid Bagnold's *National Velvet.* A fourth, rather minor, category comprises books about the young where the content but not the style or point of view is relevant (Sir James Barrie's *Sentimental Tommy,* William Golding's *Lord of the Flies,* F. Anstey's [Thomas Anstey Guthrie] *Vice Versa*). Finally, barred from central, though not all, consideration is the "nonfiction," or fact, book. Except for a handful of such books, the bright pages of which still rain influence or which possess artistic merit, this literature should be viewed from its socioeducational-commercial aspect.

THE CASE FOR A CHILDREN'S LITERATURE

Many otherwise comprehensive histories of literature slight or omit the child's reading interests. Many observers have made explicit the suspicion that children's literature, like that of detection or suspense, is "inferior." They cannot detect a sufficiently long "tradition"; distinguish an adequate number of master works; or find, to use on thoughtful critic's words, "style, sensibility, vision."

Others, holding a contrary view, assert that a tradition of two centuries is not to be ignored.

Though the case for a children's literature must primarily rest on its major writers (including a half dozen literary geniuses), it is based as well on other supports that bolster its claim to artistic stature.

Children's literature, while a tributary of the literary mainstream, offers its own identifiable, semidetached history. In part it is the issue of certain traceable social movements, of which the "discovery" of the child (see below) is the most salient. It is independent to the degree

that, while it must meet many of the standards of adult literature, it has also developed aesthetic criteria of its own by which it may be judged. According to some of its finest practitioners, it is independent, too, as the only existing literary medium enabling certain things to be said that would otherwise remain unsaid or unsayable. The nature of its audience sets it apart; it is often read, especially by pre-12-year-olds, in a manner suggesting trance, distinct from that of adult reading. Universally diffused among literate peoples, it offers a rich array of genres, types, and themes, some resembling grown-up progenitors, many peculiar to itself. Its "style, sensibility, vision" range over a spectrum wide enough to span matter-of-fact realism and tenuous mysticism.

Scholarship and criticism

Other measures of its maturity include an extensive body (notably in Germany, Italy, Sweden, the U.S.S.R., and Japan) of commentary, scholarship, criticism, history, biography, and bibliography, along with the beginnings of an aesthetic theory or philosophy of composition. Finally, one might note its power to engender its own institutions: publishing houses, theatres, libraries, itinerant storytellers, critics, periodicals, instruction in centres of higher learning, lectureships, associations and conferences, "book weeks," collections, exhibitions, and prizes. Indeed, the current institutionalizing of children's literature on an international scale has gone so far, some feel, as to cast a shadow on the spontaneity and lack of self-consciousness that should lie at its heart.

SOME GENERAL FEATURES AND FORCES

The discovery of the child. A self-aware literature flows from a recognition of its proper subject matter. The proper subject matter of children's literature, apart from informational or didactic works, is children. More broadly, it embraces the whole content of the child's imaginative world and that of his daily environment, as well as certain ideas and sentiments characteristic of it. The population of this world is made up not only of children themselves but of animated objects, plants, even grammatical and mathematical abstractions; toys, dolls, and puppets; real, chimerical, and invented animals; miniature or magnified humans; spirits or grotesques of wood, water, air, fire, and space; supernatural and fantasy creatures; figures of fairy tale, myth, and legend; imagined familiars and *Doppelgänger;* and grown-ups as seen through the child's eyes—whether Napoleon, Dr. Dolittle, parents, or the corner grocer. That writers did not detect this lively cosmos for two and a half millennia is one of the curiosities of literature. At any moment there has always been a numerous, physically visible, and audible company of children. Whether this sizable minority, appraised as literary raw material, could be as rewarding as the adult majority was never asked.

And so, almost to the dawn of the Industrial Revolution, children's literature remained recessive. The chief, though not the only, reason is improbably simple: the child himself, though there, was not seen. Not seen, that is, as a child.

In preliterate societies he was and is viewed in the light of his social, economic, and religious relationship to the tribe or clan. Though he may be nurtured in all tenderness, he is thought of not as himself but as a pre-adult, which is but one of his many forms. Among Old Testament Jews the child's place in society replicated his father's, molded by his relation to God. So too, in ancient Greece and Rome, the child, dressed in the modified adult costume that with appropriate changes of fashion remained his fate for centuries to come, was conceived as a miniature adult. His importance lay not in himself but in what Aristotle would have called his final cause: the potential citizen-warrior. A girl child was a seedbed of future citizen-warriors. Hence classical literature either does not see the child at all or misconstrues him. Astyanax and Ascanius, as well as Medea's two children, are not persons. They are stage props. Aristophanes scorns as unworthy of dramatic treatment the children in Euripides' *Alcestis.*

Throughout the Middle Ages and far into the late Renaissance the child remained, as it were, *terra incognita.* A sharp sense of generation gap—one of the motors of a children's literature—scarcely existed. The family, young and old, was a kind of homogenized mix. Sometimes children were even regarded as infrahuman: for Montaigne they had "neither mental activities nor recognizable body shape." The year 1658 is a turning point. In that year a Moravian educator, Comenius, published *Orbis Sensualium Pictus* (*The Visible World in Pictures,* 1659), a teaching device that was also the first picture book for children. It embodied a novel insight: children's reading should be of a special order because children are not scaled-down adults. But the conscious, systematic, and successful exploitation of this insight was to wait for almost a century.

Beginnings of children's literature

It is generally felt that, both as a person worthy of special regard and as an idea worthy of serious contemplation, the child began to come into his own in the second half of the 18th century. His emergence, as well as that of a literature suited to his needs, is linked to many historical forces, among them the development of Enlightenment thought (Rousseau and, before him, John Locke); the rise of the middle class; the beginnings of the emancipation of women—children's literature, unlike that for grown-ups, is in large measure a distaff product; and Romanticism, with its minor strands of the cult of the child (Wordsworth and others) and of genres making a special appeal to the young (folktales and fairy tales, myths, ballads). Yet, with all these forces working for the child, he still might not have emerged had it not been for a few unpredictable geniuses: William Blake, Edward Lear, Lewis Carroll, George MacDonald, Louisa May Alcott, Mark Twain, Collodi, Hans Christian Andersen. But, once tentatively envisaged as an independent being, a literature proper to him could also be envisaged. And so in the mid-18th century, what may be defined as children's literature was at last developing.

Shifting visions of the child. Even after the child had been recognized, his literature on occasion persisted in viewing him as a diminutive adult. More characteristically, however, "realistic" (that is, nonfantastic) fiction in all countries regarded the discovered child in a mirror that provided only a partial reflection of him. There are fewer instances of attempts to present the child whole, in the round, than there are (as in Tolstoy or Joyce) attempts to represent the whole adult. Twain's Huck Finn, Erich Kästner's Emil (in *Emil and the Detectives*), Vadim Frolov's Sasha (in *What It's All About*), and Maria Gripe's delightful Josephine all exemplify in-the-round characterization. More frequently, however, children's literature portrays the young as types. Thus there is the brand of hell of the Puritan tradition; the moral child of Mrs. Trimmer; the well-instructed child of Madame de Genlis; the small upper class benefactor of Arnaud Berquin; the naughty child, modulated variously in Catherine Sinclair's *Holiday House* and in the books of Comtesse de Ségur, E. Nesbit, Dr. Heinrich Hoffmann (*Struwwelpeter*), and Wilhelm Busch (*Max und Moritz*); the rational child of Maria Edgeworth; the little prig of Thomas Day's *Sandford and Merton;* the little angel (Frances Hodgson Burnett's *Little Lord Fauntleroy*); the forlorn waif (Hector Malot's *Sans Famille*); the manly, outdoor child (Arthur Ransome's *Swallows and Amazons*); etc. The rationale behind these shifting visions of childhood is akin to Renaissance theories of "humours" or "the ruling passion." Progress in children's literature depended partly on abandoning this mechanical, part-for-the-whole attitude. One encouraging note in realistic children's fiction of the second half of the 20th century in all advanced countries is the appearance of a more organic view.

Slow development. A third universal feature: children's literature appears later than adult and grows more slowly. Only after the trail has been well blazed does it make use of new techniques, whether of composition of illustration. As for content, only after World War II did it exploit certain realistic themes and attitudes, turning on race, class, war, and sex, that had been part of general literature at least since the 1850s. This tardiness may be due to the child's natural conservatism.

Realistic themes and attitudes

Fourth, the tempo of development varies sharply from country to country and from region to region. It is plausible that England should create a complex children's literature, while a less developed region (the Balkans, for

example) might not. Less clear is why the equally high cultures of France and England should be represented by unequal literatures.

The didactic versus the imaginative. The fifth, and most striking, general feature is the creative tension resulting from a constantly shifting balance between two forces: that of the pulpit-schoolroom and that of the imagination. The first force may take on many guises. It may stress received religious or moral doctrine, thus generating the Catholic children's literature of Spain or the moral tale of Georgian and early Victorian England. It may bear down less on morality than on mere good manners, propriety, or adjustment to the prevailing social code. It may emphasize nationalist or patriotic motives, as in Edmondo De Amicis' post-Risorgimento *Cuore* (*The Heart of a Child*) or much Soviet production. Or its concern may be pedagogical, the imparting of "useful" information, frequently sugarcoated in narrative or dialogue. Whatever its form, it is distinguishable from the shaping spirit of imagination, which ordinarily embodies itself in children's games and rhymes, the fairy tale, the fantasy, animal stories such as Kipling's *Jungle Books,* nonsense, nonmoral poetry, humour, or the realistic novel conceived as art rather than admonition.

Children's literature designed for entertainment rather than self-improvement, aiming at emotional expansion rather than acculturation, usually develops late. *Alice in Wonderland,* the first supreme victory of the imagination (except for *Mother Goose*), did not appear until 1865. Frequently the literature of delight has underground sources of nourishment and inspiration: oral tradition, nursery songs, and the folkish institutions of the chapbook and the penny romance.

While the didactic and the imaginative are conveniently thought of as polar, they need not always be inimical. *Little Women* and *Robinson Crusoe* are at once didactically moral and highly poetical. Nevertheless, many of the acknowledged classics in the field, from *Alice* to *The Hobbit,* incline to fantasy, which is less true of literature for grown-ups.

THE DEVELOPMENT OF CHILDREN'S LITERATURE

Criteria. Keeping these five general features of development in mind, certain criteria may now be suggested as helpful in making a gross estimate of the degree of that development within any given country. Some of these criteria are artistic. Others link with social progress, wealth, technological level, or the political structure. In what seems their order of importance, these criteria are:

1. Degree of awareness of the child's identity (see above).
2. Progress made beyond passive dependence on oral tradition, folklore, and legend.
3. Rise of a class of professional writers, as distinct from moral reformers, schoolteachers, clerics, or versatile journalists—all those who, for pedagogical, doctrinal, or pecuniary reasons turn themselves into writers for children. For example, a conscious Italian literature for young people may be said to have begun in 1776 with the Rev. Francesco Soave's moralistic "Short Stories," and largely because that literature continued to be composed largely by nonprofessionals, its record has been lacklustre. It took more than a century after the Rev. Francesco to produce a *Pinocchio.* And only in the 20th century, as typified by the outstanding work of a professional like Gianni Rodari (*e.g., Telephone Tales*), did children's literature in Italy seem to be getting into full stride.
4. Degree of independence from authoritarian controls: church, state, school system, a rigid family structure. Although this criterion might be rejected by historians of some nations, one must somehow try to explain why the Spanish, a great and imaginative people, took so long—indeed until 1952—to produce, in Sanchez-Silva, a children's writer of any notable talent.
5. Number of "classics" the influence of which transcends national boundaries.
6. Invention of new forms or genres and the exploitation of a variety of traditional ones.
7. Measure of dependence on translations.
8. Quantity of primary literature: that is, annual production of children's books and, more to the point, of good children's books.
9. Quantity of secondary literature: richness and scope of a body of scholarship, criticism, reviewing.
10. Level of institutional development: libraries, publishing houses, associations, etc.

To these criteria some might add a vigorous tradition of illustration. But that is arguable. While Beatrix Potter's words and pictures compose an indivisible unit, it is equally true that a country may produce a magnificent school of artists (Czechoslovakia's Jiří Trnka, Ota Janeček, and others) without developing a literature of matching depth and variety.

The criteria applied: three examples. *West versus East.* The first application of such standards reveals the expected: a gap separating the achievement of the Far East from that of the West. Some Eastern literatures (New Guinea) have not advanced beyond the stage of oral tradition. Others (India, the Philippines, Ceylon, Iran) have been handicapped by language problems. Professional children's writers are rarer than in the West: according to D.R. Kalia, former director of the Delhi Public Library, "No such class exists in Hindi." In Japan, authoritarian patterns—filial piety and ancestor worship—have operated as brakes, though far less since World War II. A low economic level and inadequate technology discourage, in such countries as Burma, Sri Lanka (Ceylon), and Thailand, the origination and distribution of indigenous writing. A towering roadblock is the tendency to imitate the children's books of the West.

It is true that this vast Eastern region, considered as a whole, has produced a number of works ranking as "classics." Most advanced is Japan. Its literature for children goes back at least to the late 19th century and by 1928 was established in its own right. Japan's "discovery" of the child seems to have been made directly after World War II. In Iwaya Sazanami, Japan has its Grimm; in Ogawa Minei, perhaps its Andersen; in the contemporary Ishii Momoko, a critic and creative writer of quality; in Takeyama Michio's *Harp of Burma* (available in English), a high-quality postwar controversial novel. But, though less markedly in Japan, the basic Oriental inspiration remains fixed in folklore (also, in China and Japan, in nursery songs and rhymes), and the didactic imperative continues to act as a hobble. By most criteria the development of Eastern (as compared with Western) children's literature still appears to be sparse and tentative.

Japanese development

North versus south. In western Europe there is a sharp variation or unevenness, as between north and south, in the tempo of development. This basic feature was first pointed out by Paul Hazard, a French critic, in *Les Livres, les enfants et les hommes* (Eng. trans. by Marguerite Mitchell, *Books, Children and Men,* 1944; 4th ed., 1960): "In the matter of literature for children the North surpasses the South by a large margin." For Hazard, Spain had no children's literature; Italy, with its *Pinocchio* and *Cuore,* could point only to an isolated pair of works of note, and even France in order to strengthen its claims had to include northern Frenchmen: Erckmann-Chatrian, Jules Verne—and the classic Comtesse de Ségur came from Russia.

Hazard wrote in the 1920s. Since then the situation has improved, not only in his own country, but in Italy and in Portugal. Yet he is essentially correct: the south cannot match the richness of England, Scotland, Germany, and the Scandinavian countries. To reinforce his position, one might also adduce the United States, noting that the Mason–Dixon line is (though not in the field of general literature) a dividing line: the American South, even including the Uncle Remus stories, has supplied very little good children's reading. As for nursery literature, though analogous rhymes are found everywhere, especially in China, the English *Mother Goose* is unique in the claims made for it as a work of art.

Superiority of northern literature

Why is the north superior to the south? The first criterion of development may be illuminating. It simply restates Hazard's dictum: "For the Latins, children have never been anything but future men. The Nordics have understood better this truer truth, that men are only grown-

up children." ("Adults are obsolete children," says the American children's author "Dr. Seuss.") Hazard does not mention other factors. Historically, the south has shown greater attachment to authoritarian controls. Also up to recent times it has depended heavily on reworked folklore as against free invention. Besides there is the mysterious factor of climate: it could be true that children in Latin countries mature faster and are sooner ready for adult literature. In France a special intellectual tradition, that of Cartesian logic, tends to discourage a children's literature. Clear and distinct ideas, excellent in themselves, do not seem to feed the youthful imagination.

Latin America. Once more applying the chosen criteria, familiar patterns are recognizable: unevenness, as compared with the United States; belatedness—in Argentina the *cuento infantil* is hardly detectable before 1900; and especially an unbalanced polarity, with didacticism decidedly the stronger magnet. The close connection of the church with the child's family and school life has tended to encourage a literature stressing piety, and this at a time when the West, at least in its northern latitudes, is concerned less with the salvation than with the imagination of the child. Fantasy emerged only in the 1930s, in Brazil and in Mexico, where a Spanish exile, Antoniorobles, continued to develop his inventive vein. And realistic writing about the actual life of the young evolved even more deliberately, being generally marked by a patriotic note. Though understandable and wholesome, this did not seem to help the cause of the imagination.

Folklore has been vigorously exploited, often by scholars of high repute. It is largely influenced by the legendry of Spain. Cuba, however, has produced interesting Afro-American tales for children; Argentina offers some indigenous folk stories and tales of gaucho life; and Central America is rich in native traditional verse enjoyed by children.

Characteristics of Latin-American literature

Latin American literature in general displays a special characteristic, part of its Iberian heritage: a partiality for linguistic decoration, which is unpalatable to the relatively straightforward taste of the young reader. Also the Latin-American view of the child remains tinged with a sentimentality from which many European countries and the United States had by 1914 more or less freed themselves. Thus verse for children, a medium specially cultivated in Latin America, has run to the soft, the sweet, even the lachrymose rather than to the gay, the humorous, or the sanguine—moods more congenial to the child's sensibility. This is true even of the children's verse of the Nobel Prize-winning poet Gabriela Mistral. To these two weaknesses one must add a third: the practical difficulty involved in the fact that most families cannot afford books. The absence of a powerful middle class has had a retarding effect.

Children in Latin America often complain that the authors write not for them but for their parents. They are given *lectura* ("reading matter") rather than *literatura,* which is but to say that in Latin America the admonitory note, considered so useful by church, state, and parent, continues to be sounded.

In summary, and applying the criteria: some less advanced Latin-American countries can hardly be said to have a children's literature at all. Others have produced notable writers: Brazil's José Bento Monteiro Lobato, Argentina's Ana Maria Berry, Colombia's Rafael Pombo, Uruguay's Horacio Quiroga. Yet the quality gap separating Latin-American children's literature from that of its northern neighbour is still wide.

HISTORICAL SKETCHES OF THE MAJOR LITERATURES

England. *Overview.* The English have often confessed a certain reluctance to say good-bye to childhood. This curious national trait, baffling to their continental neighbours, may lie at the root of their supremacy in children's literature. Yet it remains a mystery.

But, if it cannot be accounted for, it can be summed up. From the critic's vantage point, the English (as well as the Scots and the Welsh) must be credited with having originated or triumphed in more children's genres than any other country. They have excelled in the school story, two solid centuries of it, from Sarah Fielding's *The Governess; or, The Little Female Academy* (1745) to say, C. Day Lewis' *Otterbury Incident* (1948) and including such milestones as Thomas Hughes's *Tom Brown's School Days* (1857) and Kipling's *Stalky & Co.* (1899); and the boy's adventure story, with one undebatable world masterpiece in Stevenson's *Treasure Island* (1883), plus a solid line of talented practitioners, from the Victorian Robert Ballantyne (*The Coral Island*) to the contemporary Richard Church and Leon Garfield (*Devil-in-the-Fog*); the "girls' book," often trash but possessing in Charlotte M. Yonge at least one writer of exceptional vitality; historical fiction, from Marryat's vigorous but simple *Children of the New Forest* (1847) to the even more vigorous but burnished novels of Rosemary Sutcliff; the "vacation story," in which Arthur Ransome still remains unsurpassed; the doll story, from Margaret Gatty and Richard Henry Horne to the charming fancies of Rumer Godden and the remarkable serious development of this tiny genre in Pauline Clarke's *Return of the Twelves* (1962); the realism-cum-fantasy novel, for which E. Nesbit provided a classic, and P.L. Travers a modern, formulation; high fantasy (Lewis Carroll, George MacDonald, C.S. Lewis, Alan Garner); nonsense (Carroll again, Lear, Belloc); and nursery rhymes. In Jonathan Swift's *Gulliver's Travels* and Daniel Defoe's *Robinson Crusoe,* the English furnished two archetypal narratives that have bred progeny all over the world, and in Mary Norton's Tom-Thumb-and-Gulliver-born *The Borrowers* (1952) a work of art. In Leslie Brooke (*Johnny Crow's Garden*) and Beatrix Potter (*e.g., The Tale of Peter Rabbit*) they have two geniuses of children's literature (and illustration) for very small children—probably the most difficult of all the genres. In poetry they begin at the top with William Blake and continue with Christina Rossetti, Robert Louis Stevenson, Eleanor Farjeon, Walter de la Mare, A.A. Milne, and James Reeves. In the mutation of fantasy called whimsy, Milne (*Winnie-the-Pooh*) reappears as a master. In the important field of the animal story, Kipling, with his *Jungle Books* (1894, 1895) and *Just So Stories* (1902) remains unsurpassed. Finally the English have produced a number of unclassifiable masterpieces such as Kenneth Grahame's *Wind in the Willows* (which is surely more than an animal story) and several unclassifiable writers (Mayne and Lucy Boston, for example).

Genres of English children's literature

The social historian, surveying the same field from a different angle, would point out that the English were the first people in history to develop not only a self-conscious, independent children's literature but also the commercial institutions capable of supporting and furthering it. He would note the striking creative swing between didacticism and delight. He would detect the sources in ballads, chapbooks, nurses' rhymes, and street literature that have at critical moments prompted the imagination. What would perhaps interest him most is the way in which children's literature reflects, over more than two centuries, the child's constantly shifting position in society.

Prehistory (early Middle Ages to 1712). "Children's books did not stand out by themselves as a clear but subordinate branch of English literature until the middle of the 18th century." At least one critic has used "prehistorical" to designate all children's books published in England up to 1744, when John Newbery offered *A Little Pretty Pocket-Book.*

Before that, and as far back as the Middle Ages, children came in contact with schoolroom letters. There was the Anglo-Saxon theologian and historian the Venerable Bede, with his textbook on natural science, *De natura rerum.* There were the question-and-answer lesson books of the great English scholar Alcuin; the *Colloquy* of the English abbot Aelfric; the *Elucidarium* of the archbishop of Canterbury Anselm, often thought of as the first "encyclopaedia" for young people. Not until the mid-14th century was English (the genius of which somehow seems fitter than Latin for children's books) thought of as proper for literature. (For his son "litel Lowis" Geoffrey Chaucer wrote in English the "Treatise on the Astrolabe" (1391). The English child was also afflicted, in the 15th and 16th centuries, by many "Books of Courtesy" (such as *The Babees Boke, c.* 1475), the ancestors of modern, equally ineffective manuals of conduct.

Early English works

Along with these instructional works, there flourished, at least from the very early Renaissance, an unofficial or popular literature. It may not have been meant for children but—no one quite knows how—children managed to recognize it as their own. It included fables, especially those of Aesop; folk legends, such as those in the much read *Gesta Romanorum;* bestiaries, which, along with Aesop, may be ancestral to that flourishing children's genre, the animal story; romances, often clustering around King Arthur and Robin Hood; fairy tales, of which Jack the Giant Killer was the type; and nursery rhymes, probably largely orally transmitted. Perhaps the most influential underground literature consisted of the chapbooks, low-priced folded sheets containing ballads and romances (*Bevis of Southampton,* and *The Seven Champions of Christendom* [1597] were favourites), sold by wandering hawkers and peddlers. They fed the imagination of the poor, old and young, from Queen Anne's reign almost through Queen Victoria's. These native products of fancy were, in the early 18th century, reinforced by the first English translations of the classically simple French fairy tales of Charles Perrault and the more self-conscious ones of Madame D'Aulnoy.

Against this primitive literature of entertainment stands a primitive literature of didacticism stretching back to the early Middle Ages. This underwent a Puritan mutation after the Restoration. It is typified by that classic for the potentially damned child, *A Token for Children* (1671), by James Janeway. The Puritan outlook was elevated by Bunyan's *Pilgrim's Progress* (1678), which, often in simplified form, was either forced upon children or more probably actually enjoyed by them in lieu of anything better. Mrs. Overtheway (in Juliana Ewing's *Mrs. Overtheway's Remembrances,* 1869), recalling her childhood reading, refers to it as "that book of wondrous fascination." A softened Puritanism also reveals itself in Bunyan's *Book for Boys and Girls: or, Country Rhymes for Children* (1686), as well as the *Divine and Moral Songs for Children* by the hymn composer Isaac Watts, whose "How doth the little busy bee" still exhales a faint endearing charm.

Robinson Crusoe and *Gulliver's Travels*

The entire pre-1744 period is redeemed by two works of genius. Neither *Robinson Crusoe* nor *Gulliver's Travels* was meant for children. Immediately abridged and bowdlerized, they were seized upon by the prosperous young. The poorer ones, the great majority, had to wait for the beginning of the cheap reprint era. Both books fathered an immense progeny in the children's field. Defoe engendered a whole school of "Robinsonnades" in most European countries, the most famous example being Wyss's *Swiss Family Robinson* (1812–13).

On the whole, during the millennium separating Alcuin from Newbery, the child's mind was thought of, if at all, as something to be improved; his imagination as something to be shielded; his soul as something to be saved. And on the whole the child's mind, imagination, and soul resisted, persisted, and somehow, whether in a dog-eared penny history of *The Babes in the Wood* or the matchless chronicle of Gulliver among the Lilliputians, found its own nourishment.

From "T.W." to "Alice" (1712?–1865). Napoleon called the English a "nation of shopkeepers," and in England art may owe much to trade. Children's literature in England got its start from merchants such as Thomas Boreman, of whom little is known, and especially John Newbery, of whom a great deal more is known. Research has established that at least as early as 1730 Boreman began publishing for children (largely educational works) and that in 1742 he produced what sounds like a recreational story, *Cajanus, the Swedish Giant.* Beginnings of English children's literature might be dated from the first decade of the 18th century, when a tiny 12-page, undated book called *A Little Book for Little Children* by "T.W." appeared. It is instructional but, as the critic Percy Muir says, important as the earliest publication in English "to approach the problem from the point of view of the child rather than the adult." In sum, without detracting from the significance of Newbery, it may be said that he was merely the first great success in a field that had already undergone a certain amount of exploitation.

Newbery's publications

The elevation of the publisher-bookseller-editor Newbery (who also sold patent medicines) to the position of patron saint is an excusable piece of sentiment. Perhaps it originated with one of his back writers who doubled as a man of genius. In Chapter XVIII of *The Vicar of Wakefield* (1766), Oliver Goldsmith lauds his employer as "the philanthropic bookseller of St. Paul's Churchyard, who has written so many books for children, calling himself their friend, but who was the friend of all mankind." There is no reason to believe that Newbery was anything but an alert businessman who discovered and shrewdly exploited a new market: middle class children, or rather their parents. Nevertheless this was a creative act. In 1744 he published *A Little Pretty Pocket-Book.* Its ragbag of contents—pictures of children's games, jingles, fables, "an agreeable Letter to read from Jack the Giant Killer," plus a bonus in the form of "a Ball and a Pincushion"—are of interest only because, addressing itself single-mindedly to a child audience, it aimed primarily at diversion. Thus children's literature clearly emerged into the light of day.

The climate of Newbery's era was nevertheless more suited to a literature of didacticism than to one of diversion. John Locke's *Some Thoughts concerning Education* (1693) is often cited as an early Enlightenment emancipatory influence. But close inspection of this manual for the mental conditioning of gentlemen reveals a strong English stress on character building and practical learning. Locke thinks little of the natural youthful inclination to poetry: "It is seldom seen that anyone discovers mines of Gold or Silver in Parnassus." He does endorse, as a daring idea, the notion that a child should read for pleasure, and he recommends Aesop. But the decisive influence was not Locke's. It came from across the Channel with Rousseau's best-seller *Émile* (1762). What is positive in Rousseau—his recognition that the child should not be too soon forced into the straitjacket of adulthood—was more or less ignored. Other of his doctrines had a greater effect on children's literature. For all his talk of freedom, he provided his young Émile with an amiable tyrant for a teacher, severely restricting his reading to one book *Robinson Crusoe.* It was his didactic strain, exemplified in the moral French children's literature of Arnaud Berquin and Madame de Genlis, that attracted the English.

Influence of Rousseau on English children's literature

They took more easily to Rousseau's emphasis on virtuous conduct and instruction via "nature" than they did to his advocacy of the liberation of personality. Some writers, such as Thomas Day, with his long-lived *Sandford and Merton,* were avowedly Rousseauist. Others took from him what appealed to them. Sarah Kirby Trimmer, whose *Fabulous Histories* specialized in piety, opposed the presumably free-thinking Rousseau on religious grounds but was in other respects strongly influenced by him. The same is true of Anna Laetitia Barbauld, with her characteristically titled *Lessons for Children.* But Mary Martha Sherwood could hardly have sympathized with Rousseau's notion of the natural innocence of children; the author of *The History of the Fairchild Family* (1818–47) based her family chronicle on the proposition (which she later softened) that "all children are by nature evil." Of all the members of the flourishing Rousseauist or quasi-Rousseauist school of the moral tale, only one was a true writer. Maria Edgeworth may still be read.

Though the tone varies from Miss Edgeworth's often sympathetic feeling for children to Mrs. Sherwood's Savonarolan severities, one idea dominates: a special literature for the child must be manufactured in order to improve or reform him. The reigning mythology is that of reason, a mythology difficult to sell to the young.

Yet during the period from John Newbery's *Little Pretty Pocket-Book* to Lewis Carroll's *Alice in Wonderland,* children's literature also showed signs of antisolemnity. In verse there was first of all William Blake. His *Songs of Innocence* (1789) was not written for children, perhaps indeed not written for anyone. But its fresh, anti-restrictive sensibility, flowing from a deep love for the very young, decisively influenced all English verse for children. Yet the poetry the young really read or listened to at the opening of the 19th century was not Blake but *Original Poems for Infant Minds* (1804), by "Several Young Persons," in-

cluding Ann and Jane Taylor. The Taylor sisters, though adequately moral, struck a new note of sweetness, of humour, at any rate of nonpriggishness. Their "Twinkle, twinkle, little star," included in *Rhymes for the Nursery* (1806), has not only been memorized but actually liked by many generations of small children. No longer read, but in its way similarly revolutionary, was *The Butterfly's Ball and the Grasshopper's Feast* (1807), by William Roscoe, a learned member of Parliament and writer on statistics. The gay and fanciful nonsense of this rhymed satiric social skit enjoyed, despite the seeming dominance of the moral Barbaulds and Trimmers, a roaring success.

Lear's nonsense verse

Great nonsense verse, however, had to await the coming of a genius, Edward Lear, whose *Book of Nonsense* (1846) was partly the product of an emergent and not easily explainable Victorian feeling for levity and partly the issue of a fruitfully neurotic personality, finding relief for its frustrations in the noncontingent world of the absurd and the free laughter of children.

In prose may be noted, toward the end of the period under discussion, the dawn of romantic historical fiction, with Frederick Marryat's *Children of the New Forest* (1847), a story of the English Civil War; and of the manly open-air school novel, with Thomas Hughes's *Tom Brown's School Days* (1857). A prominent milestone in the career of the "realistic" children's family novel is *Holiday House* (1839), by Catherine Sinclair, in which at last there are children who are noisy, even naughty, yet not destined for purgatory. Though Miss Sinclair's book does conclude with a standard deathbed scene, the overall atmosphere is one of gaiety. The victories in the field of children's literature may seem small, but they can be decisive. It was a small, decisive victory to have introduced in *Holiday House* an Uncle David, whose parting admonition to his nieces and nephews is: "Now children! I have only one piece of serious, important advice to give you all, so attend to me!—Never crack nuts with your teeth!"

A similar note was struck by Henry (later Sir Henry) Cole with his *Home Treasury* series, featuring traditional fairy tales, ballads, and rhymes. The fairy tale then began to come into its own, perhaps as a natural reaction to the moral tale. John Ruskin's *King of the Golden River* (1851) and William Makepeace Thackeray's "fireside pantomime" *The Rose and the Ring* (1855) were signs of a changing climate, even though the Grimm-like directness of the first is partly neutralized by Ruskin's moralistic bent and the gaiety of the second is spoiled by a laborious, parodic slyness. More important than these fairy tales, however, was the aid supplied by continental allies: the English publication in 1823–26 of the Grimms' *Fairy Tales;* in 1846 of Andersen's utterly personal fairy tales and folktales; in the '40s and '50s of other importations from the country of fancy, notably Sir George Dasent's version of the stirring *Popular Tales from the Norse* (1859), collected by Peter Christen Asbjørnsen and J.E. Moe. Though the literature of improvement continued to maintain its vigour, England was readying itself for Lewis Carroll.

Coming of age (1865–1945). In 1863 there appeared *The Water-Babies* by Charles Kingsley. In this fascinating, yet repulsive, "Fairy Tale for a Land-Baby," an unctuous cleric and a fanciful poet, uneasily inhabiting one body, collaborated. *The Water-Babies* may stand as a rough symbol of the bumpy passage from the moral tale to a lighter, airier world. Only two years later that passage was achieved in a masterpiece by an Oxford mathematical don, the Reverend Charles Lutwidge Dodgson (Lewis Carroll). *Alice's Adventures in Wonderland* improved none, delighted all. It opened what from a limited perspective seems the Golden Age of English children's literature, a literature in fair part created by Scotsmen: George Macdonald, Andrew Lang, Robert Louis Stevenson, Kenneth Grahame, James Barrie.

The Golden Age in English children's literature

The age is characterized by a literary level decisively higher than that previously achieved; the creation of characters now permanent dwellers in the child's imagination (from Alice herself to Mary Poppins, and including Long John Silver, Mowgli, intelligent Mr. Toad, and—if Hugh Lofting, despite his American residence, be accepted as English—Dr. Dolittle); the exaltation of the imagination in the work of Carroll, Macdonald, Stevenson, E. Nesbit, Grahame, Barrie, Hudson, Lofting, Travers, and the early Tolkien (*The Hobbit* [1938]); the establishment of the art fairy tale (Jean Ingelow with *Mopsa the Fairy* [1869]; Dinah Maria Mulock Craik with *The Little Lame Prince* [1875]; Mrs. Ewing with *Old Fashioned Fairy Tales* [1882]; Barrie's *Peter Pan* [1904]; and the exquisite artifices of Oscar Wilde in *The Happy Prince, and Other Tales* [1888]); the transmutation and popularization, by Andrew Lang, Joseph Jacobs, and others, of traditional fairy tales from all sources; the development of a quasi-realistic school in the fiction of Charlotte M. Yonge (*Countess Kate*); Mrs. Ewing (*Jan of the Windmill*); and Mrs. Molesworth; and, furthering this trend, a growing literary population of real, or at least more real, children (by E. Nesbit and Ransome).

It is further characterized by the rapid evolution of a dozen now-basic genres, including the school story, the historical novel, the vacation story, the "group" or "gang" novel, the boy's adventure tale, the girl's domestic novel, the animal tale, the career novel (Noel Streatfeild's *Ballet Shoes,* 1936), the work of pure whimsy (A.A. Milne's *Winnie-the-Pooh,* 1926); the solution, a brilliant one by Beatrix Potter and a charming one by L. Leslie Brooke, of the problem of creating literature for pre-readers and beginning readers; and the growth of an impressive body of children's verse: the lyric delicacy of Christina Rossetti in *Sing-Song* (1872), the accurate reflection of the child's world in Stevenson's *Child's Garden of Verses,* the satirical nonsense of Hilaire Belloc in his *The Bad Child's Book of Beasts* (1896), the incantatory, other-worldly magic of Walter de la Mare with his *Songs of Childhood* (1902) and *Peacock Pie* (1913), the fertile gay invention of Eleanor Farjeon, and the irresistible charm of Milne in *When We Were Very Young* (1924).

Finally it is characterized by the dominance in children's fiction of middle and upper middle class mores; the appearance, in the late 1930s, with Eve Garnett's *The Family from One End Street,* of stories showing a sympathetic concern with the lives of slum children; the reflection, also in the 30s, of a serious interest, influenced by modern psychology, in the structure of the child's vision of the world; the rise, efflorescence, and decline of the children's magazine: *Boy's Own Magazine* (1855–74), *Good Words for the Young* (1867–77), *Aunt Judy's Magazine* (1866–85), and—famous for its outstanding contributors—*The Boy's Own Paper* (1879–1912); the beginning, with F.J.H. Darton and other scholars, of an important critical-historical literature; institutionalization, commercialization, standardization—the popularity, for example, of the "series"; and the dominating influence of the better English work on the reading taste of American, Continental, and Oriental children.

During these 80 years a vast amount of trash and treacle was produced. What will be remembered is the work of a few dozen creative writers who applied to literature for children standards as high as those ordinarily applied to mainstream literature.

Contemporary times. If the contemporary wood cannot be seen for the trees, it is in part because the number of trees has grown so great. The profusion of English, as of children's books in general, makes judgment difficult. Livelier merchandising techniques (the spread of children's bookshops, for example), the availability of cheap paperbacks, improved library services, serious and even distinguished reviewing—these are among the post-World War II institutional trends helping to place more books in the hands of more children. Slick transformation formulas facilitate the rebirth of books in other guises: radio, television, records, films, digests, cartoon versions. Such processes may also create new child audiences, but that these readers are undergoing a literary experience is open to doubt.

Post-World War II literature

Among the genres that fell in favour, the old moral tale, if not a corpse, surely became obsolescent but raised the question whether it was being replaced by a subtler form of didactic literature, preaching racial, class, and international understanding. The standard adventure story too

seemed to be dying out, though excellent examples, such as *The Cave* (U.S. title, *Five Boys in a Cave* [1950]), by Richard Church, continued to appear. The boy's school story suffered a similar fate, despite the remarkable work of William Mayne in *A Swarm in May* (1955). Children's vese by Ian Serraillier, Ted Hughes, James Reeves, and the later Eleanor Farjeon, excellent though it was, did not speak with the master tones of a de la Mare or the precise simplicity of a Stevenson. In science fiction one would have expected more of a boom; yet nothing appeared comparable to Jules Verne.

Conversely, there was a genuine boom in fact books: biographical series, manuals of all sorts, popularized history, junior encyclopaedias. Preschool and easy-to-read beginners' books, often magnificently produced, multiplied. So did specially prepared decoys for the reluctant reader. After the discovery of the child came that of the postchild: conscientiously composed teen-age and "young adult" novels were issued in quantity, though the quality still left something to be desired. A 19th-century phenomenon—experimentation in the juvenile field by those who normally write for grown-ups—took on a second life after World War II. Naomi Mitchison, Richard Church, P.H. Newby, Richard Graves, Eric Linklater, Norman Collins, Roy Fuller, C. Day Lewis, and Ian Fleming, with his headlong pop extravaganza *Chitty Chitty Bang Bang* (1964), come to mind.

Teenage novels

A post-World War II stress on building bridges of understanding was reflected both in an increase in translations and in the publication of books, whether fiction or nonfiction, dealing responsibly and unsentimentally with the sufferings of a war-wounded world. One example among many was Serraillier's *Silver Sword* (1958), recounting the trans-European adventures that befell four Polish children after the German occupation. *The Silver Sword* was a specialized instance of a general trend toward the interpretation for children of a postwar world of social incoherence, race and class conflict, urban poverty, and even mental pathology. Such novels as John Rowe Townsend's *Gumble's Yard* (1961); *Widdershins Crescent* (1965); *Pirate's Island* (1968); Eve Garnett's *Further Adventures of the Family from One End Street* (1956); and Leila Berg's *Box for Benny* (1958) represented a new realistic school, restrained in England, less so in the United States, but manifest in the children's literature of much of the world. It failed to produce a masterpiece, perhaps because the form of the realistic novel must be moderately distorted to make it suitable for children.

In two fields, however, English postwar children's literature set new records. These were the historical novel and that cloudy area comprising fantasy, freshly wrought myth, and indeed any fiction not rooted in the here and now.

Historical fiction for children

There was fair reason to consider Rosemary Sutcliff not only the finest writer of historical fiction for children but quite unconditionally among the best historical novelists using English. A sound scholar and beautiful stylist, she made few concessions to the presumably simple child's mind and enlarged junior historical fiction with a long series of powerful novels about England's remote past, especially that dim period stretching from pre-Roman times to the coming of Christianity. Among her best works are *The Eagle of the Ninth* (1954), *The Shield Ring* (1956), *The Silver Branch* (1957), *The Lantern Bearers* (1959), and especially *Warrior Scarlet* (1958).

Not as finished in style, but bolder in the interpretation of history in terms "reflecting the changed values of the age," was the pioneering Geoffrey Trease. He also produced excellent work in other juvenile fields. Typical of his highest energies is the exciting *Hills of Varna* (1948), a story of the Italian Renaissance in which Erasmus and the great printer Aldus Manutius figure prominently. Henry Treece, whose gifts were directed to depicting violent action and vigorous, barbaric characters, produced a memorable series of Viking novels of which *Swords from the North* (1967) is typical.

This new English school, stressing conscientious scholarship, realism, honesty, social awareness, and general disdain for mere swash and buckle, produced work that completely eclipsed the rusty tradition of Marryat and George Alfred Henty. Some of its foremost representatives were Cynthia Harnett, Serraillier, Barbara Leonie Picard, Ronald Welch (pseudonym of Ronald O. Felton), C. Walter Hodges, Hester Burton, Mary Ray, Naomi Mitchison, and K.M. Peyton, whose "Flambards" series is a kind of Edwardian historical family chronicle. Leon Garfield, though not working with historical characters, created strange picaresque tales that gave children a thrilling, often chilling insight into the 18th-century England of Smollett and Fielding.

In the realm of imagination England not only retained but enhanced its supremacy with such classics as *Tom's Midnight Garden* (1958), by Ann Philippa Pearce, a haunting, perfectly constructed story in which the present and Victoria's age blend into one. There is the equally haunting Green Knowe series, by Lucy M. Boston, the first of which, *The Children of Greene Knowe,* appeared when the author was 62. The impingement of a world of legend and ancient, unsleeping magic upon the real world is the basic theme of the remarkable novels of Alan Garner. Complex, melodramatic, stronger in action than in characterization, they appeal to imaginative, "literary" children. Garner's rather nightmarish narrative *The Owl Service* (1967) is perhaps the most subtle.

Novels of Tolkien and C.S. Lewis

Finally there is a trio of masters, each the architect of a complete secondary world. The vast Middle Earth trilogy *The Lord of the Rings* (1954–55), by the Anglo-Saxon and Middle English language scholar J.R.R. Tolkien, was not written with children in mind. But they have made it their own. It reworks many of the motives of traditional romance and fantasy, including the Quest, but is essentially a structure, conceivably but not inevitably allegorical, of sheer invention on a staggering scale. It is also a sociocultural phenomenon, selling 3,000,000 copies in nine languages and functioning, for a certain class of American teenagers, as a semisacred cult object.

Tolkien's fellow scholar, C.S. Lewis, created his own otherworld of Narnia. It is more derivative than Tolkien's (he owes something, for example, to Nesbit), more clearly Christian-allegorical, more carefully adapted to the tastes of children. Though uneven, the seven volumes of the cycle, published through the years 1950 to 1956, are exciting, often humorous, inventive, and, in the final scenes of *The Last Battle,* deeply moving.

The third of these classic secondary worlds is in a sense not a creation of fantasy. The four volumes (1952–61) about the Borrowers, with their brief pendant, *Poor Stainless* (1971), ask the reader to accept only a single impossibility, that in a quiet country house, under the grandfather clock, live the tiny Clock family: Pod, Homily, and their daughter Arrietty. All that follows from this premise is logical, precisely pictured, and carries absolute conviction. Many critics believe that this miniature world so lovingly, so patiently fashioned by Mary Norton will last as long as those located at the bottom of the rabbit hole and through the looking glass.

United States. *Overview.* Compared with England, the United States has fewer peaks. In *Huckleberry Finn,* of course, it possesses a world masterpiece matched in the children's literature of no other country. *Little Women,* revolutionary in its day, radiates a century later a special warmth and may still be the most beloved "family story" ever written. Though *The Wonderful Wizard of Oz* has been recklessly compared with *Alice,* it lacks Carroll's brilliance, subtlety, and humour. Nonetheless, its story and characters apparently carry, like *Pinocchio,* an enduring, near-universal appeal for children. To these older titles might be added *Stuart Little* (1945) and *Charlotte's Web* (1952), by E.B. White, two completely original works that appear to have become classics. To this brief list of high points few can be added, though, on the level just below the top, the United States bears comparison with England and therefore any other country.

American classics

The "law" of belated development applies in a special way. From Jamestown to the end of the Civil War, American children's literature virtually depended on currents in England. In the adult field Cooper and Washington Irving may stand for a true declaration of independence. But it was not until the 1860s and '70s, with Mary

Mapes Dodge's *Hans Brinker,* Louisa May Alcott's *Little Women,* Lucretia Hale's *Peterkin Papers,* Mark Twain's *Tom Sawyer,* and *St. Nicholas* magazine, that children's literature finally severed its attachment to the mother country. In the marketplace, however, a uniquely American note was sounded much earlier, the first of the Peter Parley series of Samuel Goodrich having appeared in 1827.

In certain important fields, the United States pioneered. These include everyday-life books for younger readers; the non-class-based small-town story such as *The Moffats* by Eleanor Estes; the Americanized fairy tale and folktale such as *Uncle Remus* (1880), not originally meant for children, and Carl Sandburg's *Rootabaga Stories* (1922); beginners' books such as Dr. Seuss's *The Cat in the Hat* (1957); and the "new realism." One might maintain that American children's literature, particularly that since World War II, is bolder, more experimental, more willing to try and fail, than England's. Moreover, it set new standards of institutionalization, "packaging," merchandising, and publicity, as well as mere production, especially of fact books and "subject series."

Prehistory (1646?–1865). The prehistoric annals are short and simple. Dominated by England, native creativity—to refer only to books with even the thinnest claim to literary quality—amounted to little. The Puritan view of the unredeemable child obtained almost into the era of Andrew Jackson. Jonathan Edwards put it neatly: unrepentant children were "young vipers and infinitely more hateful than vipers." More moderate notions also existed. Imported English ballads and tales, even a few "shockers," were enjoyed by the young vipers. But in general, from John Cotton's *Spiritual Milk for Boston Babes* (1646) through the Civil War, the admonitory and exemplary tract and the schoolmaster's pointer prevailed. Occasionally there is the cheerful note of non-improvement, as in Clement Moore's "Visit from St. Nicholas" (1823), sounding against the successful lesson-cum-moral tales of Peter Parley (Goodrich) and the didactic "Rollo" series of Jacob Abbott. The latter's *Franconia Stories* (1850–53), however, showing traces of Rousseau and Johann Pestalozzi, is the remote ancestor of those wholesome, humorous pictures of small-town child life in which American writers excelled after World War I. Affectionately based on the author's own memories, they occasionally reveal children rather than improvable miniatures of men.

The children's magazines of the early 19th century did their best to amuse as well as instruct the young. Sara Josepha Hale's "Mary Had a Little Lamb" appeared in *The Juvenile Miscellany* (1826–34). The atmosphere was further lightened by *Grandfather's Chair* (1841) and its sequels, retellings of stories from New England history by Nathaniel Hawthorne. These were followed in 1852–53 by his redactions, rather unacceptable today, of Greek legends in *The Wonder Book for Girls and Boys* and *Tanglewood Tales for Girls and Boys.* Hawthorne's death date (1864) coincided roughly with a qualified subsidence of the literature of the didactic.

Peaks and plateaus (1865–1940). During the period from the close of the Civil War to the turn of the century an Americanized white, Anglo-Saxon, Protestant, Victorian gentility dominated as the official, though not necessarily real, culture. At first glance such a climate hardly seems to favour the growth of a children's literature. But counterforces were at work: a vigorous upsurge of interest, influenced by European thinkers, in the education and nurture of children; the dying-out of the old Puritanism; and the accumulation of enough national history to stimulate the imagination. To these forces must be added the appearance in Louisa May Alcott of a minor genius and in Samuel Clemens (Mark Twain) of a major one.

Dominance of Victorian gentility

American materialism (and also its optimism) expressed itself in the success myth of Horatio Alger, while a softened didacticism, further modified by a mild talent for lively narrative, was reflected in the 116 novels of Oliver Optic (William Taylor Adams). But a quartet of books appearing from 1865 to 1880—heralded a happier day. These were Mary Mapes Dodge's *Hans Brinker, or the Silver Skates* (1865), which for all its Sunday-school tone, revealed to American children an interesting foreign culture and told a story that still has charm; Louisa May Alcott's *Little Women* (1868; vol. ii, 1869; and its March family sequels), which lives by virtue of the imaginative power that comes from childhood truly and vividly recalled; Lucretia Hale's *Peterkin Papers* (1880), just as funny today as a century ago, perfect nonsense produced in a non-nonsensical era; and Thomas Bailey Aldrich's *Story of a Bad Boy* (1870). This, it is often forgotten, preceded *Tom Sawyer* by seven years, offered a model for many later stories of small-town bad boys, and is a fair example of the second-class classic. But it took *Tom Sawyer* and *Huckleberry Finn* to change the course of American writing and give the first deeply felt vision of boyhood in juvenile literature.

To these names should be added Frank Stockton (whose *Ting-a-Ling Tales* [1870] showed the possibilities inherent in the invented fairy tale) and especially the writer-illustrator Howard Pyle. His reworkings of legend (*The Merry Adventures of Robin Hood,* 1883; the King Arthur stories, 1903–1910, and his novels of the Middle Ages [*Otto of the Silver Hand,* 1888; and *Men of Iron,* 1892]) exemplify perfectly the romantic feeling of his time, as does the picture of Shakespeare's England drawn by John Bennett in *Master Skylark* (1897).

Contributions of Howard Pyle

The sentimentality that is sometimes an unconscious compensatory gesture in a time of ruthless materialism expressed itself in the idyllic *Poems of Childhood* (1896), by Eugene Field, and the rural dialect *Rhymes of Childhood* (1891), by James Whitcomb Riley. These poems can hardly speak to the children of the second half of the 20th century. But it is not clear that the same is true of the equally sentimental novels of Frances Hodgson Burnett. It is easy to smile over *Little Lord Fauntleroy* (1886) or her later and superior novels, *A Little Princess* (1905) and *The Secret Garden* (1911). Back of the absurd sentimentality, however, lies an extraordinary narrative skill, as well as an ability to satisfy the perennial desire felt by children at a certain age for life to arrange itself as a fairy tale.

Development of a junior literature

The development of a junior literature from 1865 to about 1920 is ascribable less to published books than to two remarkable children's magazines: *The Youth's Companion* (1827–1929, when it merged with *The American Boy*) and the relatively nondidactic *St. Nicholas* magazine (1873–1939), which exerted a powerful influence on its exclusively respectable child readers. (It is surely needless to point out that up to the 1960s children's literature has been by and for the middle class). These magazines published the best material they could get, from England as well as the United States. For all their gentility, standards, including that of illustration, were high. The contributors' names in many cases became part of the canon of world literature. To the children of the last quarter of the 19th and first quarter of the 20th century, the periodical delivery of these magazines presumably meant something that film and television cannot mean to today's children. The magazines were not "media." They were friends.

Appropriately the new century opened with a novelty: a successful American fairy tale. *The Wonderful Wizard of Oz* (1900) is vulnerable to attacks on its prose style, incarnating mediocrity. But there is something in it, for all its doctrinaire moralism, that lends it permanent appeal: a prairie freshness, a joy in sheer invention, the simple, satisfying characterization of Dorothy and her three old, lovable companions. Several of the sequels—but only those bearing L. Frank Baum's name—are not greatly inferior.

The century underwent for the next two decades a rather baffling decline. Some institutional progress was made in library development, professional education, and the reviewing of children's books. Much useful work was also accomplished in the field of fairy-tale and folktale collections. But original literature did not flourish. There were Pyle and Mrs. Burnett and the topflight nonsense verses of Laura E. Richards, whose collected rhymes in *Tirra Lirra* (1932) will almost bear comparison with those of Edward Lear. Less memorable are the works of Lucy Fitch Perkins, Joseph Altsheler, Ralph Henry Barbour, Kate Douglas Wiggin, Eliza Orne White, and the two Burgesses—Thornton and Gelett. During these decades, de la Mare, Miss Potter, Kipling, Barrie, Grahame, and E. Nesbit were at work in England.

During the period between world wars new trails were blazed in nonfiction with van Loon's *Story of Mankind* and V.M. Hillyer's *Child's History of the World* (1922). The *Here and Now Story Book,* by Lucy Sprague Mitchell, published in the 1920s, was the first real example of the "direct experience" school of writing, but it is more properly part of the chronicle of pedagogy than of literature. The small child was far better served by a dozen talented writer-illustrators, such as Wanda Gâg, with her classic *Millions of Cats* (1928) and other delightful books; and Ludwig Bemelmans, with *Madeline* (1939) and its sequels. Other distinguished names in the important and growing picture-book field were Marjorie Flack, Hardie Gramatky, James Daugherty, the d'Aulaires, and Virginia Lee Burton.

In the field of comic verse and pictures for children of almost all ages, Dr. Seuss (Theodore Geisel), starting with *And to Think That I Saw It on Mulberry Street* (1937), continued to lead, turning out so many books that one tended to take him for granted. His talent is of a very high order.

Literature of the 1920s and '30s

The 1920s and '30s produced many well-written historical novels, striking a new note of authority and realism, such as *Drums* (1925, transformed in 1928 into a boy's book with N.C. Wyeth's illustrations), by James Boyd, and *The Trumpeter of Kracow* (1928), by Eric Kelly. The "junior novel" came to the fore in the following decade, together with an increase in books about foreign lands, minority groups, and a boom in elaborate picture books. Children's verse was well served by such able practitioners as Dorothy Aldis and Rosemary and Stephen Vincent Benét, with their stirring, hearty ballad-like poems collected in *A Book of Americans* (1933). But the only verse comparable to that of Stevenson or de la Mare was the exquisite *Under the Tree* (1922), by the novelist Elizabeth Madox Roberts, a treasure that should never be forgotten.

At least three other writers produced work of high and entirely original quality. Two of them—Florence and Richard Atwater—worked as a pair. Their isolated effort, *Mr. Popper's Penguins* (1938), will last as a masterpiece of deadpan humour that few children or adults can resist. The third writer is Laura Ingalls Wilder. Her *Little House* books, nine in all, started in 1932 with *The Little House in the Big Woods.* The entire series, painting an unforgettable picture of pioneer life, is a masterpiece of sensitive recollection and clean, effortless prose.

Work of quality was contributed during these two lively decades by authors too numerous to list. Among the best of them are Will James, with his horse story *Smoky* (1926); Rachel Field, whose *Hitty* (1929) is one of the best doll stories in the language; Elizabeth Coatsworth, with her fine New England tale *Away Goes Sally* (1934); and the well-loved story of a New York tomboy in the 1890s, *Roller Skates* (1936), by the famous oral storyteller Ruth Sawyer.

Contemporary times. Since the 1930s the quality and weight of American children's literature were sharphy affected by the business of publishing, as well as by the social pressures to which children, like adults, were subjected. Intensified commercialization and broad-front expansion had some good effects and some bad ones as well.

For any book of interest to adults, publishers constructed a corresponding one scaled to child size. The practice of automatic miniaturization stimulated a pullulation of fact books—termed by an unsympathetic observer "the information trap"—marked by a flood of subject series and simplified technology. Paperbacks and cheap reprints of juvenile favourites enlarged the youthful reading public, just as the multiplication of translations widened its horizon. More science fiction was published, a field in which the stories of Robert Heinlein and *A Wrinkle in Time* (1962), by Madeleine L'Engle, stood out. An increase was also noticeable in books for the disadvantaged child and in work of increasingly high quality by and for blacks. In the early 1950s, children's book clubs flourished, though they appeared to be on the wane little more than a decade later. Simple narration using "scientifically determined vocabulary" also seemed to decrease in popularity. There was a marked tendency to orient titles, fiction and nonfiction, to the requirements of the school curriculum. Another trend was toward collaborative "international" publishing. This had the double effect of cutting colour-plate costs and promoting blandness, since it was important that no country's readers be offended or surprised by anything in text or illustration. Still another alteration took place in the conventional notion of age and grade levels. Teenagers reached out for adult books; younger children read junior novels.

Growth of "realist" fiction

The most striking development was the growth of the "realists," most of them as earnest as Maria Edgeworth, a few of them lighter fingered, with a fringe of far-outers. The latter were fairly represented by Nat Hentoff in *Jazz Country* (1965), for example, and Maria Wojciechowska in *The Rotten Years* (1971). Teenage fiction as well as nonfiction dealt mercilessly with ethnic exploitation, poverty, broken homes, desertion, unemployment, adult hypocrisy, drug addiction, sex (including homosexuality), and death. A whole new "problem" literature became available, with no sure proof that it was warmly welcomed. The aesthetic dilemmas posed by this literature are still to be faced and resolved. The new social realist story often had the look of an updated moral tale: the dire consequences of nondiligence were replaced by those of pot smoking.

Nevertheless such original works as *Harriet the Spy* (1964) and *The Long Secret* (1965), by Louise Fitzhugh, showed how a writer adequately equipped with humour and understanding could incorporate into books for 11-year-olds subjects—even menstruation—ordinarily reserved for adult fiction. Similarly trailblazing were the semidocumentary novels of Joseph Krumgold: . . . *And Now Miguel* (1953), *Onion John* (1958), and *Henry 3* (1967), the last about a boy with an I.Q. of 154 trying to get along in a society antagonistic to brains. The candid suburban studies of E.L. Konigsburg introduced a new sophistication. Her 1968 Newbery Medal winner, *From the Mixed-Up Files of Mrs. Basil E. Frankweiler,* was original in its tone and humour.

As for the more traditional genres, a cheering number of high-quality titles rose above the plain of mediocrity. The nonfantastic animal story *Lassie Come Home* (1940), by Eric Knight, survived adaptation to film and television. In the convention of the talking animal, authentic work was produced by Ben Lucien Burman, with his wonderful "Catfish Bend" tales (1952–67). The American-style, wholesome, humorous family story was more than competently developed by Eleanor Estes, with her "Moffat" series (1941–43) and *Ginger Pye* (1951); Elizabeth Enright, with her Melendy family (1941–44); and Robert McCloskey, with *Homer Price* (1943)—to name only three unfailingly popular writers. Text-and-picture books for the very young posed an obdurate challenge: to create literature out of absolutely simple materials. That challenge, first successfully met by Beatrix Potter, attracted Americans. The modern period produced many enchanting examples of this tricky genre: *The Happy Lion* (1954) and its sequels, the joint work of the writer Louise Fatio and her artist husband, Roger Duvoisin; the "Little Bear" books, words by Else Holmelund Minarik, pictures by Maurice Sendak; and several zany tours de force by Dr. Seuss, including his one-syllable revolution *The Cat in the Hat* (1957). The picture books of Sendak, perhaps one of the few original geniuses in his restricted field, were assailed by many adults as frightening or abnormal. The children did not seem to mind.

Fiction about foreign lands

Fiction about foreign lands boasted at least one modern American master in Meindert De Jong, whose most sensitive work was drawn from recollections of his Dutch early childhood. A Hans Christian Andersen and Newbery winner, he is best savoured in *The Wheel on the School* (1954), and especially in the intuitive *Journey from Peppermint Street* (1968). The historical novel fared less well in America than in England. *Johnny Tremain* (1943), by Esther Forbes, a beautifully written, richly detailed story of the Revolution, stood out as one of the few high points, as did *The Innocent Wayfaring* (1943), a tale of Chaucer's England by the equally scholarly Marchette Chute. Poetry for children had at least two talented representatives. One was the eminent poet-critic John Ciardi, the other

David McCord, a veteran maker of nonsense and acrobat of language.

In fantasy, the farcical note was struck with agreeable preposterousness by Oliver Butterworth in *The Enormous Egg* (1956) and *The Trouble with Jenny's Ear* (1960). The prolific writer-illustrator William Pène Du Bois has given children nothing more uproariously delightful than *The Twenty-one Balloons* (1947), merging some of the appeals of Jules Verne with those of Samuel Butler's *Erewhon* and adding a sly humour all his own. Two renowned *New Yorker* writers, James Thurber and E.B. White, developed into successful fantasists, Thurber with an elaborate series of ambiguous literary fairy tales such as *The Thirteen Clocks,* White with his pair of animal stories *Stuart Little* and *Charlotte's Web* that for their humanity and uninsistent humour stand alone. The vein of "high fantasy" of the more traditional variety, involving magic and the construction of a legendary secondary world, was represented by the five highly praised volumes of the Prydain cycle (1964–68) by Newbery Medal winner Lloyd Alexander.

Two other works of pure imagination gave the 1960s some claim to special notice. The first was *The Phantom Tollbooth* (1961) by Norton Juster, a fantasy about a boy "who didn't know what to do with himself." Not entirely unjustly, it has been compared to *Alice.* The second received less attention but is more remarkable: *The Mouse and His Child* (1969), by Russell Hoban, who had been a successful writer of gentle tales for small children. But here was a different affair altogether: a flawlessly written, densely plotted story with quiet philosophical overtones. It involved a clockwork mouse, his attached son, and an unforgettable assortment of terribly real, humanized animals. Like *Alice* and *The Borrowers* —indeed like all major children's literature—it offered as much to the grown-up as to the young reader. With this moving, intellectually demanding fantasy the decade ended on a satisfactory note.

Germany and Austria. A. Merget's *Geschichte der deutschen Jugendliteratur* ("History of German Children's Literature") appeared in 1867, some years before the Germans had much children's literature to consider, a demonstration of Teutonic thoroughness. By two criteria—degree of awareness of the child's identity and level of institutional development—Germany leads the world. It has built a vast structure of history, criticism, analysis, and controversy devoted to a subject the chief property of which would appear to be its charm rather than its obscurity. One estimate has it that in West Germany alone there are over 300 associations dedicated to the study and promotion of children's literature. Such conscientiousness, nowhere else matched, such a serious desire to relate the child's reading to his nurture, education, and *Weltanschauung,* has an admirable aspect. But by attaching juvenile books too closely to the theory and demands of pedagogy, it may have constricted a marked native genius.

Historical influences

The dominant historical influences roughly coincide with those that have affected German mainstream literature, though, as expected, they were exerted more slowly. The Reformation, stressing the Bible, the catechism, and the hymnbook, bent the literature of childhood toward the didactic, the monitory, and the pious. The Enlightenment, however, did something to help toward the identification of the child as an independent being. With this insight are associated the educational theories of J.B. Basedow, J.F. Herbart, and Friedrich Froebel. One fruit of the movement was *Robinson der Jüngere* (1779; "The Young Robinson"), by Joachim Heinrich Campe, who adapted Defoe along Rousseauist lines, his eye sharply fixed on what he considered to be the natural interests of the child. Interchapters of useful moral conversations between the author and his pupils were a feature of the book. Campe's widespread activities on behalf of children, though less commercially motivated, recall Newbery's.

Rationalism, piety, and the German partiality for disciplined conduct were modified by the influence of two crucial works, not intended for children but soon taken over by them. Both are part of the Romantic movement that swept Germany and much of the Continent during the early 19th century. *Des Knaben Wunderhorn* (1805–08; "The Youth's Magic Horn"), a collection of old German songs and folk verse, included many children's songs, or songs that were so denominated by the editors, Achim von Arnim and Clemens Brentano. The effect of the book was to retrieve for Germany much of its rich folk heritage, to promote a new emotional sensibility, and to draw attention to the link, as the Romantics thought, binding folk feeling to the child's vision of the world. *Des Knaben Wunderhorn* became a part of German childhood, as La Fontaine's *Fables* in France and *Mother Goose* in England had become a part of growing up in those countries. It helped inspire several excellent writers of verse for children: A.H. Hoffmann von Fallersleben; August Kopisch; the writer-illustrator Count Franz Pocci, the first German to write nonsense verse for the young; F.W. Güll; and later poets such as Paula and Richard Dehmel.

German fairy tales

Just as *Des Knaben Wunderhorn* became a source of poetry, so the epochal folktale collection of the brothers Grimm helped to develop a school of prose fairy-tale writers. Not all of these Romantics wrote with children in mind. But some of the simplest of their tales have become part of the German child's inheritance. In today's presumably practical era, they are once more in favour. Among these masters of the "art" *Märchen* are E.T.A. Hoffmann; C.M. Brentano; Ludwig Tieck; de la Motte Fouqué, author of *Undine;* and Wilhelm Hauff, whose talents are most nearly adapted to the tastes of children.

Two curious half-geniuses of comic verse and illustration wrote and drew for the hitherto neglected small child. *Struwwelpeter* ("Shock-headed Peter"), by the premature surrealist Heinrich Hoffmann, aroused cries of glee in children across the continent. Wilhelm Busch created the slapstick buffoonery of Max and Moritz, the ancestors of the Katzenjammer Kids and indeed of many aspects of the comic strip.

The second half of the 19th century saw an increase in commercialized sentimentality and sensation and a corresponding decline in quality. The bogus Indian and Wild West tales of Karl May stand out luridly in the history of German children's literature. Up to about 1940, 7,500,000 of his books had been sold to German readers alone. (Emilio Salgari in Italy, G.A. Henty in England, and "Ned Buntline" in the United States, who were contemporaneously satisfying the same hunger for the suspenseful, did not approach's May's talent for fabrication without the slightest root in reality.)

It may have been May and others like him who roused an educator, Heinrich Wolgast, to publish in 1896 his explosive *Das Elend unserer Jugendliteratur* ("The Sad State of Our Children's Literature"). The event was an important one. It advanced for the first time the express thesis that "Creative children's literature must be a work of art"; Wolgast resolutely decried nationalistic and didactic deformations. He precipitated a controversy the echoes of which are still audible. On the whole his somewhat excessive zeal had a wholesome effect.

Two post-Wolgast poets of childhood worthy of mention are Christian Morgenstern, whose macabre, pre-Dada poetry for adults later came into vogue, and the lesser-gifted Joachim Ringelnatz. The nondidactic note they sounded in modern times was strengthened by a whole school of children's poets. No other country produced work in this difficult field superior to the finest verse of the multitalented James Krüss, and especially Josef Guggenmos, whose lyric simplicity at times recalls Blake. Guggenmos also has to his credit a translation of *A Child's Garden of Verses,* in itself an original work of art.

Between the world wars, prose showed few high points and, after the advent of Hitler, many low ones. *Der Kampf der Tertia* (1927; "The Third-Form Struggle"), by Wilhelm Speyer, was Germany's excellent contribution to the genre of the school story. Erich Kästner's *Emil and the Detectives* (1929) ranked not only as a work of art, presenting city boys with humour and sympathy, but as an immediate classic in an entirely new field, the juvenile detective story (Mark Twain's awkward *Tom Sawyer, Detective* [1896] may be ignored). Kästner, the dean of German writers for children, won an international audience with a long series of stories of which the thesis-fable *Die Konferenz der Tiere* (1949; Eng. trans. *The Animals'*

Conference, 1949) is perhaps the funniest as well as the most serious.

Post-World War II literature

Post-World War II literature, recovering from the Nazi blight, was strong in several fields. In realistic fantasy there is *Vevi* (1955), by the Austrian Erica Lillegg, an extraordinary tale of split personality, odd, exciting, even profound. Michael Ende's *Jim Knopf und Lucas der Lokomotivführer* (1961; Eng. trans., *Jim Button and Luke the Engine Driver,* 1963) has more than a touch of *Oz;* and both Kästner and Krüss have made agreeable additions to the realm of fantasy.

In the domain of the historical novel, Hans Baumann is a distinguished name. Lacking the narrative craft of Miss Sutcliff, whose story lines are always clean and clear, he matched her as a scholar and mounted scenes of great intensity in such novels as *Die Barke der Brüder* (1956; Eng. trans., *The Barque of the Brothers,* 1958) and especially *Steppensöhne* (1954; Eng. trans., *Sons of the Steppe,* 1958), a tale about two grandsons of Genghis Khan. His narrative history of some exciting archaeological discoveries, *Die Höhlen der grossen Jäger* (1953; Eng. trans., *The Caves of the Great Hunters,* 1954; rev. ed., 1962), is a minor classic. Mention should be made of Fritz Mühlenweg, a veteran of the Sven Hedin expedition of 1928–32 to Inner Mongolia and the author of *Grosser-Tiger und Kompass-Berg* (1950; Eng. trans., *Big Tiger and Christian,* 1952). A long, richly coloured narrative of a journey made by two boys, Chinese and European, through the Gobi Desert, it should stand as one of the finest adventure stories of the postwar years.

One general conclusion regarding West German children's literature after 1945 was that the native genius, which had been impeded by pedagogical theory and nationalist dogma, again appeared to be in free flow.

In East Germany, production was conditioned by the association with the Soviet Union, and it appeared to be recapitulating the developments in children's literature that had occurred in the Soviet Union after 1917. Socialist Realism was the basic food offered to the literary appetites of young East Germans.

Scandinavia. *Sweden.* Scandinavia, but especially Sweden, inevitably suggests a question as to why a group of small, sparsely populated countries ranks directly after England and the United States for the variety, vigour, and even genius of its children's literature. Hazard's north–south theory describes; it does not explain. A few possible factors may be listed: the inspiration of the master Andersen—yet he does not seem greatly to have inspired his homeland; the appearance in 1900 of the Swedish Ellen Key's two-volume *Barnets århundrede* (Eng. trans., *The Century of the Child,* 1909), pivotal in the history of the discovery that children really exist; a general modern atmosphere of social enlightenment; welfare statism tempered by regard for the individual; a school and library system, notably in Sweden, of extraordinary humanity and efficiency; perhaps even the long, lively career of the Stockholm Children's Theatre, a centre of creative activity. Yet the mystery persists. Since the first half of the 19th century, Scandinavia produced Andersen, Zacharias Topelius, Jørgen Moe, Henrik Wergeland, Helena Nyblom, Selma Lagerlöf, Elsa Beskow, Astrid Lindgren, Tove Jansson, Maria Gripe, Anna Lisa Warnlöf, Lennart Hellsing, Karin Anckarsvärd, Inger Sandberg, plus a school of critics and historians second only to that of Germany, plus many talented illustrators.

Children's literature in Sweden for centuries reflected that of Germany, of which Sweden was a cultural province during the Reformation and even through the Enlightenment period. The historian Göte Klingberg traced some kind of religious-instructive reading for children back to 1600. There is a record, though the manuscripts have vanished, of children's plays produced at the country manors during the 1700s and into the following century. The tradition of children's theatre has always been stronger in Sweden than elsewhere in Europe.

True Swedish national literature

A true native literature is usually dated from 1751-53, when the tutor Count Carl Tessin wrote his "Old Man's Letters to a Young Prince" (Gustav III), in which instruction was tempered by the first fairy tales written for Swedish children. The German influence, however, persisted until about the middle of the 19th century, when Fredrika Bremer, traveller and feminist, tried to stimulate the work of indigenous children's writers. The dominant influence of the Finnish-born but basically Swedish Topelius, of Hans Christian Andersen, and of the romantic spirit in general was felt at this time. Later in the century two followers of Andersen—Helena Nyblom and Anna Wahlenberg—enriched the tradition of the fairy tale. The former's *Sagokrans* (1903; Eng. trans., *The Witch of the Woods,* 1968), preserves a rare charm.

The great landmark, however, is Miss Lagerlöf's world classic *Nils Holgerssons underbara resa genom Sverige,* 2 vol. (1906–07; Eng. trans., *The Wonderful Adventures of Nils,* 1907; *Further Adventures of Nils,* 1911). Written (at the request of the state ministry of education) as a school geography, it is the rare example of an officially commissioned book that turned out to be a work of art. *Nils,* for all its burden of instruction, is a fantasy. At the same time, a realistic breakthrough was achieved by Laura Fitinghoff, whose historical novel about the famine of the 1860s, *Barnen från Frostmofjället* (1907; Eng. trans., *Children of the Moor,* 1927), ranks as a classic.

According to the historian Eva von Zweigbergk, didacticism ("diligence, obedience, and moderation") obtained up to the 1920s, though she also views the period 1890–1915 as Sweden's Golden Age. It included not only *Nils* but the emergence of a school of creators of picture books for small children headed by Elsa Beskow, whose work in pictures and text, extending over the years from 1897 to 1952, was decisive in its influence. This pre-modern period also saw many good writers for grown-ups devoting their talents to juvenile fiction. The sailing story *Mälarpirater* (1911; "The Pirates of Lake Mälaren"), by the novelist Sigfrid Siwertz, is a still-remembered example.

The period from 1940 on has called forth a bewildering array of talented writers and artist-writers. In the field of humour and nonsense there are Åke Holmberg, with his parodic Ture Sventon detective series; the outstanding poet Lennart Hellsing, with *Daniel Doppsko* (1959); Astrid Lindgren, successful in a half dozen genres but perhaps best known as the creator of the supergirl Pippi Longstocking; Gösta Knutsson, with her well-liked *Pelle svanslös* (1939; Eng. trans., *The Adventures of the Cat Who Had No Tail*). The psychological realistic novel, delving deeply into the inner lives of children, has been developed by Maria Gripe, whose *Hugo and Josephine* trilogy may become classic; Gunnel Linde's *Tacka vet jag Skorstensgränd* (1959; Eng. trans., *Chimney-Top Lane,* 1965); and Anna Lisa Warnlöf, writing under the pseudonym of "Claque," whose two series about Pella and Fredrika show an intuitive understanding of lonely and misunderstood children.

Harry Kullman and Martha Sandwall-Bergström are among the few Swedish writers who have used working class industrial backgrounds successfully. Kullman is also a historical novelist. The prolific Edith Unnerstad has written charming family stories, with a touch of fantasy, as has Karin Anckarsvärd, whose *Doktorns pojk'* (1963; Eng. trans., *Doctor's Boy,* 1965) is a quietly moving tale of small-town life in the horse-and-buggy days. The Sandbergs, Inger and Lasse, have advanced the Beskow tradition in a series of lovely picture books. Fantasy has been well served by Lindgren, Edith Unnerstad, Holmberg, Hellsing, and others. Children's poetry is a lively contemporary art, one distinguished poet being Britt G. Hallqvist.

By most criteria of development the Swedes rank high among those creating a children's literature that is both broad and deep.

Norway. Norway cannot boast a genius of worldwide fame. But, beginning with the 1830s when a new literary language, based on spoken Norwegian, was forged, Norway has possessed an identifiable children's literature. From 1837 to 1844 Asbjørnsen and Moe, the Grimms of Norway, published their remarkable collection of folk stories, and thus created not only a literary base on which the future could build but a needed sense of national identity. Moe also wrote specifically for children. His poems are part of Norwegian childhood, and his nature fantasy *I brønden*

og i tjernet ("In the Well and the Lake," 1851) made Viggo and his little sister Beate familiar for more than a century. Equally enduring are the fairy tales and children's verse of Norway's greatest poet Henrik Wergeland.

Golden Age of Norwegian literature

The Norwegian critic Jo Tenfjord believes that the 30 years from 1890 to 1920 represented a golden age. With this period are associated Dikken Zwilgmeyer, author of the "Inger Johanne" series about a small-town little girl; Barbra Ring, creator of the popular "Peik" stories and of a play *The Princess and the Fiddler,* which was produced yearly at the National Theatre in Oslo; Gabriel Scott; and the fairy-tale writer Johan Falkberget.

Among the more prominent and well-loved moderns are Halvor Floden, whose most famous work, centred on a gypsy waif, is *Gjenta fra lands vegen* ("The Girl from the Road"); the nonsense versifier Zinken Hopp; the poet Jan-Magnus Bruheim, three of whose collections have won state prizes; Finn Havrevold, whose toughminded boys' teenage novel *Han Var Min Ven* became available in English translation as *Undertow* in 1968, and who also wrote successfully for girls; Leif Hamre, specializing in air force adventures; the prolific, widely translated Aimée Sommerfelt, whose works range from "puberty novels" to faraway stories set in Mexico City and northern India; Thorbjørn Egner, who is the author of, among other books, a tiny droll fantasy, *Karius and Baktus* (1958; Eng. trans. 1962), which will actually persuade small children to brush their teeth; and Alf Prøysen, creator of Mrs. Pepperpot, a delightful little old lady who never knows when she is going to shrink to pepperpot size. Fantasy of this kind seems less characteristic of contemporary Norway than does the realistic novel, especially that designed for older children.

Denmark. Without Hans Christian Andersen, Danish children's literature might have fared better. It is not that his countrymen deify him, as much as it is that the outside world does. Indeed, because modernized versions of his tales do not exist, his now rather antiquated Danish tends to outmode him. Yet his gigantic shadow must have intimidated his literary descendants, just as Dante and Cervantes intimidated theirs. Doubtless other forces also account for the sparseness and relative conventionality of Danish children's literature.

The earliest books were written for the children of the nobility. Not till the passage of the Education Act of 1814 did the poorer ones have access to any suitable reading matter, and this, obedient to the prevailing European fashion, was dour in tone. The climate, of course, relaxed when Andersen appeared with his phenomenal series, still the finest of their kind, of invented or reworked fantastic tales. In 1884 H.V. Kaalund published a picture book of "Fables for Children" based on the popular verse narratives (1833) of a Thüringian pastor, Wilhelm Hey. Three years later an unidentified Danish humorist added three cautionary tales to a translation of six *Struwwelpeter* stories. Though it does not seem to have appeared as a picture book until 1900, Christian Winther in 1830 wrote a pleasing trifle, with an unusual fantastic touch, called "Flugten til Amerika" ("Flight to America"). It is still ranked as a classic. Such are some of the 19th-century oases.

Denmark's general tendency has been to over-rely on translations or adaptations, drawn especially from its neighbour Germany. As against this, it can point to an excellent original tradition of nursery and nonsense rhymes.

Early nursery and nonsense rhymes

The first such collection, made as early as 1843, stimulated not only Andersen but such other 19th-century figures as Johan Krohn, whose "Peter's Christmas" remains a standard seasonal delight. The tradition is relayed to the 20th century by Halfdan Rasmussen, whose collected *Bjørnerim* ("Verse for Children") won the 1964 Danish Children's Book Prize, and Ib Spang Olsen, with his nonsense picture book *The Boy in the Moon* (1962). As for the complementary prose tradition of fireside tales, Denmark had to wait (Andersen was artist, not scholar) for its Grimm until 1884, when a collection made by Svend Grundtvig, the son of N.F.S. Grundtvig, a great bishop-educator, was posthumously published.

As compared with other Scandinavian countries, post-World War II developments lagged. Picture books exhibited much more originality than did teenage literature. Jytte Lyngbirk's girls' novels, notably the love story "Two Days in November," however, are well reputed, as are the realistic fictions, laid against an industrial background, of Tove Ditlevsen. Perhaps Denmark's boldest original talent is Anne Holm, who aroused healthy controversy with her (to some) shocking narrative of a displaced boy's journey to Denmark, the novel *David* (1963; Eng. trans., *North to Freedom,* 1965).

Some informed observers ascribe Denmark's only moderate performance to domination by the teaching profession, to the lingering influence of conventional didacticism, and to the lack of the economic-social forces that stimulate professional writers. As late as 1966 the Minister of Culture commented on the scarcity of Danish juvenile authors, and this at a time when the rest of Scandinavia was, as it remained, in the full flood of the modern movement.

Finland. Although its language and people are not of European origin, Finland is loosely conceived as part of the Scandinavian bloc. Only since December 6, 1917, has it been formally independent. During much of its history Swedish was the language of the educated class. Thus its two outstanding premodern children's writers, the father figure Zacharias Topelius and Anni Swan, wrote their fairy tales and folktales primarily for a Swedish-reading audience. Their works however were promptly translated into Finnish and became part of the native heritage. The same is true of the contemporary Tove Jansson, 1966 Andersen Medal winner, whose series of novels about the fantastic self-contained world of Moomintrolls, though less successful with English-reading children, enchants young readers throughout northern and central Europe.

The labours of Topelius in the children's field and of Elias Lönnrot (compiler of the great Finnish epic-miscellany the *Kalevala,* 1835) in the field of national folklore constituted the soil from which Finnish children's literature was eventually to derive nutriment. But that literature emerged as an identifiable whole only after World War I. It is largely folktale rooted. Indeed this small country became an international focus of folklore research. One student has said that it probably possesses the largest number of folktales in existence, some 30,000 of them. In the early 1960s a fairy tale competition yielded 795 manuscripts, a phenomenal statistic in view of Finland's sparse population.

Finland, despite the fact that its language tends to limit its audience, is part of the main current of children's literature, even though only Jansson has won anything like an international reputation. Two children's poets, Aila Meriluoto and Kirsi Kunnas, have achieved renown.

France. *Overview.* The French themselves are not happy with their record. Writing in the late 1940s, critic Jean de Trigon, in *Histoire de la littérature enfantine, de ma Mère l'Oye au Roi Babar* (Paris, Librairie Hachette, 1950) said: "The French have created little children's literature. They have received more than they have given, but they have assimilated, adapted, transformed. The two are not the same thing, for one must love childhood in general if one is to please children other than one's own." In 1923 Marie-Thérèse Latzarus tolled the passing bell in *La littérature enfantine en France dans la seconde moitié du XIX^e siècle* (Paris; Les Presses Universitaires de France): "Children's literature, more's the pity, is dying." And in 1937, in their introduction to *Beaux livres, belles histoires,* the compilers Marguerite Gruny and Mathilde Leriche wrote: "Children's literature in France is still poor, despite the earnest efforts of the last decade."

Criticism of French children's literature

Surely Trigon was too severe. Even more surely Mlle Latzarus has proved a false Cassandra. As for the compilers, the very decade they scorned saw at least three magnificent achievements. The first was Jean de Brunhoff's. Equally talented as author and artist, in 1931 he gave the world that enlightened monarch Babar the Elephant, one of the dozen or so immortal characters in children's literature. The next year saw the start of Paul Faucher's admirable Père Castor series, imaginatively conceived, beautifully designed educational picture books for the very young—not literature, perhaps, but historically comparable to Comenius. Finally, in 1934 appeared the first of Marcel Aymé's miraculous stories about two little girls and the talking animals whose adventures they shared. These

grave-comic fantasies were later collected as *Les Contes du chat perché* (1939; Eng. trans., *The Wonderful Farm,* 1951; *Return to the Wonderful Farm,* 1954), and, along with de Brunhoff and Faucher, were enough to make the decade great.

But there are no other decades to match it. There does exist a disproportion between French literary genius as a whole and the children's literature it has been able to produce. The explanation is uncertain. Mme Le Prince de Beaumont, an adventurous 18th-century lady who wrote over 70 volumes for the young, thought that children's stories should be pervaded by "the spirit of geometry." It is possible that the blame for France's showing might in part be laid on a persistent Cartesian spirit, reinforced by rationalist and positivist philosophies. The Cartesian does not readily surrender to fancy, especially of the more wayward variety. And so, even counting Charles Perrault, the later Charles Nodier, and the contemporary Simone Ratel and Maurice Vauthier, a dearth of first-rate fairy tales may be noted. Cartesians would tend to be weak also in children's verse, in nonsense of any sort, in humour (despite Babar), even in the more imaginative kind of historical novel exemplified by Hans Baumann in Germany and Rosemary Sutcliff in England. Perhaps French children's literature has been restrained by a Catholicism or by a Protestantism that continued to insist on the edifying when mainstream literature had already freed itself from explicit moralism. It may not even be true, as Trigon thinks, that the French have fruitfully assimilated the children's literature of foreign countries. *Alice* has more or less bewildered them; *Huckleberry Finn* has never been digested. The child's cause was not much aided by the triumph of a post-Napoleonic bourgeois cast of thought—or by the wave of post-1871 nationalism.

It is a complicated problem. But perhaps the heart of it lies in the value the French set on maturity. For them childhood at times has seemed less a normal human condition than a handicap. The children themselves have often seemed to feel the pressure, which may account for the fact that they absorb French adult books precociously. The French came much later than did many other countries to the discovery of the child as a figure worthy of the most sensitive understanding; that is what makes Père Castor so important. One is not surprised to note the comparatively recent date (1931) of a study by Aimé Dupuy, translatable as *The Child: A New Character in the French Novel.*

History. If one skips Jean de La Fontaine, whose *Fables* (1668; 1678–79; and 1693), though read by the young, were not meant for them, French children's literature from one point of view begins with the classic fairy tales of Charles Perrault. These were probably intended for the salon rather than the nursery, but their narrative speed and lucidity commended them at once to children. The fairy tales of his contemporary Mme d'Aulnoy, like many others produced in the late 17th and early 18th centuries, are hardly the real thing. With a Watteau-like charm, they taste of the court, as does the *Télémaque* of François Fénelon, a fictionalized lecture on education.

Perrault's fairy tales

Rousseau, as has been noted, did make a difference. *Émile* at least drew attention to what education might be. But the effect on children's literature was not truly liberating. His disciple, Mme de Genlis, set a stern face against make-believe of any sort; all marvels must be explained rationally. Her stories taught children more than they wanted to know, a circumstance that endeared her to a certain type of parent. Sainte-Beuve, to be fair, called her "the most gracious and gallant of pedagogues." One of her qualities, priggishness, was energetically developed by Arnaud Berquin in his *Ami des enfants.* Berquin created the French equivalent of the concurrent English bourgeois morality. In effect, he unconsciously manufactured an adult literature for the young, loading the dice in favour of the values held by parents to be proper for children. Yet one must beware of judging Berquin or his equally moralistic successor Jean-Nicolas Bouilly by today's standards. Children accepted them because they were the best reading available; and Anatole France's tribute in *Le Petit Pierre* (1918) shows that they must have exerted some charm.

The didactic strain, if less marked than in England or Germany, persisted throughout most of the 19th century. To it, Mme de Ségur, in her enormously popular novels, added sentimentality, class snobbery, but also some liveliness and occasional fidelity to child nature. Her "Sophie" series (1850s and 60s), frowned on by modern critics, is still loved by obstinate little French girls. *Sans Famille* (1878), by Hector Malot, a minor classic of the "unhappy child" school, also continues to be read and is indeed a well-told story. But the century's real writer of genius is of course Jules Verne, whose first book, *Un Voyage en ballon,* was originally published in 1851 in a children's magazine, *Le Musée des Familles.*

The period was lively enough. Production was vast. Children's magazines flourished, particularly the remarkable *Magasin d'Éducation et de Récréation,* brilliantly edited by Jules Hetzel. Writers of the stature of George Sand, Alphonse Daudet, and Alexandre Dumas *père* were not too proud to write for children. Much worthy, though transient, work was produced along with a mass of mediocrity, as was the case also in England and the United States. But on the whole, as the century drew to a close, French children might have been better served, even though one critic sees the apogee as occurring between 1860 and 1900.

From the turn of the century to the close of World War II, a number of superior works were produced. The books of de Brunhoff and Faucher have already been cited. A remarkable picture of prehistoric life by J.-H. Rosny (pseudonym of J.-H.-H. Boex) appeared in 1911 and has proved so durable that in 1967 an English translation, *The Quest for Fire,* appeared. *Patapoufs et filifers,* by André Maurois, a gentle satire on war, has lasted (Eng. trans. *Pattypuffs and Thinifers,* 1948; reissued 1968). His fantastic *Le Pays des 36,000 volontés* is almost as popular. The famous dramatist Charles Vildrac has done much to advance the cause of French children's literature. Two pleasant stories of his, remotely decended from *Robinson Crusoe, L'Isle rose* and its sequel *La Colonie,* appeared in the 1920s and 1930s. In 1951 his now-classic comic animal tale *Les Lunettes du lion* won immediate success (Eng. trans., *The Lion's Eyeglasses,* 1969). On a high literary level, not accessible to all children, was *Le Petit Prince* (1943, both French and English, *The Little Prince*) by the famous aviator-author Antoine de Saint-Exupéry. The very vagueness of this mystical parable has lent it a certain magnetism. Finally, it is necessary to mention a field in which the French proved incomparable: the comic strip combining action and satire, conceived on a plane of considerable sophistication. Hergé's *Tintin* started in the 1930s and sold over 25,000,000 copies. Also successful was the later and even more unconventional *Astérix* series.

Early 20th-century French literature

Production after 1945 so multiplied that to single out names is bound to involve some injustice. A few, however, by reason either of the originality of their talent or the scope of their achievement, stand out. One is Maurice Druon, whose *Tistou of the Green Fingers* (1957; Eng. trans. 1958), a kind of children's *Candide,* demonstrated how the moral tale, given sufficient sensitivity and humour, can be transmuted into art. Perhaps the most original temperament was that of Henri Bosco, author of four eerie, haunting Provençal novels about the boy Pascalet and his strange involvements with a gypsy companion, a fox, and a dog in a shifting, legend-shrouded natural world. It may be that time will rate these books, like those of the English writer Walter de la Mare, among the finest of their kind. Bosco's *L'Enfant et la rivière* (1955; Eng. trans., *The Boy and the River,* 1956), *Le Renard dans l'île* (1956; Eng. trans., *The Fox in the Island,* 1958), and *Barboche* (1957; Eng. trans. 1959) are notable.

Sound, realistic novels, almost free of excess moralism, were written by at least a dozen reputable authors. Among them Colette Vivier (*The House of the Four Winds*), Paul-Jacques Bonzon (*The Orphans of Simitra*), and Étienne Cattin (*Night Express!*) were distinguished. The domain of the imaginative tale was well represented by Maurice Vauthier, especially by his *Ecoute, petit loup.* Among those noted for their prolific output as well as the high level of their art two names emerged. One is Paul Berna, who has worked in half a dozen genres, including detective stories and science fiction. His *Cheval sans tête* (1955) was pub-

lished in England as *A Hundred Million Francs,* and in the United States as *The Horse Without a Head,* and was made into a successful Disney film. A "gang" story, using a hard, unemotional tone that recalls Simenon, it may be the best of its kind since *Emil and the Detectives.*

The death of René Guillot removed a deeply conscientious and responsible artist. Guillot, though probably not of the first rank, was not far below it. He left over 50 widely translated novels for the young and about 10 nonfiction works. For his entire body of work he received in 1964 the Andersen Prize. His finest achievements in the adventure novel, based on his experiences in Africa, include *The White Shadow* (1948) and *Riders of the Wind* (1953).

Children's verse has at least one delightful practitioner in Pierre Gamarra. His *Mandarine et le Mandarin* contains Fontainesque fables of notable drollery and high technical skill. The Belgian author Maurice Carême also has some repute as a children's poet. In summary, contemporary French activity seems a bit lacking in colour and versatility. But one solid achievement must be registered: the 19th century's legacy was decisively rejected and at last a natural child prevailed in the imaginative work of the best French contemporaries.

Russia/Soviet Union. Here history breaks cleanly into two periods: pre-1917 and post-1917. In pre-Revolutionary Russia may be observed a most dramatic illustration of the disproportion that may exist between a children's and a mainstream literature. Beyond question the latter is one of the greatest of the modern world. But Russia's pre-1917 children's literature is anemic. It does include the fables of Ivan Krylov; a great treasury of Russian folktales (*skazki*) assembled by A.N. Afanasyev; the epic tales (*byliny*) sung or told to children; the classic by Pyotr Yrshov, *Konyok gorbunok* (1834; English adaption by Ireene Wicker, *The Little Hunchback Horse,* 1942); and other stories and poems enjoyed by young Russians but not originally designed for them. To this folk material should be added the McGuffeyish moral tales that Tolstoy wrote for his four graded readers. There is also the poet-translator Vasily Zhukovsky, praised by the respected critic Vissarion Belinsky as one of the few poets of the century, part of whose work was dedicated to children.

Pre-Revolutionary Russian literature

On the whole, however, pre-Revolutionary Russia could make only a few feeble gestures toward the creation of an independent children's literature. The submerged peasantry relied on the fireside tale teller. The middle class, while far stronger than is generally recognized, was in no position to stimulate or support a literature for its children. The privileged class looked to the West: the children read Mme de Genlis. Thus it came about that the child was recognized later in Russia than in other parts of western Europe. The critic and children's writer Korney Chukovsky speaks of the "indifference" with which "early childhood was regarded in the past." He then points out that attitudes have changed, so that now the child is "an adored hero."

The Revolution was the watershed. After 1917 Soviet children's literature developed more or less in accord with the necessities of the state. This is not to say that it became identical with Soviet propaganda. Indeed one of the finest teenage novels, Vadim Frolov's *Chto k chemu* (Eng. trans., *What It's All About,* 1965), is quite untouched by dogma of any kind. Soviet children's literature, and especially its vast body of popularized science and technology for the young, however, was in general governed by the ideals of socialist realism, the idolization of the "new Soviet man" (as in the widely read works of Boris Zhitkov and Arkady Gaydar), the exaltation of the machine over the irresponsible furniture of fairyland, and especially a revised version of the pre-18th-century miniature adult view of the child: he now becomes a potential Soviet citizen and architect of the Communist future.

Juvenile fiction and biography have naturally tended to cue themselves into the crucial episodes of Soviet history. But the theory underlying this basically nationalist literature (suggesting similar developments in Italy and England in the latter half of the 19th century) is by no means clear-cut. The most influential thinker was Maksim Gorky, who during the 1920s called for "creative fantasy," for children's stories "which make out of the human being, instead of a will-less creature or an indifferent workman, a free and active artist, creator of a new culture." He asked for books that would encourage the child to become "a knight of the spirit." Gorky's essays are a curious, endearing mixture of Marxist doctrine (with a utopian slant) and quite standard Western humanistic ideas. It is in Korney Chukovsky's remarkable book *Malenkiye deti* (1925) or *Ot dvukh do pyati* (Eng. trans., *From Two to Five,* 1963), however, that the opposition of two familiar forces, entertainment and instruction, can be sensed most clearly. The tension is typically expressed in Chukovsky's account of the Soviet war over the fairy tale, the opposition to which reached its high point in the 1920s and 1930s. "We propose," wrote one journalist in a Moscow magazine in 1924, "to replace the unrealistic folktales and fantasies with simple realistic stories taken from the world of reality and from nature." Chukovsky, himself a writer full of humour and invention, opposed this view, as had Gorky before him. The struggle still continues, in Soviet Russia as elsewhere.

Gorky's contributions

Though rich in folklore drawn from its many peoples and languages, the Soviet Union remains weak in the realm of fantasy. A fairy play such as Marshak's *Krugly god* (Eng. trans., *The Month Brothers,* 1967) seems (at least in English) fatally heavy handed. That children's literature in the Soviet Union is vigorous, varied, and motivated by a genuine concern for the child is undoubted. But it is hard to imagine a Soviet "Narnia" series, a Soviet *Borrowers.*

It is not difficult to see that contemporary children's literature in Russia is lively, copious, and probably enjoyed. It is much more difficult for those who have no Russian to judge its value. Occasionally in translation one will come across something as superb as the beautiful nature and animal tales in *Arcturus the Hunting Hound and Other Stories* (1968) by Yury Kazakov. But one can only record, without judging, the vast production of such popular children's writers as Samuil Marshak, Sergey Mikhalkov, Lev Kassil, and N. Nosov. Especially notable is the popularity of poetry, whether it be the work of such past generation writers as Vladimir Mayakovsky or that of the contemporary Agniya Barto. Apparently Russian children read poetry with more passion and understanding than do English-speaking children. The mind of the Soviet child is carefully looked after. He is provided with books, often beautifully illustrated, at prices the West cannot match. "Demand from them as much as possible, respect them as much as possible," says Anton Makarenko, the theorist of children's literature. (C.Fa.)

BIBLIOGRAPHY

The scope of literature. *General works:* KENNETH BURKE, *The Philosophy of Literary Form,* 2nd ed. (1967); I.A. RICHARDS, *Practical Criticism: A Study of Literary Judgment* (1929, reprinted 1968) and *Principles of Literary Criticism* (1924, reprinted 1961); GEORGE SAINTSBURY, *A History of Criticism and Literary Taste in Europe from the Earliest Texts to the Present Day,* 3 vol. (1900–04, reprinted 1961); NOWELL C. SMITH (ed.), *Literary Criticism* (1905); KONSTANTIN KOLENDA, *Philosophy in Literature* (1982).

Ancient to modern: A Translation of the Latin Works of Dante Alighieri (1904), see Letter X to Can Grande; CHARLES SEARS BALDWIN, *Ancient Rhetoric and Poetic, Interpreted from Representative Works* (1924), *Medieval Rhetoric and Poetic to 1400, Interpreted from Representative Works* (1928, reprinted 1971), and *Renaissance Literary Theory and Practice* (1939); EDWARD H. BLAKENEY (ed.), *Horace on the Art of Poetry* (1928); CECIL MAURICE BOWRA, *Primitive Song* (1962); S.H. BUTCHER, *Aristotle's Theory of Poetry and Fine Art,* with a critical text and translation of the *Poetics,* 4th ed. (1907; reprinted with corrections, 1932); INGRAM BYWATER, *Aristotle on the Art of Poetry* (1909), reprinted in Aristotle's *Poetics* and *Longinus on the Sublime,* ed. by CHARLES SEARS BALDWIN (1930); PIERRE CORNEILLE, *Oeuvres,* 3 vol. (1862), contains the "Discourse on Dramatic Poetry" and "Discourse on the Three Unities"; ALBERT S. COOK (ed.), *The Art of Poetry* (1892, reprinted 1926), contains a translation of Horace's *Art of Poetry;* J.D. DENNISTON, *Greek Literary Criticism* (1924, reprinted 1971), translations, beginning with Aristophanes; *Dryden's Essays on the Drama,* ed. by WILLIAM STRUNK (1898); ALLAN H. GILBERT (ed.), *Literary Criticism: Plato to Dryden* (1940); EDMUND D.

JONES (ed.), *English Critical Essays (Sixteenth, Seventeenth, and Eighteenth Centuries)* (1922); *Ben Jonson: Timber, Discoveries, and Conversations with Drummond of Hawthornden* in his *Works,* ed. by C.H. HERFORD and PERCY SIMPSON, 11 vol. (1925–52); LONGINUS, *On the Sublime,* Greek text with an English translation by W. HAMILTON FYFE ("Loeb Classical Library," 1927); *Plato,* trans. by LANE COOPER (1938), contains the *Phaedrus,* the *Symposium,* the *Ion,* the *Gorgias,* and parts of the *Republic* and the *Laws;* GEORGE PUTTENHAM, *The Arte of English Poesie,* ed. by GLADYS D. WILLCOCK and ALICE WALKER (1936); PAUL RADIN, *Primitive Man as Philosopher* (1927); *Sidney's Apologie for Poetrie,* ed. by J. CHURTON COLLINS (1907); G. GREGORY SMITH (ed.), *Elizabethan Critical Essays,* 2 vol. (1904); GAY WILSON ALLEN and H.H. CLARK (eds.), *Literary Criticism: Pope to Croce* (1941 and 1962); MATTHEW ARNOLD, *Essays in Criticism,* 2 vol., First and Second Series complete (1902), and *Essays in Criticism,* with an introduction by E.J. O'BRIEN, Third Series (1910); EDWIN BERRY BURGUM (ed.), *The New Criticism: An Anthology of Modern Aesthetics and Literary Criticism* (1930). SAMUEL TAYLOR COLERIDGE, *Biographia Literaria, or Biographical Sketches of My Literary Life and Opinions,* 2 vol., reprinted from the original plates (1907); BENEDETTO CROCE, *The Defence of Poetry, Variations on the Theme of Shelley,* trans. by E.F. CARRITT (1933); T.S. ELIOT, *Selected Essays, 1917–1932* (1932); *Hazlitt on English Literature: An Introduction to the Appreciation of Literature,* ed. by JACOB ZEITLIN (1913, reprinted 1970); E.R. HUGHES (trans.), *The Art of Letters: Lu Chi's "Wen Fu," A.D. 302* "Bollingen Series XXIX" (1951); THOMAS ERNEST HULME, *Speculations: Essays on Humanism and the Philosophy of Art,* ed. by HERBERT READ (1924); JAMES GIBBONS HUNEKER, *Essays* (1929); EDMUND D. JONES (ed.), *English Critical Essays of the Nineteenth Century* (1922); PHYLLIS M. JONES (ed.), *English Critical Essays: Twentieth Century* (1933); WILLIAM PATON KER, *Collected Essays,* 2 vol. (1925, reprinted 1968); KARL MARX and FRIEDRICH ENGELS, *Sur la littérature et l'art,* ed. and trans. by JEAN FREVILLE (1936); H.L. MENCKEN, *A Mencken Chrestomathy* (1949); PAUL ELMER MORE, *The Demon of the Absolute* (1928) and *Shelburne Essays,* 11 series (1904–21); FRIEDRICH NIETZSCHE, *Ecce Homo* and *The Birth of Tragedy,* trans. by CLIFTON FADIMAN (1927); HORATIO, *Works,* 10 vol. (1910); GEORGY V. PLEKHANOV, *Art and Society,* trans. by ALFRED GOLDSTEIN (1936), a Marxist analysis; EDGAR ALLAN POE, *Selections from the Critical Writings of Edgar Allan Poe,* ed. with an introduction by F.C. PRESCOTT (1909); EZRA POUND, *ABC of Reading* (1934) and *Literary Essays* (1954); HERBERT READ, *Reason and Romanticism* (1926); CHARLES AUGUSTIN SAINTE-BEUVE, *Causeries du lundi,* 15 vol. (1852–62; Eng. trans., 8 vol., 1909–11), contains "What Is a Classic?". *Shelley's Literary and Philosophical Criticism,* ed. by JOHN SHAWCROSS (1909); LEO TOLSTOY, *What Is Art?,* trans. by AYLMAR MAUDE (1932); LEON TROTSKY, *Literature and Revolution,* trans. by ROSE STRUNSKY (1925, reprinted 1957); PAUL VALERY, *Littérature* (1929) and *Variété,* trans. by MALCOLM COWLEY (1927); WILLIAM CARLOS WILLIAMS, *In the American Grain* (1925); EDMUND WILSON, *Axel's Castle* (1931) and *The Triple Thinkers,* rev. ed. (1952 and 1963); ÉMILE ZOLA, *The Experimental Novel and Other Essays,* trans. by BELLE M. SHERMAN (1893).

Contemporary: CECIL MAURICE BOWRA, *In General and Particular* (1964); STANLEY BURNSHAW (ed.), *The Poem Itself,* rev. ed. (1967); CYRIL CONNOLLY, *The Modern Movement* (1965); PAUL GOODMAN, *The Structure of Literature* (1954); MARSHALL MCLUHAN, *Understanding Media* (1964); R.E. SCHOLES and R.L. KELLOGG, *The Nature of Narrative* (1966); HERBERT READ, *The Nature of Literature* (1956).

The nature of poetry. The most convenient way to get to know poetry is to read poetry. It would be invidious for the writer of a general article on the subject to prejudice the reader by making a selection of poems or poets; in experience, anyhow, one's acquaintance with poetry comes about chiefly by love and accident, supported, when not undermined, by schools, colleges, and libraries. Beyond that, the bibliographical temptation is to put before the reader numerous learned works that are not poetry but about poetry; whatever their usefulness at various stages of study, and it may be great, they must not substitute for the reading of poetry itself. Therefore no such list is attempted.

The beginning reader, however, may well be able to use some help in interpreting, such as a critical or explanatory anthology. CLEANTH BROOKS and ROBERT PENN WARREN, *Understanding Poetry,* 4th ed. (1976), is still probably the best of its kind, as numerous imitations amply attest. See also TZVETAN TODOROV, *Introduction to Poetics* (1981; originally published in French, 1973), a comprehensive introduction to modern poetics.

Prosody. *Greek and Latin prosody:* PAUL MAAS, *Greek Metre,* trans. by HUGH LLOYD-JONES (1962); ULRICH VON WILAMOWITZ-MOLLENDORFF, *Griechische Verskunst* (1921), the definitive work on the subject but difficult for beginners.

Prose rhythm: MORRIS W. CROLL, *Style, Rhetoric, and Rhythm* (1966), contains classic essays on the period styles of prose and on musical scansion of verse; GEORGE SAINTSBURY, *A History of English Prose Rhythm* (1912, reprinted 1965).

History and uses of English prosody: WILLIAM BEARE, *Latin Verse and European Song: A Study in Accent and Rhythm* (1957); ROBERT BRIDGES, *Milton's Prosody,* rev. ed. (1921); HARVEY S. GROSS, *Sound and Form in Modern Poetry: A Study of Prosody from Thomas Hardy to Robert Lowell* (1964); T.S. OMOND, *English Metrists* (1921, reprinted 1968); GEORGE SAINTSBURY, *A History of English Prosody from the Twelfth Century to the Present,* 3 vol. (1906–10); JOHN THOMPSON, *The Founding of English Metre* (1961).

Theories of prosody: SEYMOUR B. CHATMAN, *A Theory of Meter* (1965); OTTO JESPERSEN, "Notes on Metre," in *The Selected Writings of Otto Jespersen* (1962); WILLIAM K. WIMSATT, JR., and MONROE C. BEARDSLEY, "The Concept of Metre: An Exercise in Abstraction," in WILLIAM K. WIMSATT, JR., *Hateful Contraries* (1965); YVOR WINTERS, "The Audible Reading of Poetry," in *The Function of Criticism* (1957).

Non-Western prosody: ROBERT H. BROWER and EARL MINER, *Japanese Court Poetry* (1961); JAMES LEGGE (ed. and trans.), *The Chinese Classics,* vol. 4 (1960).

General works: PAUL FUSSELL, JR., *Poetic Meter and Poetic Form* (1965); HARVEY S. GROSS (ed.), *The Structure of Verse: Modern Essays on Prosody* (1966); JOSEPH MALOF, *A Manual of English Meters* (1970).

Epic. H.M. CHADWICK, *The Heroic Age* (1912), and H.M. and N.K. CHADWICK, *The Growth of Literature,* vol. 1, *The Ancient Literatures of Europe* (1932), are two classic works on European heroic poetries that are still valuable. A more comprehensive and up-to-date general survey is given in C.M. BOWRA, *Heroic Poetry* (1952); and J. DE VRIES, *Heldenlied und Heldensage* (1961; Eng. trans., *Heroic Song and Heroic Legend,* 1963). A.B. LORD, *The Singer of Tales* (1960), was written by an authority on the Yugoslav oral epic of the *guslari.* On Homer, see G.S. KIRK, *The Songs of Homer* (1962); W. SCHADEWALDT, *Von Homers Welt und Werk,* 4th ed. (1965); C.M. BOWRA, *Homer and His Forerunners* (1955); and R. CARPENTER, *Folk Tale, Fiction and Saga in the Homeric Epics* (1946). F.R. SCHRODER, *Germanische Heldendichtung* (1935), is a basic reference book for the study of the Germanic epic. J.B. PRITCHARD, *The Ancient Near East* (1958), gives summaries and translations of Akkadian and Ugaritic epics. A theory of the epics of the Indo-Europeans is presented by G. DUMEZIL in *Mythe et épopée,* vol. 1–2 (1968–71). For the use of mythical themes in the epics of the Indo-Europeans, see also D. WARD, *The Divine Twins: An Indo-European Myth in Germanic Tradition* (1968). MICHAEL MURRIN, *The Allegorical Epic* (1980), is a survey of the European tradition.

Fable, parable, and allegory. *Fable:* E. CHAMBRY, *Fables* (1927), in Greek and French; S.A. HANDFORD, *Fables of Aesop* (1956); B. PARES, *Krylov's Fables* (1926); MARIANNE MOORE, *Fables of La Fontaine* (1954). For commentary on fables, see: P. CLARAC, *La Fontaine, l'homme et l'oeuvre* (1947); B.E. PERRY, *Aesopica* (1952).

Parable: A.M. HUNTER, *The Parables: Then and Now* (1971); ETA LINNEMANN, *Gleichnisse Jesu,* 3rd ed. (1964; Eng. trans., *The Parables of Jesus,* 1966); T.W. MANSON (ed.), *The Sayings of Jesus as Recorded in the Gospels According to St. Matthew and St. Luke* (1949); D.C. ALLEN, *The Legend of Noah: Renaissance Rationalism in Art, Science and Letters* (1963); HEINZ POLITZER, *Franz Kafka: Parable and Paradox* (1962).

Allegory: (*General theory and history*): D.C. ALLEN, *Mysteriously Meant: The Rediscovery of Pagan Symbolism and Allegorical Interpretation in the Renaissance* (1970); C.H. DODD, *The Authority of the Bible* (1958); A.S. FLETCHER, *Allegory: The Theory of a Symbolic Mode* (1964); R.M. GRANT, *The Letter and the Spirit* (1958); EDWIN HONIG, *Dark Conceit: The Making of Allegory* (1959); C.S. LEWIS, *The Allegory of Love* (1936); JEAN PEPIN, *Mythe et allégorie* (1958); ROSEMOND TUVE, *Allegorical Imagery* (1966); MAUREEN QUILLIGAN, *The Language of Allegory: Defining the Genre* (1979).

(*Pagan and Christian interpretation*): KENNETH BURKE, *The Rhetoric of Religion* (1961); HENRY CHADWICK, *Early Christian Thought and the Classical Tradition* (1966); C.H. DODD, *The Interpretation of the Fourth Gospel* (1968); A.O. LOVEJOY, *The Great Chain of Being* (1936); H. DE LUBAC, *Exégèse médiévale: les quatre sens de l'Écriture* (1959–64); A. MOMIGLIANO (ed.), *The Conflict Between Paganism and Christianity in the Fourth Century* (1963); G. VON RAD, *Theologie des Alten Testaments,* 2nd ed. (1958; Eng. trans., *Old Testament Theology,* 2 vol., 1962–65); RENE ROQUES, *L'Univers dionysien* (1954); B. SMALLEY, *The Study of the Bible in the Middle Ages,* 2nd ed. (1952); H.A. WOLFSON, *The Philosophy of the Church Fathers,* vol. 1, *Faith, Trinity, Incarnation* (1956); *Philo: Foundations*

of Religious Philosophy in Judaism, Christianity, and Islam, 2 vol. (1947).

(*Typology and typological symbolism*): ERICH AUERBACH, "Figura," in *Scenes from the Drama of European Literature: Six Essays* (1959); A.C. CHARITY, *Events and Their Afterlife: The Dialectics of Christian Typology in the Bible and Dante* (1966); JEAN DANIELOU, *Sacramentum futuri: études sur les origines de la typologie biblique* (1950; Eng. trans., *From Shadows to Reality: Studies in the Biblical Typology of the Fathers,* 1960); AUSTIN FARRER, *A Rebirth of Images: The Making of St. John's Apocalypse* (1949); R.P.C. HANSON, *Allegory and Event* (1959); W.G. MADSEN, *From Shadowy Types to Truth: Studies in Milton's Symbolism* (1968).

(*Medieval allegory*): ERICH AUERBACH, *Dante als Dichter der irdischen Welt* (1929; Eng. trans., *Dante: Poet of the Secular World,* 1961); M.W. BLOOMFIELD, "Symbolism in Medieval Literature," *Modern Philology,* 56:73–81 (1958), and *Piers Plowman As a Fourteenth-Century Apocalypse* (1962); EDGAR DE BRUYNE, *Études d'esthétique médiévale,* 3 vol. (1946); M.D. CHENU, *La Théologie au douzième siècle* (1957; Eng. trans. of nine selected essays, *Nature, Man, and Society in the Twelfth Century,* 1968); E.R. CURTIUS, *Europäische Literatur und lateinisches Mittelalter* (1948; Eng. trans., *European Literature and the Latin Middle Ages,* 1953); RAYMOND KLIBANSKY, *The Continuity of the Platonic Tradition During the Middle Ages* (1939); C.S. LEWIS, *The Discarded Image: An Introduction to Medieval and Renaissance Literature* (1964); JOSEPH A. MAZZEO, *Medieval Cultural Tradition in Dante's Comedy* (1960); D.W. ROBERTSON and B.F. HUPPE, *Piers Plowman and Scriptural Tradition* (1951); CHARLES SINGLETON, *Dante Studies,* vol. 1, *Commedia* (1954).

(*Renaissance and modern allegory*): DOUGLAS BUSH, *Mythology and the Renaissance Tradition in English Poetry,* rev. ed. (1963); WALTER BENJAMIN, *Ursprung des deutschen Trauerspiels* (1928); HAROLD BLOOM, *The Visionary Company* (1961); A.S. FLETCHER, *The Prophetic Moment: An Essay on Spenser* (1971); ALASTAIR FOWLER, *Triumphal Forms: Structural Patterns in Elizabethan Poetry* (1970); NORTHROP FRYE, *Fearful Symmetry: A Study of William Blake* (1947); U.M. KAUFMAN, *The Pilgrim's Progress and Traditions in Puritan Meditation* (1966); MICHAEL MURRIN, *The Veil of Allegory: Some Notes Toward a Theory of Allegorical Rhetoric in the English Renaissance* (1969); JEAN SEZNEC, *La Survivance des dieux antiques* (1939; Eng. trans., *The Survival of the Pagan Gods,* rev. ed., 1953); E.M.W. TILLYARD, *The Elizabethan World Picture* (1943); EDGAR WIND, *Pagan Mysteries in the Renaissance,* new ed. (1968).

Ballad. F.J. CHILD (ed.), *The English and Scottish Popular Ballads,* 5 vol. (1882–98), is the canon of traditional balladry; the tunes for which are supplied in B.H. BRONSON (ed.), *Traditional Tunes of the Child Ballads,* 4 vol. (1959–72). Important broadside collections include *The Roxburghe Ballads,* ed. by W. CHAPPELL and J.W. EBSWORTH, 9 vol. (1871–99); and *The Pepys Ballads,* ed. by H.E. ROLLINS, 8 vol. (1929–32). See also *The Common Muse: An Anthology of Popular British Ballad Poetry, XVth–XXth Century,* ed. by V. DE SOLA PINTO and A.E. RODWAY (1957); C.M. SIMPSON, *The British Broadside Ballad and Its Music* (1966); T.P. COFFIN, *The British Traditional Ballad in North America,* rev. ed. (1963); and G.M. LAWS, *Native American Balladry,* rev. ed. (1964). Ballad criticism and scholarship are analyzed in S.B. HUSTVEDT, *Ballad Books and Ballad Men* (1930); D.K. WILGUS, *Anglo-American Folksong Scholarship Since 1898* (1959); A.B. FRIEDMAN, *The Ballad Revival: Studies in the Influence of Popular on Sophisticated Poetry* (1961); C.J. SHARP, *English Folk-Song: Some Conclusions* (1907); G.H. GEROULD, *Ballad of Tradition* (1932); and M.J.C. HODGART, *Ballads* (1950). A.T. QUILLERCOUCH (ed.), *The Oxford Book of Ballads* (1910, reissued 1951); M. LEACH (ed.), *The Ballad Book* (1955); and A.B. FRIEDMAN (ed.), *Folk Ballads of the English Speaking World* (1956), are the standard anthologies.

Romance. Among older works the most notable are RICHARD HURD, *Letters on Chivalry and Romance* (1764); GEORGE ELLIS, *Specimens of Early English Metrical Romances,* 3 vol. (1805); and SIR WALTER SCOTT, "Essay on Romance" in the Supplement to the 1815–24 edition of the *Encyclopædia Britannica.* The academic study of romance as a form of imaginative narrative may be said to have begun in 1897 with the publication of W.P. KER, *Epic and Romance* (2nd ed. 1908, reprinted 1957), and of GEORGE SAINTSBURY, *The Flourishing of Romance and the Rise of Allegory.* (*Origins and sources*): EDMOND FARAL, *Recherches sur les sources latines des contes et romans courtois du moyen âge* (1913); JESSIE L. WESTON, *From Ritual to Romance* (1920, reprinted 1957); ROGER S. LOOMIS, *Arthurian Tradition and Chrétien de Troyes* (1949); and JEAN MARX, *La Légende arthurienne et le Graal* (1952). (*Nature and development*): FANNI BOGDANOW, *The Romance of the Grail* (1966); EUGENE VINAVER, *The Rise of Romance* (1971); and ROSEMOND TUVE, *Allegorical Imagery* (1966). J.D. BRUCE, *The Evolution of Arthurian Romance,* 2nd ed., 2 vol. (1928), at one time the standard work in this field, has now been largely superseded by R.S. LOOMIS (ed.), *Arthurian Literature in the Middle Ages: A Collaborative History* (1959). Since 1949 the International Arthurian Society has been publishing an annual *Bibliographical Bulletin* covering the whole range of Arthurian literature in all languages.

Saga. The best guide to current research is the annual *Bibliography of Old Norse-Icelandic Studies* (from 1964); for earlier works on the sagas, see *Islandica* (from 1908). Standard editions of important texts include *Íslenzk fornrit* (from 1933); *Altnordische Saga-Bibliothek,* ed. by G. CEDERSCHIOLD *et al.,* 18 vol. (1892–1929); *Editiones Arnamagnaeanae* (from 1958); *Fornaldar Sögur Nordurlanda,* 4 vol. (1950); *Sturlunga saga,* 2 vol. (1946); and *Nelson's Icelandic Texts* (from 1957), with English translations. Useful general surveys of the sagas are PETER HALLBERG, *Den Isländska Sagan* (1956; Eng. trans., *The Icelandic Saga,* 1962); S. NORDAL, *Sagalitteraturen* (1953); and KURT SCHIER, *Sagaliteratur* (1969). For criticism and interpretation of the saga, see WALTER BAETKE, *Über die Entstehung der Isländersagas* (1956); THEODORE M. ANDERSSON, *The Problem of Icelandic Saga Origins* (1964); GABRIEL TURVILLE-PETRE, *Origins of Icelandic Literature* (1953); THEODORE M. ANDERSSON, *The Icelandic Family Saga: An Analytic Reading* (1967); HERMANN PALSSON, *Art and Ethics in Hrafnkel's Saga* (1971); EINAR O. SVEINSSON, *Á Njálsbúd, bok um mikid listaverk* (1943; Eng. trans., *Njáls Saga: A Literary Masterpiece,* 1971); GABRIEL TURVILLE-PETRE, *The Heroic Age of Scandinavia* (1951) and *Myth and Religion of the North* (1964); and HERMANN PALSSON and PAUL EDWARDS, *Legendary Fiction in Medieval Iceland* (1970). See also MARGARET SCHLAUCH, *Romance in Iceland* (1934); and E.F. HALVORSEN, *The Norse Version of the Chanson de Roland* (1959).

Of translations into English, the following may be mentioned: L.M. HOLLANDER (ed. and trans.), *Heimskringla: History of the Kings of Norway* (1964) and *The Sagas of Kormák and the Sworn Brothers* (1949); GWYN JONES (ed. and trans.), *The Vatnsdalers' Saga* (1944), *Egil's Saga* (1960), and *Eirik the Red, and Other Icelandic Sagas* (1961); GEORGE JOHNSTON (ed. and trans.), *The Saga of Gisli* (1963); HERMANN PALSSON (ed. and trans.), *Hrafnkel's Saga and Other Icelandic Stories* (1971); PAUL EDWARDS and HERMANN PALSSON (eds. and trans.), *Arrow-Odd: A Medieval Novel* (1970), *Gautrek's Saga, and Other Medieval Tales* (1968), *Hrolf Gautreksson: A Viking Romance* (1972), and *Eyrbyggja Saga* (1973); MAGNUS MAGNUSSON and HERMANN PALSSON (eds. and trans.), *Njal's Saga* (1960), *The Vinland Sagas* (1965), *King Harald's Saga* (1966), and *Laxdaela Saga* (1969); M.H. SCARGILL and MARGARET SCHLAUCH (eds. and trans.), *Three Icelandic Sagas* (1950); J.I. YOUNG (ed. and trans.), *The Prose Edda of Snorri Sturluson: Tales from Norse Mythology* (1954); J.H. MCGREW (ed. and trans.), *Sturlunga Saga,* vol. 1 (1970); DENTON FOX and HERMANN PALSSON (eds. and trans.), *Grettir's Saga* (1973).

Novel. The following works deal in general terms with the reader's approach to the novel: WALTER ALLEN, *Reading a Novel,* rev. ed. (1963); VAN METER AMES, *Aesthetics of the Novel* (1928, reprinted 1966); CLEANTH BROOKS and R.P. WARREN (eds.), *Understanding Fiction,* 3rd. ed. (1979); ALEXANDER COMFORT, *The Novel and Our Time* (1948); PELHAM EDGAR, *The Art of the Novel* (1933, reprinted 1966); WILSON FOLLETT, *The Modern Novel: A Study of the Purpose and Meaning of Fiction,* rev. ed. (1923); E.M. FORSTER, *Aspects of the Novel* (1927, many reprintings); PERCY LUBBOCK, *The Craft of Fiction,* new ed. (1957).

The following are concerned with the problems of writing fiction and are all the work of novelists: PHYLLIS BENTLEY, *Some Observations on the Art of Narrative* (1946); *Conrad's Prefaces to His Works,* with an essay by EDWARD GARNETT (1937); HENRY JAMES, *The Art of Fiction and Other Essays,* ed. by MORRIS ROBERTS (1948), and *The Art of the Novel,* introduction by R.P. BLACKMUR (1934); EDITH WHARTON, *The Writing of Fiction* (1925); THOMAS WOLFE, *The Story of a Novel* (1936).

The various elements of the novel are dealt with in the following: BONAMY DOBREE, *Modern Prose Style,* 2nd ed. (1964); MAREN ELWOOD, *Characters Make Your Story* (1942); MANUEL KOMROFF, *How to Write a Novel* (1950); W. VAN O'CONNOR (ed.), *Forms of Modern Fiction* (1948); GEORGE G. WILLIAMS (ed.), *Readings for Creative Writers* (1938).

The following studies deal with the style and philosophy of the novel in the wider sense: DAVID DAICHES, *The Novel and the Modern World,* rev. ed. (1960); AGNES HANSEN, *Twentieth Century Forces in European Fiction* (1934); ALFRED KAZIN, *On Native Grounds* (1942); Y. KRIKORIAN (ed.), *Naturalism and the Human Spirit* (1944); GEORGE LUKACS, *Studies in European Realism* (1950), and *The Historical Novel* (1962); H.J. MULLER, *Modern Fiction* (1937); and MAS'UD ZAVARZADEH, *The Mythopoeic Reality: The Postwar American Nonfiction Novel* (1976).

Short story. K.P. KEMPTON examines the genre, emphasizing

theme and meaning, in *The Short Story* (1947). An excellent analysis of story techniques is offered in SEAN O'FAOLAIN, *The Short Story* (1948). BRANDER MATTHEWS, *The Philosophy of the Short Story* (1901, reprinted 1931), is predicated on Poe's theories. A provocative thesis regarding the nature of stories is presented in FRANK O'CONNOR, *The Lonely Voice: A Study of the Short Story* (1963). H.S. SUMMERS has collected some of the more important discussions of the form in *Discussions of the Short Story* (1963). *Storytellers and Their Art,* ed. by G. SAMPSON and C. BURKHART (1963), contains comments of various authors on the form. More specialized discussions of the form are contained in F.L. PATTEE, *The Development of the American Short Story* (1923); R.B. WEST, *Short Story in America, 1900–1950* (1952); E.K. BENNETT, *A History of the German Novelle,* 2nd ed. rev. by H.M. WAIDSON (1961); and S. TRENKNER, *Greek Novella in the Classical Period* (1958).

Dramatic literature. ALLARDYCE NICOLL, *World Drama* (1949), offers the best survey of the whole field, but should be supplemented by JOHN GASSNER and EDWARD QUINN, *The Reader's Encyclopaedia of World Drama* (1969); and PHYLLIS HARTNOLL, *The Oxford Companion to the Theatre,* 3rd ed. (1967). The classical texts of dramatic theory and criticism may be found in a collection by B.H. CLARK, *European Theories of the Drama,* rev. ed. (1965), which also contains an extensive bibliography. The *Natya Shastra* of BHARATA, the classic source for Indian dramatic theory, was translated by M.M. GHOSE in 1951.

Books of importance in the development of modern theory on drama are BERNARD BECKERMAN, *Dynamics of Drama* (1970); E.R. BENTLEY, *The Life of the Drama* (1964); KENNETH BURKE, *A Grammar of Motives* (1945); FRANCIS FERGUSSON, *The Idea of a Theatre* (1949); S.K. LANGER, *Feeling and Form* (1953); ELDER OLSON, *Tragedy and the Theory of Drama* (1961); RONALD PEACOCK, *The Art of Drama* (1957); J.L. STYAN, *The Elements of Drama* (1960); and KEIR ELAM, *The Semiotics of Theatre and Drama* (1980), a technical semiotic approach.

The finest study of the classical drama of Greece is probably H.D.F. KITTO, *Greek Tragedy,* 3rd ed. (1961); and for the medieval drama are recommended KARL YOUNG, *The Drama of the Medieval Church,* 2 vol. (1933); HARDIN CRAIG, *English Religious Drama of the Middle Ages* (1955); and O.B. HARDISON, *Christian Rite and Christian Drama in the Middle Ages* (1965). Oriental theatre is surveyed in FAUBION BOWERS, *Japanese Theatre* (1952); F.A. LOMBARD, *An Outline History of the Japanese Drama* (1928), which should be read in conjunction with ARTHUR WALEY's classic *The Noh Plays of Japan* (1922); A.C. SCOTT, *The Classical Theatre of China* (1957); A.B. KEITH, *The Sanskrit Drama* (1924); with H.W. WELLS's comparative studies, *The Classical Drama of India* (1963), and *The Classical Drama of the Orient* (1965).

M.C. BRADBROOK, *Themes and Conventions of Elizabethan Tragedy* (1935); and U.M. ELLIS-FERMOR, *Jacobean Drama* (1936), are standard surveys of the English Renaissance drama; and for standard Shakespearean criticism the reader should consult A.M. EASTMAN, *A Short History of Shakespearean Criticism* (1968). The classic source books for the commedia dell'arte are P.L. DUCHARTRE, *La Comédie italienne* (Eng. trans., *The Italian Comedy,* 1929, reprinted 1966); and ALLARDYCE NICOLL, *Masks, Mimes and Miracles* (1931). On the French classical drama H.C. LANCASTER, *A History of French Dramatic Literature in the Seventeenth Century,* 9 vol. (1929–42), is standard; but MARTIN TURNELL, *The Classical Moment* (1947), deals more briefly with Corneille, Racine, and Molière. On Restoration comedy J.L. PALMER, *The Comedy of Manners* (1913); and BONAMY DOBREE, *Restoration Comedy* (1924), remain the best.

American drama is surveyed briefly in W.J. MESERVE, *An Outline History of American Drama* (1965); and A.S. DOWNER, *Fifty Years of American Drama, 1900–1950* (1951). U.M. ELLIS-FERMOR, *The Irish Dramatic Movement,* 2nd ed. (1954), is a comprehensive study of the early years at Dublin's Abbey Theatre; and on Western drama after Ibsen the reader should begin by consulting ERIC BENTLEY, *The Playwright as Thinker* (1946, reprinted 1955); ROBERT BRUSTEIN, *The Theatre of Revolt* (1964); and J.L. STYAN, *The Dark Comedy,* 2nd ed. (1968), an account of the blending of tragic and comic elements in the post-Ibsen theatre.

Comedy. C.L. BARBER, *Shakespeare's Festive Comedy: A Study of Dramatic Form and Its Relation to Social Custom* (1959), Shakespearean comedy considered in relation to archetypal patterns of folk ritual and games; LANE COOPER, *An Aristotelian Theory of Comedy, with an Adaptation of the Poetics and a Translation of the Tractatus Coislinianus* (1922), the only modern text of the *Tractatus,* with an introductory essay relating it to Aristotle's theory of tragedy, and a conjectural reconstruction of the lost treatise on comedy based on the example of the *Poetics;* F.M. CORNFORD, *The Origin of Attic Comedy* (1914; ed. by T.H. GASTER, 1961), an account of the development of Greek comedy from primitive fertility rites, and of the survival of traces of these ceremonials in the extant plays of Aristophanes; CYRUS HOY, *The Hyacinth Room: An Investigation into the Nature of Comedy, Tragedy, and Tragicomedy* (1964), an examination of the plays of Euripides, Shakespeare, Jonson, Molière, Ibsen, Strindberg, Pirandello, Beckett, and Ionesco; J.W. KRUTCH, *Comedy and Conscience After the Restoration* (1924, reprinted with a new preface and additional bibliographic material, 1949), a study of the decline of Restoration comedy and the rise of sentimental comedy at the end of the 17th and the beginning of the 18th century; K.M. LYNCH, *The Social Mode of Restoration Comedy* (1926), the best available account of the relation of the plays of Dryden, Etherege, Wycherley, Congreve, and their contemporaries to their social milieu; A.W. PICKARD-CAMBRIDGE, *Dithyramb, Tragedy, and Comedy* (1927; 2nd ed. rev. by T.B.L. WEBSTER, 1962), and *The Dramatic Festivals of Athens* (1953; 2nd ed. rev. by J. GOULD and D.M. LEWIS, 1968), the definitive accounts of the origins of Greek comedy and tragedy; and F.H. RISTINE, *English Tragicomedy: Its Origin and History* (1910), the only full-scale account of the subject through the 17th century.

Tragedy. A lengthier development of many of the points made in this section may be found in R.B. SEWALL, *The Vision of Tragedy* (1959). J. JONES, *On Aristotle and Greek Tragedy* (1962), examines the origins of Greek tragedy. Works concentrating on modern tragedy include GEORGE STEINER, *The Death of Tragedy* (1961); WALTER KERR, *Tragedy and Comedy* (1968); and RAYMOND WILLIAMS, *Modern Tragedy* (1966). Special aspects of tragedy are treated in J.M.R. MARGESON, *The Origins of English Tragedy* (1967); EUGENE VINAVER, *Racine and Poetic Tragedy* (1955); and A.C. BRADLEY, *Shakespearean Tragedy* (1904). A useful anthology of writings on tragedy is LIONEL ABEL (ed.), *Moderns on Tragedy* (1967). Other recent works on the subject include RICHMOND HATHORN, *Tragedy, Myth, and Mystery* (1962); MURRAY KRIEGER, *The Tragic Vision: Variations on a Theme in Literary Interpretation* (1960); DOROTHY KROOK, *Elements of Tragedy* (1969); and TIMOTHY J. REISS, *Tragedy and Truth* (1980).

Satire. DAVID WORCESTER, *The Art of Satire* (1940), a study of rhetorical techniques available to the satirist; JAMES R. SUTHERLAND, *English Satire* (1958), a sound scholarly history; ALVIN B. KERNAN, *The Cankered Muse: Satire of the English Renaissance* (1959), valuable theory and criticism; ROBERT C. ELLIOTT, *The Power of Satire: Magic, Ritual, Art* (1960), on the origins of satire in magic and its development into an art; RONALD PAULSON, *The Fictions of Satire* (1967), a study of satire in fiction from Lucian to Swift, and (ed.), *Satire: Modern Essays in Criticism* (1971), an authoritative and indispensable collection; MATTHEW J.C. HODGART, *Satire* (1969), a well-illustrated, readable survey of satire in many forms and in many countries.

Other genres. On the essay, see R.D. O'LEARY, *The Essay* (1928), which analyzes the essay theoretically and examines several categories of essayists. DAVID DAICHES gives a brief and lively introduction to his anthologies of essays: *A Century of the Essay: British and American* (1951), and *More Literary Essays* (1968). On letter writing in general, the best piece is by GUSTAVE LANSON, in French: Introduction to *Choix de lettres du XVII^e siècle,* 5th rev. ed. (1898), reprinted in Lanson's *Essais de méthode, de critique et d'histoire littéraire,* pp. 243–258 (1965). On personal literature, the diary, autobiography, and the questions raised by those forms of prose, see HENRI PEYRE, *Literature and Sincerity* (1963).

Biography. *Critical and scholarly books:* JAMES L. CLIFFORD, *From Puzzles to Portraits: Problems of a Literary Biographer* (1970), examples of the author's own research followed by an analysis of biographical problems; LEON EDEL, *Literary Biography* (1959), essentially an account of the methods, psychological and narrative, used by the author in his multivolume life of Henry James; JOHN A. GARRATY, *The Nature of Biography* (1957), a historical survey coupled with a study of biographical methods, with emphasis on aids offered by psychology; PAUL M. KENDALL, *The Art of Biography* (1965), a historical survey, with emphasis on contemporary biography, and a study of biographical problems from the viewpoint of a practicing biographer; ANDRE MAUROIS, *Aspects de la biographie* (1928; Eng. trans. 1930) and HAROLD NICOLSON, *The Development of English Biography* (1928), particularly interesting for complementary views of the "new" biography of the 1920s by two eminent biographers; ROY PASCAL, *Design and Truth in Autobiography* (1960), a historical survey and a study of the chief problems, aspects, and varieties of autobiography; WILLIAM M. RUNYAN, *Life Histories and Psychobiography* (1982), a discussion of methodologies used in conducting psychobiographical research.

Anthologies: JAMES L. CLIFFORD (ed.), *Biography as an Art: Selected Criticism 1560–1960* (1962); WILLIAM H. DAVENPORT and BEN SIEGEL (eds.), *Biography Past and Present* (1965), contains a number of critical essays as well as biographical

selections; EDGAR JOHNSON (ed.), *A Treasury of Biography* (1941); JOHN C. METCALFE (ed.), *The Stream of English Biography* (1930).

Literary criticism. A useful compilation of essential texts is MARK SCHORER, JOSEPHINE MILES, and GORDON MCKENZIE (eds.), *Criticism,* rev. ed. (1958). The best survey of critical history is WILLIAM K. WIMSATT, JR., and CLEANTH BROOKS, *Literary Criticism: A Short History* (1957); G.M.A. GRUBE, *The Greek and Roman Critics* (1965); JOEL E. SPINGARN, *A History of Literary Criticism in the Renaissance,* 5th ed. (1925, paperback edition 1963); WALTER J. BATE, *From Classic to Romantic* (1946, reprinted 1961); and RENE WELLEK, *A History of Modern Criticism, 1750–1950,* 4 vol. (1955–65), are more specialized historical studies. Important theoretical statements are M.H. ABRAMS, *The Mirror and the Lamp* (1953); RENE WELLEK and AUSTIN WARREN, *Theory of Literature,* 3rd rev. ed. (1966); NORTHROP FRYE, *Anatomy of Criticism* (1957); and WAYNE C. BOOTH, *The Rhetoric of Fiction* (1961). WILLIAM EMPSON, *Seven Types of Ambiguity,* 3rd ed. (1956, reprinted 1963); ERICH AUERBACH, *Mimesis* (1946; Eng. trans. 1953); and LIONEL TRILLING, *The Liberal Imagination* (1950), are representative examples of modern criticism, combining theory with analysis of a wide variety of texts. See also DOUWE W. FOKKEMA and ELRUD KUNNE-IBSCH, *Theories of Literature in the Twentieth Century: Structuralism, Marxism, Aesthetics of Reception, Semiotics* (1978).

Children's literature. *Historical, critical:* (*Europe*): BETTINA HURLIMANN, *Europäische Kinderbücher in drei Jahrhunderten,* 2nd ed. (1963; Eng. trans., *Three Centuries of Children's Books in Europe,* 1968). (*England*): GILLIAN AVERY, *Nineteenth Century Children: Heroes and Heroines in English Children's Stories 1780–1900* (1965); FLORENCE V. BARRY, *A Century of Children's Books* (1922, reprinted 1968); MARCUS CROUCH, *Treasure Seekers and Borrowers: Children's Books in Britain 1900–1960* (1962); F.J. HARVEY DARTON, *Children's Books in England: Five Centuries of Social Life* (1932); ROGER LANCELYN GREEN, *Tellers of Tales: British Authors of Children's Books from 1800 to 1964,* rev. ed. (1965); PERCY MUIR, *English Children's Books, 1600 to 1900* (1954); M.F. THWAITE, *From Primer to Pleasure: An Introduction to the History of Children's Books in England, from the Invention of Printing to 1900* (1963); JOHN ROWE TOWNSEND, *Written for Children: An Outline of English Children's Literature* (1965). (*Canada*): SHEILA EGOFF, *The Republic of Childhood: A Critical Guide to Canadian Children's Literature in English* (1967). (*Anglo-American mainly*): CORNELIA MEIGS *et al., A Critical History of Children's Literature: A Survey of Children's Books in English from Earliest Times to the Present,* rev. ed. (1969). (*Germany*): IRENE DYHRENFURTH-GRAEBSCH, *Geschichte des deutschen Jugendbuches,* 3rd rev. ed. (1967); H.L. KOSTER, *Geschichte der deutschen Jugendliteratur* (1968). (*Sweden*): EVA VON ZWEIGBERGK, *Barnboken I Sverige 1750–1950* (1965). (*France*): MARIE-THERESE LATZARUS, *La Littérature enfantine en France dans la seconde moitié du XIX^e siècle* (1923); JEAN DE TRIGON, *Histoire de la littérature enfantine de ma Mère l'Oye au Roi Babar* (1950). (*Italy*): PIERO BARGELLINI, *Canto alle rondini: panorama storico della letteratura infantile,* 6th ed. (1967); GIUSEPPE FANCIULLI, *Scrittori per l'infanzia,* 3rd ed. (1968); LOUISE RESTIEAUX HAWKES, *Before and After Pinocchio: A Study of Italian Children's Books* (1933). (*Spain*): CARMEN BRAVO VILLASANTE (ed.), *Historia de la literatura infantil española* (1963). (*Latin America*): CARMEN BRAVO VILLASANTE, *Historia y antología de la literatura infantil iberoamericana,* 2 vol. (1966); DORA PASTORIZA DE ETCHEBARNE, *El cuento en la literatura infantil, ensayo crítico* (1962).

General: RICHARD BAMBERGER, *Jugendlektüre,* 2nd ed. (1965); ELEANOR CAMERON, *The Green and Burning Tree: On the Writing and Enjoyment of Children's Books* (1969); KORNEI CHUKOVSKY, *From Two to Five,* rev. ed. (1968; Eng. trans. of the 20th Russian ed. of 1968); HANS CORNIOLEY, *Beiträge zur Jugendbuchkunde* (1966); MARGERY FISHER, *Intent upon Reading: A Critical Appraisal of Modern Fiction for Children* (1961); PAUL HAZARD, *Les Livres, les enfants et les hommes* (1932; Eng. trans., *Books, Children and Men,* 4th ed., 1960); ENZO PETRINI, *Avviamento critico alla letteratura giovanile* (1958); LILLIAN H. SMITH, *The Unreluctant Years: A Critical Approach to Children's Literature* (1953); DOROTHY M. WHITE, *Books Before Five* (1954); KAY E. VANDERGRIFT, *Child and Story* (1981).

Bibliographic: VIRGINIA HAVILAND, *Children's Literature: A Guide to Reference Sources* (1966); ANNE PELLOWSKI, *The World of Children's Literature* (1968).

Biographical: BRIAN DOYLE (ed.), *The Who's Who of Children's Literature* (1968); MURIEL FULLER (ed.), *More Junior Authors* (1963); STANLEY J. KUNITZ and HOWARD HAYCRAFT (eds.), *The Junior Book of Authors,* 2nd ed. rev. (1951).

Illustration: BETTINA HURLIMANN, *Die Welt im Bilderbuch* (1965; Eng. trans., *Picture-Book World,* 1968); LEE KINGMAN, JOANNA FOSTER, and RUTH GILES LONTOFT (comps.), *Illustrators of Children's Books: 1957–1966* (1968); DIANA KLEMIN, *The Art of Art for Children's Books: A Contemporary Survey* (1966); BERTHA E. MAHONY *et al.* (comps.), *Illustrators of Children's Books 1744–1945* (1947); BERTHA MAHONY MILLER *et al.* (comps.), *Illustrators of Children's Books, 1946–1956* (1958).

The History of Western Literature

Diverse as they are, European literatures, like European languages, are parts of a common heritage. Greek, Latin, Germanic, Baltic and Slavic, Celtic, and Romance languages are all members of the Indo-European family. (Finnish and Hungarian and Semitic languages of the eastern Mediterranean, such as Hebrew, are not Indo-European. Literatures in these languages are, however, closely associated with major Western literatures and are often included among them.) The common literary heritage is essentially that originating in ancient Greece and Rome. It was preserved, transformed, and spread by Christianity and thus transmitted to the vernacular languages of the European Continent, the Western Hemisphere, and other regions that were settled by Europeans. To the present day, this body of writing displays a unity in its main features that sets it apart from the literatures of the rest of the world. Such common characteristics are considered here.

For specific information about the major national literatures or literary traditions of the West, see such *Macropædia* articles as AMERICAN LITERATURE; ENGLISH LITERATURE; GERMAN LITERATURE; GREEK LITERATURE; LATIN-AMERICAN LITERATURE; and SCANDINAVIAN LITERATURE. Various other Western literatures—including those in the Armenian, Bulgarian, Estonian, Lithuanian, and Romanian languages—are treated in separate entries of the *Micropædia.*

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Six, Division II, Section 621.

The article is divided into the following sections:

Ancient literature 224
Medieval literature 225
The Renaissance 226
The 17th century 226
The 18th century 227
The 19th century 227
  Romanticism 227
  Post-Romanticism 227
The 20th century 228
Bibliography 228

ANCIENT LITERATURE

The stark fact about ancient Western literature is that the greater part of it has perished. Some of it had been forgotten before it was possible to commit it to writing; fire, war, and the ravages of time have robbed posterity of most of the rest; and the restitutions that archaeologists and paleographers achieve from time to time are small. Yet surviving writings in Greek and far more in Latin have included those that on ancient testimony marked the heights reached by the creative imagination and intellect of the ancient world.

Five ancient civilizations, Babylon and Assyria, Egypt, Greece, Rome, and the culture of the Israelites in Palestine, each came into contact with one or more of the others. The two most ancient, Assyro-Babylonia, with its broken clay tablets, and Egypt, with its rotted papyrus rolls, make no direct literary signal to the modern age; yet Babylon produced the first full code of laws and two epics of archetypal myth, which came to be echoed and re-echoed in distant lands, and Egypt's mystical intuition of a supernatural world caught the imagination of the Greeks and Romans. Hebrew culture exerted its greatest literary influence on the West because of the place held by its early writings as the Old Testament of the Christian Bible; and this literature profoundly influenced Western consciousness through translation from about the time of St. Augustine onward into every vernacular language as well as into Latin. Until then, Judaism's concentrated spirituality set it apart from the Greek and Roman world.

Though influenced by the religious myths of Mesopotamia, Asia Minor, and Egypt, Greek literature has no direct literary ancestry and appears as self-originated. Roman writers looked to Greek precept for themes, treatment, and choice of verse and metre. Rome eventually passed the torch on to the early Middle Ages, by which time Greek had been subsumed under a wholly Latin tradition and was only rediscovered in its own right at the Renaissance—the "classical" tradition afterward becoming a threat to natural literary development, particularly when certain critics of the 17th century began to insist that the subjects and style of contemporary writing should conform with those employed by Greece and Rome.

All of the chief kinds of literature—epic, tragedy, comedy, lyric, satire, history, biography, and prose narrative—were established by the Greeks and Romans, and later developments have for the most part been secondary extensions. The Greek epic of Homer was the model for the Latin of Virgil; the lyric fragments of Alcaeus and Sappho were echoed in the work of Catullus and Ovid; the history of Thucydides was succeeded by that of Livy and Tacitus; but the tragedy of the great Athenians of the 5th century BC had no worthy counterpart in Roman Seneca nor had the philosophical writings of Plato and Aristotle in those of any ancient Roman, for the practical Romans were not philosophers. Whereas Greek writers excelled in abstraction, the Romans had an unusually concrete vision and, as their art of portraiture shows, were intensely interested in human individuality.

In sum, the work of these writers and others and perhaps especially that of Greek authors expresses the imaginative and moral temper of Western man. It has helped to create his values and to hand on a tradition to distant generations. Homer's epics extend their concern from the right treatment of strangers to behaviour in situations of deep involvement among hero rivals, their foes, and the overseeing gods; the tragedies of Aeschylus and Sophocles are a sublime expression of man's breakthrough into moral awareness of his situation. Among Roman authors an elevated Stoicism stressing the sense of duty is common to many, from Naevius, Ennius, and Cato to Virgil, Horace, and Seneca. A human ideal is to be seen in the savage satire of Juvenal and in Anacreon's songs of love and wine, as it is in the philosophical thought of Plato and Aristotle. It is given voice by a chorus of Sophocles, "Wonders are many, but none is more wonderful than man, the power that crosses the white sea. . . ." The human ideal held up in Greek and Latin literature, formed after civilization had emerged from earlier centuries of barbarism, was to be transformed, before the ancient world came to its close, into the spiritual ideal of Judeo-Christianity, whose writers foreshadowed medieval literature.

MEDIEVAL LITERATURE

Medieval, "belonging to the Middle Ages," is used here to refer to the literature of Europe and the eastern Mediterranean from as early as the establishment of the Eastern Roman, or Byzantine, Empire about AD 300 for medieval Greek, from the period following upon the fall of Rome in 476 for medieval Latin, and from about the time of Charlemagne and the Carolingian Renaissance he fostered in France (*c.* 800) to the end of the 15th century for most written vernacular literatures.

The establishment of Christianity throughout the territories that had formed the Roman Empire meant that Europe was exposed to and tutored in the systematic approach to life, literature, and religion developed by the early Church Fathers. In the West, the fusion of Christian and classical philosophy formed the basis of the medieval habit of interpreting life symbolically. Through St. Augustine, Platonic and Christian thought were reconciled: the permanent and uniform order of the Greek universe was given Christian form; nature became sacramental, a symbolic revelation of spiritual truth. Classical literature was invested with this same symbolism; exegetical, or interpretative, methods first applied to the Scriptures were extended as a general principle to classical and secular writings. The allegorical or symbolic approach that found in Virgil a pre-Christian prophet and in the *Aeneid* a narrative of the soul's journey through life to paradise (Rome) belonged to the same tradition as Dante's allegorical conception of himself and his journey in *The Divine Comedy.*

The role of the church in preserving ancient literatures

The church not only established the purpose of literature but preserved it. St. Benedict's monastery at Monte Cassino in Italy was established in 529, and other monastic centres of scholarship followed, particularly after the 6th- and 7th-century Irish missions to the Rhine and Great Britain and the Gothic missions up the Danube. These monasteries were able to preserve the only classical literature available in the West through times when Europe was being raided by Goths, Vandals, Franks, and, later, Norsemen in succession. The classical Latin authors so preserved and the Latin works that continued to be written predominated over vernacular works throughout most of the period. St. Augustine's *City of God,* the Venerable Bede's *Ecclesiastical History,* the Danish chronicle of Saxo Grammaticus, for example, were all written in Latin, as were most major works in the fields of philosophy, theology, history, and science.

The main literary values of the period are found in vernacular works. The pre-Christian literature of Europe belonged to an oral tradition that was reflected in the *Poetic Edda* and the sagas, or heroic epics, of Iceland, the Anglo-Saxon *Beowulf,* and the German *Song of Hildebrand.* These belonged to a common Germanic alliterative tradition, but all were first recorded by Christian scribes at dates later than the historical events they relate, and the pagan elements they contain were fused with Christian thought and feeling. The mythology of Icelandic literature was echoed in every Germanic language and clearly stemmed from a common European source. Only the Scandinavian texts, however, give a coherent account of the stories and personalities involved. Numerous ballads in different countries also reflect an earlier native tradition of oral recitation. Among the best known of the many genres that arose in medieval vernacular literatures were the romance and the courtly love lyric, both of which combined elements from popular oral traditions with those of more scholarly or refined literature and both derived largely from France. The romance used classical or Arthurian sources in a poetic narrative that replaced the heroic epics of feudal society, such as *The Song of Roland,* with a chivalrous tale of knightly valour. In the romance, complex themes of love, loyalty, and personal integrity were united with a quest for spiritual truth, an amalgam that was represented in every major western European literature of the time. The love lyric has had a similarly heterogeneous background. The precise origins of courtly love are disputed, as is the influence of a popular love poetry tradition; it is clear, however, that the idealized lady and languishing suitor of the poets of southern and northern France were imitated or reinterpreted throughout Europe—in the Sicilian school of Italy, the minnesingers (love poets) of Germany, and in a Latin verse collection, *Carmina Burana.*

Medieval drama began in the religious ceremonies that took place in church on important dates in the Christian calendar. The dramatic quality of the religious service lent itself to elaboration that perhaps first took the form of gestures and mime and later developed into dramatic

interpolations on events or figures in the religious service. This elaboration increased until drama became a secular affair performed on stages or carts in town streets or open spaces. The players were guild craftsmen or professional actors and were hired by towns to perform at local or religious festivals. Three types of play developed: the mystery, the miracle, and the morality. The titles and themes of medieval drama remained religious but their pieces' titles can belie their humorous or farcical and sometimes bawdy nature. One of the best known morality plays was translated from Dutch to be known in English as *Everyman.* A large majority of medieval literature was anonymous and not easily dated. Some of the greatest figures—Dante, Chaucer, Petrarch, and Boccaccio—came late in the period, and their work convincingly demonstrates the transitional nature of the best of medieval literature, for, in being master commentators of the medieval scene, they simultaneously announced the great themes and forms of Renaissance literature.

THE RENAISSANCE

The name Renaissance ("Rebirth") is given to the historical period in Europe that succeeded the Middle Ages. The awakening of a new spirit of intellectual and artistic inquiry, which was the dominant feature of this political, religious, and philosophical phenomenon, was essentially a revival of the spirit of ancient Greece and Rome; in literature this meant a new interest in and analysis of the great classical writers. Scholars searched for and translated "lost" ancient texts, whose dissemination was much helped by developments in printing in Europe from about 1450.

Art and literature in the Renaissance reached a level unattained in any previous period. The age was marked by three principal characteristics: first, the new interest in learning, mirrored by the classical scholars known as humanists and instrumental in providing suitable classical models for the new writers; second, the new form of Christianity, initiated by the Protestant Reformation led by Martin Luther, which drew men's attention to the individual and his inner experiences and stimulated a response in Catholic countries summarized by the term Counter-Reformation; third, the voyages of the great explorers that culminated in Christopher Columbus' discovery of America in 1492 and that had far-reaching consequences on the countries that developed overseas empires, as well as on the imaginations and consciences of the most gifted writers of the day.

To these may be added many other factors, such as the developments in science and astronomy and the political condition of Italy in the late 15th century. The new freedom and spirit of inquiry in the Italian city-states had been a factor in encouraging the great precursors of the Renaissance in Italy, Dante, Petrarch, and Boccaccio. The flowering of the Renaissance in France appeared both in the poetry of the poets making up the group known as the Pléiade and in the reflective essays of Michel de Montaigne, while Spain at this time produced its greatest novelist, Miguel de Cervantes. Another figure who stood out above his contemporaries was the Portuguese epic poet Luís Camões, while drama flourished in both Spain and Portugal, being represented at its best by Lope de Vega and Gil Vicente. In England, too, drama dominated the age, a blend of Renaissance learning and native tradition lending extraordinary vitality to works of Christopher Marlowe, Ben Jonson, John Webster, and others, while Shakespeare, England's greatest dramatic and poetic talent, massively spanned the end of the 16th century and the beginning of the 17th.

In the 16th century the Dutch scholar Desiderius Erasmus typified the development of humanism, which embodied the spirit of critical inquiry, regard for classical learning, intolerance of superstition, and high respect for men as God's most intricate creation. An aspect of the influence of the Protestant Reformation on literature was the number of great translations of the Bible, including an early one by Erasmus, into vernacular languages during this period, setting new standards for prose writing. The impetus of the Renaissance carried well into the 17th century, when John Milton reflected the spirit of Christian humanism.

THE 17TH CENTURY

The 17th century was a period of unceasing disturbance and violent storms, no less in literature than in politics and society. The Renaissance had prepared a receptive environment essential to the dissemination of the ideas of the new science and philosophy. The great question of the century, which confronted serious writers from Donne to Dryden, was Michel de Montaigne's "What do I know?" or, in expanded terms, the ascertainment of the grounds and relations of knowledge, faith, reason, and authority in religion, metaphysics, ethics, politics, economics, and natural science.

The questioning attitude that characterized the period is seen in the works of its great scientists and philosophers: Descartes's *Discourse on Method* (1637) and Pascal's *Pensées* (written 1657–58) in France; Bacon's *Advancement of Learning* (1605) and Hobbes's *Leviathan* (1651) in England. The importance of these works has lain in their application of a skeptical, rationalist mode of thought not only to scientific problems but to political and theological controversy and general problems of understanding and perception. This fundamental challenge to both thought and language had profound repercussions in man's picture of himself and was reflected in what T.S. Eliot described as "the dissociation of sensibility," which Eliot claimed took root in England after the Civil War, whereby, in contrast to the Elizabethan and Jacobean writers who could "devour any kind of experience," later poets in English could not think and feel in a unified way.

The effects of war on 17th-century literature

A true picture of the period must also take into account the enormous effect of social and political upheavals during the early and middle parts of the century. In England, where the literary history of the period is usually divided into two parts, the break seems to fall naturally with the outbreak of the Civil War (1642–51), marked by a closure of the theatres in 1642, and a new age beginning with the restoration of the monarchy in 1660. In France the bitter internecine struggle of the Fronde (1648–53) similarly divided the century and preceded possibly the greatest period of all French literature—the age of Molière, Racine, Boileau, and La Fontaine. In Germany the early part of the century was dominated by the religious and political conflicts of the Thirty Years' War (1618–48) and thereafter by the attempts of German princes to emulate the central power and splendour of Louis XIV's French court at Versailles. The Netherlands was also involved in the first part of the century in a struggle for independence from Spain (the Eighty Years' War, 1568–1648) that resulted not only in the achievement of this but also in the "Golden Age" of Dutch poetry—that of Henric Spieghel, Daniël Heinsius, and Gerbrand Bredero.

The civil, political, and religious conflicts that dominated the first half of the century were in many ways also the characteristic response of the Counter-Reformation. The pattern of religious conflict was reflected in literary forms and preoccupations. One reaction to this—seen particularly in Italy, Germany, and Spain but also in France and England—was the development of a style in art and literature known as Baroque. This development manifested itself most characteristically in the works of Giambattista Marino in Italy, Luis de Góngora in Spain, and Martin Opitz in Germany. Long regarded by many critics as decadent, Baroque literature is now viewed in a more favourable light and is understood to denote a style the chief characteristics of which are elaboration and ornament, the use of allegory, rhetoric, and daring artifice.

Baroque literature

If Baroque literature was the characteristic product of Italy and Germany in this period, Metaphysical poetry was the most outstanding feature in English verse of the first half of the century. This term, first applied by Dryden to John Donne and expanded by Dr. Johnson, is now used to denote a range of poets who varied greatly in their individual styles but who possessed certain affinities with Baroque literature, especially in the case of Richard Crashaw.

Perhaps the most characteristic of all the disputes of the 17th century was that in which the tendency to continue to develop the Renaissance imitation of the classics came into conflict with the aspirations and discoveries of new

The conflict between the classics and the new science

thinkers in science and philosophy and new experimenters with literary forms. In France this appeared in a struggle between the Ancients and Moderns, between those who thought that literary style and subject should be modeled on classical Greek and Latin literature and supporters of native tradition. In Spain a similar conflict was expressed in a tendency toward ornament, Latinization, and the classics (*culteranismo*) and that toward a more concise, profound, and epigrammatic style (*conceptismo*). This conflict heralded through the Moderns in France and the idea of *conceptismo* in Spain a style of prose writing suitable to the new age of science and exploration. The Moderns in France were largely, therefore, followers of Descartes. In England a similar tendency was to be found in the work of the Royal Society in encouraging a simple language, a closer, naked, natural way of speaking, suitable for rational discourse, paralleled by the great achievements in prose of John Milton and John Dryden.

THE 18TH CENTURY

To call the 18th century the Age of Reason is to seize on a useful half-truth but to cause confusion in the general picture, because the primacy of reason had also been a mark of certain periods of the previous age. It is more accurate to say that the 18th century was marked by two main impulses: reason and passion. The respect paid to reason was shown in pursuit of order, symmetry, decorum, and scientific knowledge; the cultivation of the feelings stimulated philanthropy, exaltation of personal relationships, religious fervour, and the cult of sentiment, or sensibility. In literature the rational impulse fostered satire, argument, wit, plain prose; the other inspired the psychological novel and the poetry of the sublime.

The cult of wit, satire, and argument is evident in England in the writings of Alexander Pope, Jonathan Swift, and Samuel Johnson, continuing the tradition of Dryden from the 17th century. The novel was established as a major art form in English literature partly by a rational realism shown in the works of Henry Fielding, Daniel Defoe, and Tobias Smollett and partly by the psychological probing of the novels of Samuel Richardson and of Laurence Sterne's *Tristram Shandy*. In France the major characteristic of the period lies in the philosophical and political writings of the Enlightenment, which had a profound influence throughout the rest of Europe and foreshadowed the French Revolution. Voltaire, Jean-Jacques Rousseau, Charles de Montesquieu, and the Encyclopédistes Denis Diderot and Jean d'Alembert all devoted much of their writing to controversies about social and religious matters, often involving direct conflict with the authorities. In the first part of the century, German literature looked to English and French models, although innovative advances were made by the dramatist and critic Gotthold Ephraim Lessing. The great epoch of German literature came at the end of the century, when cultivation of the feelings and of emotional grandeur found its most powerful expression in what came to be called the *Sturm und Drang* ("Storm and Stress") movement. Associated with this were two of the greatest names of German literature, Johann Wolfgang von Goethe and Friedrich Schiller, both of whom in drama and poetry advanced far beyond the turbulence of *Sturm und Drang*.

THE 19TH CENTURY

The 19th century in Western literature—one of the most vital and interesting periods of all—has special interest as the formative era from which many modern literary conditions and tendencies derived. Influences that had their origins or were in development in this period—Romanticism, Symbolism, Realism—are reflected in the current of modern literature, and many social and economic characteristics of the 20th century were determined in the 19th.

Romanticism. The predominant literary movement of the early part of the 19th century was Romanticism, which in literature had its origins in the *Sturm und Drang* period in Germany. An awareness of this first phase of Romanticism is an important correction to the usual idea of Romantic literature as something that began in English poetry with William Wordsworth and Samuel Taylor Coleridge and the publication of *Lyrical Ballads* in 1798. Moreover, although it is true that the French Revolution of 1789 and the Industrial Revolution were two main political and social factors affecting the Romantic poets of early 19th-century England, many characteristics of Romanticism in literature sprang from literary or philosophical sources. A philosophical background was provided in the 18th century chiefly by Jean-Jacques Rousseau, whose emphasis on the individual and the power of inspiration influenced Wordsworth and also such first-phase Romantic writers as Friedrich Hölderlin and Ludwig Tieck in Germany and the French writer Bernardin de Saint-Pierre, whose *Paul et Virginie* (1787) anticipated some of the sentimental excesses of 19th-century Romantic literature. Positive as it was, the influence of Rousseau must also be seen as a partly negative reaction against 18th-century rationalism with its emphasis on intellect.

Belief in self-knowledge was, indeed, a principal article of Romantic faith. Late 18th-century French writers such as Fabre d'Olivet sought to explain the physical world by an idea of a "breath of life" similar to the "inspiration" of Wordsworth and Coleridge. The Romantics believed that the real truth of things could be explained only through examination of their own emotions in the context of nature and the primitive. Because of this emphasis on inspiration, the poet came to assume a central role—that of seer and visionary. Simultaneously, such formal conventions as imitation of the classics were rejected as binding rules. A new directness of the poet's role emphasized the language of the heart and of ordinary men, and Wordsworth even tried to invent a new simplified diction. Poetry became divorced from its 18th-century social context, and a poet was answerable only to ultimate truth and himself. Two classic poses of the Romantic poet were the mystic visionary of John Keats and the superman of Lord Byron—indeed, satirization of the Byronic hero was to become a theme of later novelists such as Fyodor Dostoyevsky, even though he himself had Romantic antecedents.

The fact that Dostoyevsky was a Russian showed how the Romantic stream flowed across Europe. In Spain and Italy, Hungary, Poland, and the Balkans, it took the form of drama, which in England failed to produce great works. The early and middle 19th century was a time of poetry and prose rather than of drama. The Romantic style in poetry was seen everywhere in Europe—in José de Espronceda in Spain; Ugo Foscolo and Giacomo Leopardi in Italy, where it became identified with nationalist sentiments; Aleksandr Pushkin and Mikhail Lermontov in Russia; Adam Mickiewicz in Poland. In America, a Romantic thread also allied with the emergence of national feeling could be seen in the adventurous stories of James Fenimore Cooper; in the supernatural and mystic element in Edgar Allan Poe; in the poetry of Walt Whitman and Henry Wadsworth Longfellow; and in the Transcendentalist theories of Ralph Waldo Emerson and Henry David Thoreau, which, as Wordsworth's pronouncements had done, affirmed the power of "insight" to transcend ordinary logic and experience.

The impetus of Romantic poetry began to slacken after about 1830 and gave way to more objective styles, although many of its themes and devices, such as the misunderstood artist or the unhappy lover, continued to be employed.

Post-Romanticism. Arguably the first post-Romantic poet was a German, Heinrich Heine, but German poetry in the mid-19th century mostly followed Wordsworth, though new tendencies were to be found in August von Platen Hallermünde and an Austrian, Nikolaus Lenau. The principal development was to be seen in France in the growth of a movement known as Parnassianism. Originating with Théophile Gautier, Parnassianism in some ways was an offshoot of Romanticism rather than a reaction against it. In concentrating on the purely formal elements of poetry, on aesthetics, and on "art for art's sake," it changed the direction of French poetry and had much influence abroad. Its most illustrious representative was Charles Baudelaire, who believed that "everything that is not art is ugly and useless." Another branch of new development was the growth of Impressionism and

Impressionism and Symbolism

the Symbolist movement, a result of "borrowing" from movements in painting, sculpture, and music. Paul Verlaine, foremost of the Impressionists, used suggestion, atmosphere, and fleeting rhythms to achieve his effects. Symbolism, a selective use of words and images to evoke tenuous moods and meanings, is conveyed in the work of Stéphane Mallarmé and Arthur Rimbaud. The advance of French poetry in the middle and later part of the century was an achievement of individuals, based on invention of a personal idiom.

The spread of education and, in England, of circulating libraries increased a demand for novels. At the beginning of the 19th century Jane Austen had already satirized the excesses of the Gothic novel, a harbinger of medievalizing Romanticism in the latter part of the 18th century, in *Northanger Abbey* and the conflict of sense and Romantic sensibility in *Sense and Sensibility.* In France the conflict of intelligence and emotion appeared in the work of Benjamin Constant (*Adolphe,* 1816) and most notably in *Le Rouge et le noir* (1830) of Stendhal and later in Gustave Flaubert's *Madame Bovary* (1857). The detailed verbal scrupulousness and Realism exhibited in the work of Flaubert and of Honoré de Balzac were carried forward by Guy de Maupassant in France and Giovanni Verga in Italy; they culminated in the extreme Naturalism of Émile Zola, who described his prose in novels such as *Thérèse Raquin* (1867) as "literary surgical autopsy."

But Realism and nationalism seem irrelevant as descriptions of the great writers of the period—for example, George Eliot, Charles Dickens, and Thomas Hardy in England and Nikolay Gogol, Ivan Turgenev, Leo Tolstoy, Fyodor Dostoyevsky, and Anton Chekhov in Russia. In such writers there was a distinct bias toward literature with a social purpose, stimulated by awakening forces of liberalism, humanism, and socialism in many Western countries.

A decline of the Romantic theatre into melodrama was fairly general in Europe, and it was slower than the novel to take up problems of contemporary life. When revival came, through the work of a Norwegian, Henrik Ibsen, Romantic conflicts of visionary and realist, individual and society were restated, and this was true also of the plays of August Strindberg in Sweden. In Russia a modern theatre became a vital influence that could trace its beginnings back to Gogol's *Government Inspector* (1836) but was to be felt later in the century in Turgenev's *Month in the Country* (1850) and, above all, in the work of Anton Chekhov, a great dramatist of the period.

THE 20TH CENTURY

When the 20th century began, social and cultural conditions that prevailed in Europe and America were not too different from those of the middle and late 19th century. Continuity could be seen, for example, in the work of four novelists writing in English at the turn of the century and after. Joseph Conrad, Thomas Hardy, Henry James, and D.H. Lawrence all demonstrated in the progress of their work the transition from a relatively stable world at the end of the 19th century to a new age that began with World War I. The awakening of a new consciousness in literature was also to be traced in such works of fiction as the first volume of Marcel Proust's *Remembrance of Things Past* (*Swann's Way,* 1913), André Gide's *Vatican Cellars* (1914), James Joyce's *Ulysses* (1922), Franz Kafka's *Trial* (published posthumously in 1925), and Thomas Mann's *Magic Mountain* (1924).

Various influences that characterized much of the writing from the 1920s were at work in these writers. An interest in the unconscious and the irrational was reflected in their work and that of others of about this time. Two important sources of this influence were Friedrich Nietzsche, a German philosopher to whom both Gide and Mann, for example, were much indebted, and Sigmund Freud, whose psychoanalytical works, by the 1920s, had had a telling influence on Western intellectuals. A shift away from 19th-century assumptions and styles was not limited to writers of fiction. André Breton's first *Manifeste du surréalisme* (1924; "Manifesto of Surrealism") was the first formal statement of a movement that called for spontaneity and a complete rupture with tradition. Surrealism showed the influence of Freud in its emphasis on dreams, automatic writing, and other antilogical methods and, although short-lived as a formal movement, had a lasting effect on much 20th-century art and poetry. The uncertainty of the new age and the variety of attempts to deal with it and give it some artistic coherence can be seen also in Rainer Maria Rilke's *Duino Elegies* and *Sonnets to Orpheus* (1923); in T.S. Eliot's *Waste Land* (1922); and in Luigi Pirandello's play about the instability of identity, *Henry IV* (1922).

The influence of Nietzsche and Freud

The international and experimental period of Western literature in the 1910s and 1920s was important not only for the great works it produced but also because it set a pattern for the future. What was clearly revealed in the major works of the period was an increasing sense of crisis and urgency, doubts as to the 19th century's faith in the psychological stability of the individual personality, and a deep questioning of all philosophical or religious solutions to human problems. In the 1930s these qualities of 20th-century thought were not abandoned but, rather, were expanded into a political context, as writers divided into those supporting political commitment in their writing and those reacting conservatively against such a domination of art by politics. Nor did World War II resolve the debate concerning political commitment—issues similar to those that exercised major creative imaginations of the 1930s were still very much alive during the last quarter of the century.

The scarcity of great writers after World War II

It would be tempting to explain what seemed to be a relative scarcity of great writers in the period after World War II as an inevitable result of the cumulative pressure of disturbing social and technological developments accelerated by that war. Under such fluctuating and doubtful circumstances, it would not seem altogether strange if writing and reading, as traditionally understood, should cease. Indeed, in certain technologically highly developed countries, such as the United States, the printed word itself seemed to some critics to have lost its central position, having been displaced in the popular mind by a visual and aural electronic culture that did not need the active intellectual participation of its audience. Thus the communications media that helped to create something resembling an international popular culture in many Western countries did nothing to make the question of literary value easier to answer. Given the extraordinary conditions in which a modern writer works, it was not surprising that reputations were difficult to judge, that radical experimentation characterized many fields of literature, and that traditional forms of writing were losing their definition and were tending to dissolve into one another. Novels might acquire many features of poetry or be transformed into a kind of heightened nonfictional reportage, while experimentation with typography gave poems an appearance of verbal paintings, and dramatic works, shorn of anything resembling a traditional plot, became a series of carefully orchestrated gestures or events. But formal experimentation was only part of the picture, and to say that modern writing since World War II has been primarily experimental would be to ignore other characteristics that writing acquired earlier in the century and that still continued to be issues. Most good critics felt that there was no lack of good literature being written, despite the lack of major reputations and despite the possibly transitional nature of much of the period's work in its variety of styles and subjects.

BIBLIOGRAPHY

General introductory works: FERNAND BALDENSPERGER and WERNER P. FRIEDERICH, *Bibliography of Comparative Literature* (1950, reprinted 1960), comprehensive coverage of literatures, classified by author, country, and according to genre, theme, etc.; ANTONY BRETT-JAMES, *The Triple Stream: Four Centuries of English, French, and German Literature, 1531–1930* (1953, reprinted 1977), contains useful comparative tables arranged chronologically; JOHN M. COHEN, *A History of Western Literature,* rev. ed. (1963), a useful factual and critical account; JEAN-ALBERT BEDE and WILLIAM B. EDGERTON (gen. eds.), *Columbia Dictionary of Modern European Literature,* 2nd ed. (1980), contains over 1,800 articles, including surveys, critical essays, and biographies, with bibliographies; WILLIAM F. THRALL and ADDISON HIBBARD, *A Handbook to Literature,* 4th ed. by C. HUGH

HOLMAN (1980), good coverage of literary terms and movements; RENE WELLEK, *A History of Modern Criticism: 1750–1950,* 4 vol. (1955–), surveys theory in the major European nations since the Renaissance.

Ancient period: JULIUS A. BEWER, *The Literature of the Old Testament in Its Historical Development,* 3rd ed. (1962); SIR PAUL HARVEY (ed.), *The Oxford Companion to Classical Literature* (1937, reprinted with corrections 1969), a basic reference work; MAURICE PLATNAUER (ed.), *Fifty Years of Classical Scholarship* (1954), contains valuable bibliographies; C.M. BOWRA, *Landmarks in Greek Literature* (1966); ROBERT FLACELIERE, *Histoire littéraire de la Grèce* (1962; Eng. trans., *A Literary History of Greece,* 1964), a systematic, readable survey of ancient literature; KENNETH J. DOVER *et al., Ancient Greek Literature* (1980); HERBERT J. ROSE, *A Handbook of Greek Literature from Homer to the Age of Lucian,* 4th ed. rev. (1961); and *A Handbook of Latin Literature from the Earliest Times to the Death of St. Augustine,* 3rd ed. (1954, reprinted with suppl. bibliog. 1966), two standard works; MICHAEL GRANT, *Roman Literature* (1954), a good short account; WILLIAM A. LAIDLAW, *Latin Literature* (1951), a useful introduction; JOHN WIGHT DUFF, *A Literary History of Rome,* 2nd–3rd ed., 2 vol., ed. by A.M. DUFF (1960–64, 3rd ed. reprinted 1979), standard introduction containing comprehensive and scholarly surveys, with supplementary bibliographies.

The Middle Ages: H.M. CHADWICK, *The Heroic Age* (1912, reprinted 1974), a comparative study covering Greek, Teutonic, Slavonic, and Celtic literatures and cultures; E.K. CHAMBERS, *The Mediaeval Stage,* 2 vol. (1903), a standard reference work; W.T.H. JACKSON, *The Literature of the Middle Ages* (1960), and (ed.), *The Interpretation of Medieval Lyric Poetry* (1980); CHARLES W. JONES (ed.), *Medieval Literature in Translation* (1950), a good, readable anthology; C.S. LEWIS, *The Allegory of Love* (1936, reprinted 1958), a critical study of courtly love in medieval literature; ROGER S. LOOMIS (ed.), *Arthurian Literature in the Middle Ages* (1959), an excellent survey.

The Renaissance: ERNST CASSIRER, PAUL O. KRISTELLER, and JOHN H. RANDALL (eds.), *The Renaissance Philosophy of Man* (1948), a collection of excerpts from the writings of Renaissance philosophers; JOSEPH A. MAZZEO, *Renaissance and Revolution: The Remaking of European Thought* (1966), useful background reading; HENRY O. TAYLOR, *Thought and Expression in the Sixteenth Century,* 2nd rev. ed., 2 vol. (1959), a good standard work.

The 17th century: MARIO PRAZ, *Studi sul concettismo,* 2 vol. (1934–46; Eng. trans., *Studies in Seventeenth-Century Imagery,* 2 vol., 1939–48), a standard and readable work; JOEL E. SPINGARN (ed.), *Critical Essays of the Seventeenth Century,* 3 vol. (1909, reprinted 1957), a valuable collection.

The 18th century: LILIAN R. FURST, *Romanticism in Perspective,* 2nd ed. (1979), on the Romantic movement in France, England, and Germany, and *The Contours of European Romanticism* (1979); JOHN B. HALSTED (ed.), *Romanticism* (1969), a collection of extracts and key essays by the leading figures of western European Romanticism, with a long introduction and chronological table; PAUL HAZARD, *La Pensée européenne au XVIII^e siècle* (1963; Eng. trans., *European Thought in the Eighteenth Century,* 1963); DANIEL MORNET, *The Development of Literature and Culture in the XVIIIth Century* (1954), an extensive survey; MARIO PRAZ, *La carne, la morte e il diavolo nella letteratura romantica* (1930; 5th ed., 1976; Eng. trans., *The Romantic Agony,* 2nd ed., 1951; reissued with corrections, 1970), a remarkable work, now a classic in this field; JOHN G. ROBERTSON, *Studies in the Genesis of Romantic Theory in the Eighteenth Century* (1923, reprinted 1962).

The 19th century: GEORG BRANDES, *Creative Spirits of the Nineteenth Century* (1923, reprinted 1967); BENEDETTO CROCE, *Poesia e non poesia,* 7th ed. (1964; Eng. trans., *European Literature in the Nineteenth Century,* 1924, reprinted 1967); JANKO LAVRIN, *Aspects of Modernism: From Wilde to Pirandello* (1935, reprinted 1968); JOHN LUCAS (ed.), *Literature and Politics in the Nineteenth Century* (1971).

The 20th century: ERICH AUERBACH, *Mimesis* (1946; Eng. trans. 1953), traces the development of Realism from ancient to modern times; CLEANTH BROOKS and ROBERT PENN WARREN (eds.), *Understanding Poetry,* 4th ed. (1976), and *Understanding Fiction,* 3rd ed. (1979); and CLEANTH BROOKS and ROBERT B. HEILMAN, *Understanding Drama,* rev. ed. (1948), detailed anthologies, with explications of texts—invaluable introductions, especially to the study of modern writing, though not limited to 20th-century literature; GYORGY LUKACS, *Studies in European Realism* (1950, reprinted 1972), an important work by a major Marxist critic; RENE WELLEK and AUSTIN WARREN, *Theory of Literature,* 3rd ed. rev. (1970), a general and theoretical work, especially useful as a guide to modern literature; CLAUDE MAURIAC, *L'Alittérature contemporaine* (1958; Eng. trans., *The New Literature,* 1959); STEPHEN SPENDER, *The Struggle of the Modern* (1963) and *The Thirties and After* (1978); EDMUND WILSON, *Axel's Castle* (1931), *The Triple Thinkers,* rev. ed. (1948, reprinted 1977), and *The Twenties: From Notebooks and Diaries of the Period,* ed. by LEON EDEL (1975).

(Ed.)

Locke

The English philosopher John Locke was an initiator of the Enlightenment in England and France, an inspirer of the U.S. Constitution, and the author of, among other works, *An Essay Concerning Human Understanding,* his account of human knowledge, including the "new science" of his day—*i.e.,* modern science.

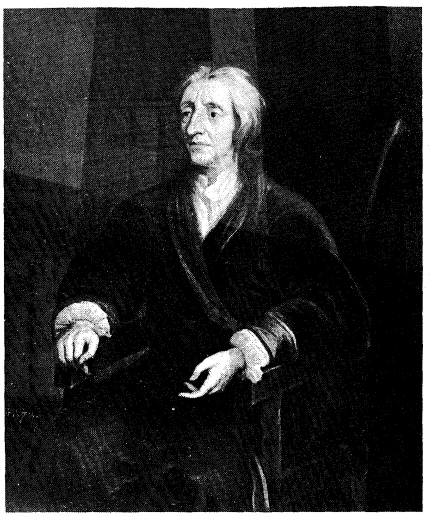
THE LIFE OF JOHN LOCKE

Early years. Locke was born in Wrington, Somerset, on Aug. 29, 1632, and reared in Pensford, six miles south of Bristol. His family was Anglican with Puritan leanings. His father, a country attorney of modest means, fought on the Parliamentarian side in the Civil War—a fact that later helped him to find a place for his son in Westminster School, then controlled by a Parliamentarian committee (though its headmaster, Richard Busby, was a Royalist). The training there was thorough, but Locke later complained of the severity of its discipline. In 1652 he entered Christ Church, Oxford. Puritan reforms at Oxford had not yet altered the traditional Scholastic curriculum of rhetoric, grammar, moral philosophy, geometry, and Greek; Locke found the course insipid and interested himself in studies outside the traditional program, particularly experimental science and medicine. He was graduated with a B.A. degree in 1656 and an M.A. two years later, around which time he was elected a student (the equivalent of fellow) of Christ Church. In 1660, as a newly appointed tutor in his college, Locke enthusiastically welcomed the end of the Puritan Commonwealth and the restoration of Charles II to the throne.

Student years

In 1661 Locke inherited a portion of his father's estate, which ensured a modest annual income. His studentship would eventually be subject to termination unless he took holy orders, which he declined to do. Not wishing to make teaching his permanent vocation, he taught undergraduates for four years only. He served as secretary to a diplomatic mission to Brandenburg in 1665, and on his return he was immediately offered, but refused, another diplomatic post. His papers of this period, his correspondence, and his commonplace books all testify to his chief interests at the time, viz., natural science, on the one hand, and the study of the underlying principles of moral, social, and political life, on the other. To remedy the narrowness of his education he read contemporary philosophy, particularly that of René Descartes, the father of modern philosophy. But more than all, experimental science engaged his interest. He collaborated with Robert Boyle, one of the founders of modern chemistry, who was a close friend, and, toward the end of the period, with another friend, Thomas Sydenham, an eminent medical scientist.

Association with Shaftesbury. It was as a physician that Locke first came to the notice of the statesman Lord Ashley (later to become the 1st earl of Shaftesbury). On a visit to Oxford in the summer of 1666, Lord Ashley required some medical attention and was introduced to Locke by a mutual acquaintance; the two immediately



Locke, oil painting by Sir Godfrey Kneller. In Christ Church, Oxford.
By courtesy of the Governing Body of Christ Church, Oxford

became friends. A royal mandate of that November secured Locke's studentship indefinitely. The following year, despite his having no medical degree and no desire to practice medicine, he joined Ashley's household at Exeter House in the Strand in London as family physician. He became Ashley's personal adviser not merely on medical matters but on his general affairs as well.

Ashley was a forceful, aggressive politician who had many enemies (some of them men of letters—for instance, Locke's schoolfellow, the poet laureate John Dryden). It is doubtful, however—if only in view of Locke's respect for him—whether Ashley was as evil as his enemies sometimes made him out to be. It is known that he stood firmly for a constitutional monarchy, for a Protestant succession, for civil liberty, for toleration in religion, for the rule of Parliament, and for the economic expansion of Britain; and that he continued to make this stand when many influential men were working against these aims. Since these were already aims to which Locke had dedicated himself, there existed from the first a perfect understanding between the statesman and his adviser, one that meant much to both. Ashley entrusted Locke with the task of negotiating his son's marriage with the daughter of the Earl of Rutland; he also made him secretary of the group that he had formed to increase trade with America, particularly with the southern colonies. Locke helped to draft a constitution for the new colony of Carolina, a document that extended freedom of worship to all colonists, denying admission only to atheists.

Private studies and discussions

During the following decades, Locke persevered in his private studies, and many of his social meetings were in effect meetings with friends to discuss philosophical and scientific problems. As early as 1668 he had become a fellow of the newly formed (1663) Royal Society, which kept him in touch with scientific advances. It is known, too, that groups of friends (Lord Ashley; the physician John Mapletoft; Thomas Sydenham; Sydenham's physician colleague, James Tyrrell, who was also a divine; and others) met in his rooms, for one such meeting is mentioned in the preface of his *Essay Concerning Human Understanding,* in which he reports that, because of the difficulties that beset the participants, they resolved to devote their next meeting to discussing the powers of the mind in order, as they said, "to examine our own abilities and see what objects our understandings were, or were not, fitted to deal with." Locke himself opened the discussion and, following the meeting, set out his view of human knowledge in two drafts (1671), still extant, which show the beginnings of the thinking that 19 years later would blossom into his famous *Essay.* In these London years, too, Locke encountered representatives of Cambridge Platonism, a school of Christian humanists, who, though sympathetic to empirical science, nonetheless opposed materialism because it failed to account for the rational element in human life. They tended to be liberal in both politics and religion. Insofar as they taught a Platonism that rested on belief in innately known Ideas, Locke could not follow them; but their tolerance, their emphasis on practical conduct as a part of the religious life, and their rejection of materialism were features that he found most attractive. This school was closely related in spirit to another school that influenced Locke at this time, viz., that of latitudinarianism. For the latter school, if a man confessed Christ, that alone should be enough to entitle him to membership in the Christian Church; conformity in nonessentials should not be demanded. These movements prepared Locke for the antidogmatic, liberal school of theology that he would later encounter in Holland, a school in revolt against the narrowness of traditional Calvinism.

In 1672 Ashley was raised to the peerage as the 1st earl of Shaftesbury and at the end of that year was appointed lord high chancellor of England. Though he soon lost favour and was dismissed, he did, while in office, establish the Council of Trade and Plantations, of which Locke was secretary for two years. Locke, however, who suffered greatly from asthma, found the London air and his heavy duties unhealthy, and in 1675 he had to return to Oxford.

Intellectual contacts in France

Six months later he departed for France, where he stayed for four years (1675–79), spending most of his time in Paris and Montpellier. In France during the 1670s, Locke made contacts that deeply influenced his view of metaphysics and epistemology, viz., with the Gassendist school and, particularly, with its leader, François Bernier. Pierre Gassendi, a philosopher and scientist, had rejected overspeculative elements in Descartes's philosophy and had advocated a return to Epicurean doctrines—*i.e.,* to empiricism (stressing sense experience), to hedonism (holding pleasure to be the good), and to corpuscular physics (according to which reality consists of atomic particles). Knowledge of the external world, Gassendi held, depends upon the senses, though it is through reasoning that man may derive much further information from empirically gained evidence.

Upon Locke's return to England, he found the country torn by dissension. The heir to the throne, James (the brother of Charles II), was a Roman Catholic, whom the Protestant majority led by Shaftesbury wished to exclude from the succession. For a year Shaftesbury had been imprisoned in the Tower, but by the time Locke returned he was back in favour once more as lord president of the Privy Council. When he failed, however, to reconcile the interests of the King and Parliament, he was dismissed; in 1681 he was arrested, tried, and finally acquitted by a London jury. A year later he fled to Holland, where, in 1683, he died.

Later life. No one of Shaftesbury's known friends was now safe in Great Britain. Locke himself, who was being closely watched, crossed to Holland in September 1683.

Exile in Holland. Locke's sojourn in Holland was happier than he had expected it to be: his health improved, he made many new friends, and he found the leisure that enabled him to bring his thoughts on many subjects to fruition. Locke spent his first winter in Amsterdam and soon became friendly with a distinguished Arminian theologian, Philip van Limborch, pastor of the Remonstrants' church there—a friendship that lasted until Locke's death. The companionship of Philip and other friends made it easier to bear bad news from home: at Charles II's express command, Locke (in 1684) was deprived of his studentship at Christ Church. The next year his name appeared on a list sent to The Hague that named 84 traitors wanted by the English government. Locke went into hiding for a while but soon was able to move freely over Holland and became familiar with its different provinces.

Return to England and retirement to Oates. Locke remained abroad for more than five years, until James II, who had become king in 1685, was overthrown. In the autumn of 1688, after it was announced that James had been presented with a male heir (and thus a Roman Catholic successor), the King's opponents invited his Protestant nephew and son-in-law, William of Orange in the Netherlands, to seize the throne. The King offered

little resistance. Locke himself in February 1689 crossed in the party that accompanied the Princess of Orange, now to be crowned Queen Mary II of England. The triumph was complete; Locke was home again, although not without a nostalgia for the Holland that he had come to love. He now took little part in public life. He refused ambassadorial posts but accepted a membership in the Commission of Appeals. (Much later, in 1696, he was appointed a commissioner in the resuscitated Board of Trade and Plantations, however, and for four years played a leading part in its deliberations.) But the London air again bothered him, and he was forced to leave the city for long visits to his friends in the country. In 1691 he retired to Oates, the house of his friends Sir Francis and Lady Masham in Essex, and subsequently made only occasional visits to London. Nonetheless, he was not without influence in these last years of his life, for he was the intellectual leader of the Whigs. Their principal parliamentarians were frequently old friends of Locke, and the younger generation—particularly the ablest of them all, John Somers, who soon became lord chancellor—turned to him constantly for guidance. In "the glorious, bloodless revolution," the main aims for which Shaftesbury and Locke had fought were achieved—even though in William's reign strong Tory pressures limited the extent of the reform. First and foremost, England became a constitutional monarchy, controlled by Parliament. Second, real advances were made in securing the liberty of subjects in the law courts, in achieving a greater (though far from complete) measure of religious toleration, and in assuring freedom of thought and expression. Locke himself drafted the arguments that his friend Edward Clarke used in the House of Commons in arguing for the repeal of the restrictive Act for the Regulation of Printing. The act was abolished in 1695 and the freedom of the press was secured.

Influence during retirement

Publication of his works. The main task of this last period of his life, however, was the publication of his works, which had been the product of long years of gestation. The *Epistola de Tolerantia* (*A Letter Concerning Toleration,* 1689) was published anonymously at Gouda in 1689. Locke had been reflecting on this topic from his early days at Oxford. Though his correspondence and a paper that he wrote in 1667 show his support for toleration in religion, in 1660–61 he wrote two tracts on this theme (not published until 1967) that are surprisingly conservative. *Two Treatises of Government* (1690) was also the fruit of years of reflection upon the true principles in politics, a reflection resting on Locke's own observations. In all of these social and political issues, Locke saw that the ultimate factor is man's nature. To understand man, however, it is not enough to observe his actions; one must also inquire about his capacities for knowledge. Locke had been conscious of this point in writing his paper on the "Law of Nature" as early as 1663. In 1671, as has been seen, he set out to write a book about human knowledge, *An Essay Concerning Human Understanding,* which was not published, however, until December 1689 (all copies dated 1690)—nor was it wholly completed even then, for Locke made changes, sometimes substantial ones, in three of the four following editions. (See below, *Locke's philosophy.*)

Last years. Locke's last years were spent in the peaceful retreat of Oates. His hostess was a woman with whom he had been acquainted for many years—Lady Masham, or Damaris, the daughter of Ralph Cudworth, one of the Cambridge Platonists, by whom Locke had been significantly influenced. He found friendship and comfort in this household. Many of his friends visited him there: Sir Isaac Newton, who came to discuss the Epistles of St. Paul, a subject of great interest to both; his nephew and heir Peter King, destined to become lord high chancellor of England; and Edward Clarke with his wife and children, for whom Locke had great affection. Locke had written a series of letters to Edward Clarke from Holland, advising him on the best upbringing for his son. These letters formed the basis of his influential *Some Thoughts Concerning Education* (1693), setting forth new ideals in that field. He wrote and published pamphlets on matters of economic interest, on rates of interest, on the coinage of the realm, and, more widely, on trade (defending mercantilist views). In 1695 he published a dignified plea for a less dogmatic Christianity in *The Reasonableness of Christianity.*

Later writings

John Locke died on Oct. 28, 1704, and was buried in the parish church of High Laver. "His death," wrote Lady Masham, "was like his life, truly pious, yet natural, easy and unaffected." This account of his character by one who knew him well seems singularly appropriate. He was orderly, careful about money, occasionally parsimonious, abstemious, and, though naturally emotional and hot-tempered, controlled and disciplined. He had a great love of children, and friendship was for him a necessity. Both in his books and in his life are found the marks of the prudence and wisdom for which he was famed.

LOCKE'S PHILOSOPHY

Theory of knowledge. Locke was thoroughly suspicious of the view that a thinker could work out by reason alone the truth about the universe. Much as he admired Descartes, he feared this speculative spirit in him, and he despised it in the Scholastic philosophers. In this sense he rejected metaphysics. Knowledge of the world could only be gained by experience and reflection on experience, and this knowledge was being gained by Boyle, Sydenham, Christiaan Huygens, and Newton. They were the true philosophers who were advancing knowledge. Locke set himself the humbler task, as he conceived it, of understanding how this knowledge was gained. What was "the original, certainty, and extent of human knowledge, together with the grounds and degrees of belief, opinion, and assent"?

Empiricism. As for "the original," the answer was plain. Knowledge of the world began in sense perception, and self-knowledge in introspection, or "reflection" in Locke's language. It did not begin in innate knowledge of maxims or general principles, and it did not proceed by syllogistic reasoning from such principles. In the 17th century there had been much vague talk about innate knowledge, and in Book I of his *Essay Concerning Human Understanding* Locke examines this talk and shows its worthlessness. In Book II of his *Essay* he begins by claiming that the sources of all knowledge are sense experience and reflection; these are not themselves, however, instances of knowledge in the strict sense, but they provide the mind with the material of knowledge. Locke calls the material so provided "ideas." Ideas are objects "before the mind," not in the sense that they are physical objects, but that they represent them. Locke distinguishes ideas that represent actual qualities of objects (such as size, shape, or weight) from ideas that represent perceived qualities, which do not exist in objects except as they affect observers (such as colour, taste, or smell). Locke designates the former primary qualities and the latter secondary qualities.

Locke proceeds to group and classify the ideas, with a view to showing that the origin of all of them lies in sensation and reflection. Although ideas are immediately "before the mind," not all of them are simple. Many of them are compounded, and their simple parts can be revealed on analysis. It is these simple ideas alone that are given in sensation and reflection. Out of them the mind forms complex ideas, though Locke is ambiguous on this point. For while he uses the language of "forming" or "compounding" and speaks of the "workmanship" of the mind, the compounding is frequently in accordance with what is perceived "to go together" and is not arbitrary.

Simple and complex ideas

Locke's reflections upon cause and effect, had they been elaborated, would undoubtedly have led him into acute difficulties. He does admit one failure. As an empiricist he can give no account of the idea of substance; it is, he thinks, essential and not to be denied, and yet it is not a simple idea given in sensation or reflection nor is it derived from simple ideas so given. In fact he can say little of it; it is "a-something-I-know-not-what." Thus, the case for empiricism cannot be said to be entirely established by Book II, but Locke thinks it strong enough for him to persist in the view that knowledge of the physical world is wholly derived from sense perception.

Self-knowledge. Some ideas are not of things outside

the mind but are reflexive and internal. Locke finds it necessary to classify these in Book II and in doing so sets down the foundations of empirical psychology. His source of information is introspection and rarely the observation of behaviour. His account of sense perception is celebrated for its appreciation of the part that the interpretative mind plays in perceiving, and some of his farsighted observations on the relations between the senses, particularly vision and touch, have profoundly affected subsequent thought. He makes valuable remarks on memory, on discerning, on comparing, on madness, on pleasure and pain, on the emotions, and on the association of ideas.

Locke holds that man has an intuitive knowledge of his own existence and supposes that man exists as material and immaterial substance, but he is none too clear about this and at one point plays with the idea that man is simply material substance to which God has "superadded" a power of thinking. Locke's most valuable contribution, however, is his account of personal identity. Having distinguished between different types of identity, he argues that personal identity depends on self-consciousness (that is, I am the person who did so-and-so 20 years ago because I can remember myself doing it).

Language. According to Locke, Book III, on language, "cost [him] more pains" than any other book of his *Essay;* yet it is the book that has been most neglected. To understand thinking and knowing one must understand language as the means of thought and communication. Words are conventional signs, but signs, according to Locke, not immediately of things but of ideas of things, so that he carries his theory of ideas into his account of language. Frequently, the idea signified by the word is not clear, and sometimes words are used even when there are no ideas corresponding to them. This is particularly so in the case of general words, without which language would be so impoverished as to lose most of its worth. The use of general words, in Locke's mind, is bound up with the theory of universals. Does the general word stand for a particular idea that is used in a representative capacity? Or is the universal nothing more than a creation of the mind, through abstraction, to which is attached a name? In considering natural substances Locke is inclined strongly toward a conceptualism according to which the use of general words is possible only because they signify "nominal essences." In this view what is meant is not the real essence but an abstract concept, something brought about through the "workmanship of the understanding." Locke also discusses the names of simple ideas and of relations, and it is interesting to find the crude beginnings of a discussion of what were later to be called logical or operative words. Book III contains also a valuable account of definition, which denies the theory that all definition must be *per genus et differentiam* (by comparison and contrast). The final chapters deal with the inevitable imperfections of language and with avoidable abuses.

General words

Conclusions. In Book IV, Locke discusses the nature and extent of human knowledge. The tone is more rationalistic than that of the previous books because the skepticism that emanated from his empiricism drove him to find the ideal of knowledge in the indubitable certainties of mathematics. There he was on common ground with the rationalists of his day, and indeed the direct influence of Descartes seems to be observable in the opening chapters of Book IV. Knowledge is perception, not sense perception but intellectual perception or intuition, frequently gained by a deliberate process of demonstration. But even when this is so, each step in the demonstration is observed intuitionally, so that knowledge in the strict sense is essentially intuitive.

Knowledge and intuition

Unfortunately, what can be intuited and demonstrated is limited. Strict knowledge is not confined entirely to mathematics, but the intuition of relations within the physical world is impossible. Books II and III have shown that ideas and nominal essences can be grasped directly and that the inner nature of real things cannot be known, so that "science," in the exact sense of perfectly certain knowledge, is not possible in this sphere. The only possibility of intuiting is that within the world of ideas, an ideal world that is for Locke empirically derived and not intellectual in character. Knowledge in general terms he accordingly defines as the intuition or "perception of the connection of and agreement, or disagreement and repugnancy, of any of our ideas." Within the realm of ideas indubitable knowledge can be gained, but when dealing with ideas "whose archetypes are without them" the position is uncertain.

In spite of this, and somewhat inconsistently, Locke thinks that knowledge approaches certainty in the "sensitive knowledge" of the existence of physical things. Further, knowledge of one's own existence is intuited. In these cases knowledge that is not an apprehension of a relation between ideas is nonetheless certain. But Locke makes it clear that, for the most part, knowledge of the physical world or of oneself is probable and rests not on intuition but on judgment; it is assenting to a proposition on the strength of the evidence, and there may be degrees of assent and wrong assent or error. Locke recognizes the need for a logic of probability, though he does little himself to meet that need. Yet it should be added that the important regular-sequence theory of induction, afterward developed by George Berkeley and David Hume, is put forward in the pages of Locke's *Essay.*

Political theory. Locke's most important work on political philosophy is that entitled *Two Treatises of Government.* The first treatise is a refutation of Sir Robert Filmer's *Patriarcha,* a defense of the divine right of kings that was written in the mid-17th century; the second and more important treatise refutes the absolutist theory of government as such.

Locke defines political power as "a right of making laws, with penalties of death, and consequently all less penalties for the regulating and preserving of property and of employing the force of the community in the execution of such laws, and in the defence of the commonwealth from foreign injury, and all this only for the public good." Government is thus a trust, forfeited by a ruler who fails to secure the public good. The ruler's authority, that is to say, is conditional rather than absolute. Nor does the individual surrender all his rights when he enters a civil society. He has established his right to property by "mixing his labour" with things originally given to mankind in common but now made his own by his labour. (Here in germ is the labour theory of value.) He has the right to expect political power to be used to preserve his property, in his own person and in his possessions, and the right to freedom of thought, speech, and worship. In fact the one right that he gives up in entering a civil society is the right to judge and punish his fellow man, which is his right in the state of nature. He quits his "executive power of the law of Nature" and "resigns it to the public"; he himself makes himself subject to the civil law and finds his freedom in voluntary obedience. To secure this freedom, Locke favoured a mixed constitution—the legislative should be an elected body, whereas the executive is usually a single person, the monarch—and he argues for a separation of legislative and executive powers. The people are ultimately sovereign, though it is not always clear in Locke's theory where the immediate sovereignty lies. But the people always have the right to withdraw their support and overthrow the government if it fails to fulfill its trust.

Freedom through law

Moral philosophy. One searches in vain for a consistent moral theory in Locke. His view that morality can be a science, as certain as mathematics, is well known. This might imply a rationalism, and there are indeed rationalist trends in his moral philosophy—although sometimes when advocating a science of morals he seems to have in mind simply the possibility of an exact analysis of the terms used in moral discourse and the clarification of moral statements. At other times, he puts forward a hedonist theory. "That we call *good* which is apt to cause or increase pleasure or diminish pain in us." But not every good is moral good: "*Moral* good and evil is only the conformity or disagreement of our voluntary actions to some law, whereby good or evil is drawn on us, from the will and power of the law-maker." In this view law rests on God's will, "the true ground of morality," though in saying this Locke does not appear to be consistent with what he says elsewhere of the law of nature.

Theory of education. A good education, as set forth by Locke in *Some Thoughts Concerning Education,* attends to both the physical and the mental. The body is not to be coddled; on the contrary, it is necessary that it should be hardened in various ways. The good educator insists on exercise, play, and plentiful sleep, "the great cordial of nature." Young children should be allowed to give vent to their feelings and should be restrained rarely. As for mental training, character comes first before learning; the educator's aim is to instill virtue, wisdom, and good breeding into the mind of the young. Parents, too, must interest themselves in their children's upbringing and, as far as possible, have them near; for no educative force is more powerful than the good example of parents. A stock of useful knowledge must be imparted: modern languages and Latin; geography and history; mathematics, as "the powers of abstraction develop"; and later civil law, philosophy, and natural science. For recreation, training in the arts, crafts, and useful hobbies should be available.

Religion. Locke's reaction against the "enthusiasm" of the sects in his youth had been sharp, and he disliked religious fanaticism throughout his life. He was a broad, tolerant Anglican anxious to heal the breach in English Protestant ranks. His own views on church government and on the priesthood were close to those of the dissenters, and he favoured the liberal views of the latitudinarians, of the Cambridge Platonists, and of the Remonstrants of Holland. This becomes manifest in *The Reasonableness of Christianity.* Two essentials, and two alone, he thinks, are involved in being a Christian: first, that a man should accept Christ as God's Messiah and, second, that he should live in accordance with Christ's teaching. His point of view is not far removed from that of the Deists on the one hand and the Unitarians on the other, yet he cannot be grouped with them. Christianity, though reasonable, needs revelation as well as reason, for human reason alone is inadequate: there is an experience of God "through His Spirit" without which all religion is empty. However, any act of persecution in the name of religious truth is wholly unjustified, since our knowledge and understanding are so confined. Each individual is a moral being, responsible before God, and this presupposes freedom. By the same token, no compulsion that is contrary to the will of the individual can secure more than an outward conformity.

Reason and revelation

Influence. Locke's faith in the salutary, ennobling powers of knowledge justifies his reputation as the first philosopher of the Enlightenment. In a broader context, he founded a tradition of thought that would span three centuries, in the schools of British empiricism and American pragmatism. In developing the Whig ideology underlying the Exclusion Controversy and the Revolution of 1688, Locke formulated the classic expression of liberalism, which was to inspire both the shapers of the American Revolution and the authors of the U.S. Constitution. Locke's influence remained strongly felt in the West in the 20th century, as notions of mind, freedom, and authority continued to be challenged and explored.

MAJOR WORKS

PHILOSOPHY, RELIGION, AND EDUCATION: *An Essay Concerning Humane* [sic] *Understanding* (1690); *Epistola de Tolerantia* (1689; *A Letter Concerning Toleration,* trans. by William Popple, 1689); *A Second Letter Concerning Toleration* (1690); *A Third Letter for Toleration* (1692); *Some Thoughts Concerning Education* (1693); *The Reasonableness of Christianity, as Delivered in the Scriptures* (1695); *A Vindication of the Reasonableness of Christianity* (1695); *A Second Vindication of the Reasonableness of Christianity* (1697); *Of the Conduct of the Understanding,* in *Posthumous Works of Mr. John Locke* (1706).

POLITICAL PHILOSOPHY AND ECONOMICS: *Two Treatises of Government* (1690); *Some Considerations of the Consequences of the Lowering of Interest, and Raising the Value of Money* (1692); *Short Observations on a Printed Paper, Intituled, for Encouraging the Coining Silver Money in England, and After for Keeping It Here* (1695); *Further Considerations Concerning Raising the Value of Money* (1695).

RECOMMENDED EDITIONS: A complete edition is *The Works of John Locke,* new ed. corrected, 10 vol. (1823, reprinted 1963). There is no complete modern edition of Locke's works, although several volumes have appeared in the Oxford Press series, "The Clarendon Edition of the Works of John Locke"; the first of these was a critical edition of *An Essay Concerning Human Understanding,* edited by Peter H. Nidditch (1975, reprinted 1979). Useful editions of other single works include *A Letter Concerning Toleration,* edited by James Tully (1983); *Two Treatises of Government,* edited by Peter Laslett, 2nd ed. (1967, reprinted 1970), a critical edition; and *The Educational Writings of John Locke: A Critical Edition with Introduction and Notes,* edited by James L. Axtell (1968).

BIBLIOGRAPHY

Biographies: Early works include LORD KING, *The Life and Letters of John Locke,* new ed. (1858, reissued 1984), an amateurish work but based on the Lovelace Collection of Locke papers in the possession of Peter King's family; and H.R. FOX BOURNE, *The Life of John Locke,* 2 vol. (1876, reprinted 1969), a detailed study, based on secondary sources. MAURICE W. CRANSTON, *John Locke: A Biography* (1957, reissued 1985), is now the standard biography. An outstanding resource is E.S. DE BEER (ed.), *The Correspondence of John Locke* (1976–), part of "The Clarendon Edition of the Works of John Locke"; 7 of 8 vol. have appeared to 1986.

Commentaries: JOHN W. YOLTON, *Locke: An Introduction* (1985); and JOHN DUNN, *Locke* (1984), provide general accounts of Locke's life and work. For Locke's theory of knowledge, see R.S. WOOLHOUSE, *Locke* (1983); and JAMES GIBSON, *Locke's Theory of Knowledge and Its Historical Relations* (1917, reprinted 1968), another useful introductory essay, if somewhat old-fashioned in its approach. For a survey of Locke's thought, see RICHARD I. AARON, *John Locke,* 3rd ed. (1971, reprinted 1973); D.J. O'CONNOR, *John Locke* (1952, reissued 1967); and JOHN W. YOLTON, *John Locke and the Way of Ideas* (1956, reprinted 1968), a study based on Locke's unpublished as well as his published writings.

Specialized commentaries on Locke's epistemology are found in JOHN W. YOLTON, *Locke and the Compass of Human Understanding: A Selective Commentary on the "Essay"* (1970); J.L. MACKIE, *Problems from Locke* (1976); and I.C. TIPTON (ed.), *Locke on Human Understanding: Selected Essays* (1977). Political theory is covered in STERLING POWER LAMPRECHT, *The Moral and Political Philosophy of John Locke* (1918, reprinted 1962); GERAINT PARRY, *John Locke* (1978); J.W. GOUGH, *John Locke's Political Philosophy: Eight Studies,* 2nd ed. (1973); and M. SELIGER, *The Liberal Politics of John Locke* (1968), an exposition and a defense of Locke's arguments for political freedom. W. VON LEYDEN, *Hobbes and Locke: The Politics of Freedom and Obligation* (1981); RICHARD H. COX, *Locke on War and Peace* (1960, reprinted 1982); and C.B. MACPHERSON, *The Political Theory of Possessive Individualism: Hobbes to Locke* (1962, reprinted 1983), explore the relationship between Locke's political thought and that of Thomas Hobbes. See also JOHN DUNN, *The Political Thought of John Locke: An Historical Account of the Argument of the "Two Treatises of Government"* (1969, reprinted 1982), a survey of Locke's thought in the context of his intellectual environment; and RAYMOND POLIN, *La Politique morale de John Locke* (1960, reprinted 1984), on Locke's liberalism from the perspective of a French historian of ideas. JAMES TULLY, *A Discourse on Property: John Locke and His Adversaries* (1980, reissued 1982); GORDON J. SCHOCHET, *Life, Liberty and Property: Essays on Locke's Political Ideas* (1971); and J.G.A. POCOCK and RICHARD ASHCRAFT, *John Locke* (1980), discuss Locke's defense of the natural right to property. See also KAREN IVERSEN VAUGHN, *John Locke: Economist and Social Scientist* (1980), for Locke's ideas on economics; and KENNETH DEWHURST, *John Locke, 1631–1704, Physician and Philosopher* (1963, reprinted 1984), on his career as a practitioner and theorist of medical science. Research in progress, queries, and corrections to published work on Locke are reported in *The Locke Newsletter* (annual).

Bibliographies: H.O. CHRISTOPHERSEN, *A Bibliographical Introduction to the Study of John Locke* (1930, reprinted 1968), is still useful, although the whole content has been assimilated into a larger, more recent work, JEAN S. YOLTON and JOHN W. YOLTON, *John Locke: A Reference Guide* (1985)—both cover mainly secondary sources. A useful companion to these is JOHN C. ATTIG (comp.), *The Works of John Locke: A Comprehensive Bibliography from the Seventeenth Century to the Present* (1985), tracking the various editions and translations of Locke's writings and placing them in historical context. See also ROLAND HALL and R.S. WOOLHOUSE, *80 Years of Locke Scholarship: A Bibliographical Guide* (1983); and P. LONG, *A Summary Catalogue of the Lovelace Collection of the Papers of John Locke in the Bodleian Library* (1959), a valuable guide to the most important source of manuscript material.

(R.Aa./Ed.)

The History and Kinds of Logic

Traditionally, logicians have distinguished between deductive logic, whose principles are used in drawing new propositions out of premises in which they lie latent, and inductive logic, which ventures conclusions from particular facts that appear to serve as evidence for them. But this division is obsolete, because the problems earlier subsumed under induction are now apportioned to the methodology of the natural sciences. Thus logic is here taken to mean deductive logic with all of its ramifications.

This article traces the development of deductive logic from its origins among the ancient Greeks to the complex systems of the 20th century. Although much of the article is devoted to logic in Western thought, some attention is given to its manifestations in Indian and Chinese writings. Other points of coverage include the principal systems of deductive logic and their place among such disciplines as mathematics and philosophy. For details on the relation of logic to mathematics, see PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE: *Foundations of mathematics.* Another section of the same article deals with the issues and developments in the philosophy of logic.

The present article is divided into the following sections:

History of logic 235
- Ancient logic 235
  - Precursors of ancient logic
  - Aristotle and the logic of noun expressions
  - Theophrastus and the logic of noun expressions
  - Founding of the logic of propositions
- Medieval logic 238
  - Development of medieval logic
  - Theories of language
  - Formal logic
- Logic in the East 240
  - Indian logic
  - Chinese logic
- Modern logic 242
  - Logic in the Renaissance and in Humanism
  - Rise of mathematical logic during the Enlightenment
  - Development of mathematical logic in the 19th century
- 20th-century logic 247
  - Logicism, intuitionism, and formalism
  - Logic of propositions
  - Logic of noun expressions
  - Metalogical studies
- Role of logic as an element within a culture 250

Logic systems 250
- Formal logic 250
  - General observations
  - The propositional calculus
  - The predicate calculus
  - Syllogistic
  - Modal logic
  - Set theory
  - Logical foundations of mathematics
- Metalogic 272
  - Nature, origins, and influences of metalogic
  - Nature of a formal system and of its formal language
  - Discoveries about formal mathematical systems
  - Discoveries about logical calculi
  - Model theory
- Applied logic 279
  - The critique of forms of reasoning
  - Epistemic logic
  - Practical logic
  - Logics of physical application
  - Hypothetical reasoning and counterfactual conditions

Divisions and nature of deductive logic. In its narrower sense deductive logic divides into (1) the logic of propositions (also called sentential logic) and (2) the logic of noun expressions. The logic of propositions or sentences is so named because, in it, propositions or sentences form the only basic semantic category (or part of speech). Some of them are simple and remain unanalyzed, others are compound and are analyzed into propositional connectives ("if . . . then," "and," "or," "it is not the case that," . . .) and their arguments, which themselves are propositions. In logical systems they are represented by *p, q, r,* . . . . In the logic of noun expressions simple propositions are broken down into their parts, which comprise a second basic semantic category; namely, the category of names ("Socrates," "the father of Socrates") or the category of common nouns ("philosopher," "Athenian citizen"). In simple propositions names often occur as arguments in concatenation with predicates; that is to say, verbs ("moves," "lives") or verblike expressions ("is wise," "is a philosopher"). The theory whose vocabulary comprises propositional connectives, names (represented by *a, b, c,* . . . or *x, y, z* . . .), predicates (represented by *F, G, H,* . . .), and the quantifiers ("for all . . ." and "for some . . .") is known as the logic of predicates or lower functional calculus or logic of quantification. Verbs or verblike expressions that require one name to form a proposition are called one-place predicates. There are also two-place predicates, three-place predicates, etc.; *i.e.,* verbs or verblike expressions that require two or three or more names to form a proposition ("*a* is bigger than *b,*" "*a* gives *b* to *c*"). It is the task of the logic of relations, another part of the logic of noun expressions, to exhibit the use of such predicates. Neither the logic of predicates nor the logic of relations accommodates common nouns as a semantic category in its own right. Both treat common nouns as integral parts of predicates. Common nouns, however, can be given the status of a basic semantic category, as was, in fact, done in Aristotle's syllogistic and in that part of the logic of noun expressions that is known as the logic of classes.

Logic of propositions and logic of noun expressions

In a wider sense, logic comprises, in addition to the above, various theories of language—such as logical syntax and semantics. Systems of the logic of propositions as well as systems of the logic of noun expressions can become, in turn, objects of special studies that are part of what is now known as metalogic or the methodology of deductive systems, a logic of logics, as it were. Theories of modalities—*i.e.,* theories that are concerned with the notions of necessity, impossibility, possibility, and contingency—fall also within the scope of logic in its wider sense. Finally, the crucial role of logical fallacies and paradoxes cannot be overlooked.

Direct method: theses and expressions

As already mentioned, the principles or theses of the logic of propositions exhibit directly the use of such expressions as "if . . . then," "it is not the case that," "and," and "or," employed in concatenation with propositions, *p, q, r.* The principle of the hypothetical syllogism, which says that

> For all *p, q, r,* if [(if *p* then *q*) and (if *q* then *r*)] then (if *p* then *r*),

is an example of a thesis belonging to the logic of propositions. The theses of the logic of noun expressions exhibit the use of such constant terms as "is the same object as," "is," "every . . . is," "some . . . is," "no . . . is," and "some . . . is not," to mention only a few. In concatenation with noun expressions, *a, b, c,* these terms form propositions. Examples of statements belonging to the logic of noun expressions are the law that asserts the symmetry of identity:

For all *a* and *b*, if *a* is the same object as *b* then *b* is the same object as *a*

and the Aristotelian syllogism (called *Darii*):

For all *a*, *b*, and *c*, if every *b* is *c* and some *a* is *b* then some *a* is *c*.

It should be noted that, in the theses of the logic of noun expressions, a vocabulary belonging primarily to the logic of propositions is employed, which reflects the fact that the logic of propositions is logically prior to that of noun expressions.

Some logicians have wondered whether there are any logical theories that involve the use of three or more basic semantic categories; and the answer is "yes," though so far such theories have not been worked out in detail.

Alternative formats employed in presenting logical principles. The foregoing examples of logical theses have been formulated with the aid of propositional variables, *p, q, r,* and with the aid of nominal variables, *a, b, c,* which stand for noun expressions. This formulation, first used by Aristotle, may be described as (1) the direct method of expressing logical principles. But Aristotle himself and many of his ancient and medieval followers often referred to logical principles also by means of (2) standard examples. For instance, the law of the hypothetical syllogism could be represented by the following example, adapted from a Greek commentator to Aristotle's *Analytica priora:*

Standard examples, metalogical terms, and schemata

If [(if a man exists then an animal exists) and (if an animal exists then a substance exists)] then (if a man exists then a substance exists).

Instead of expressing the law of the hypothetical syllogism directly or illustrating it with the aid of a standard example, it can be described in (3) metalogical terms:

For all *a, b,* and *c,* if (*a* implies *b* and *b* implies *c*) then *a* implies *c*. (Example: If [the universal affirmation (*a*) implies the corresponding particular affirmation (*b*), and the particular affirmation (*b*) implies the negation of the corresponding universal denial (*c*)], then the universal affirmation (*a*) implies the negation of the corresponding universal denial (*c*). For terminology see below *Theory of opposition.*)

Here the variables *a, b,* and *c* are nominal instead of propositional variables, and the term "implies"—which is not the same as the connective "if . . . then," because it requires noun expressions, and not propositions, as arguments—belongs to the vocabulary of metalogic. Many logical principles give rise to valid inferences, and some logicians have preferred using (4) schemata of valid inferences in presenting their systems of logic. The schema

If *p* then *q*; but if *q* then *r*; therefore if *p* then *r*

gives the form of inferences the validity of which is guaranteed by the law of the hypothetical syllogism. A less abstract way of referring to logical principles consists in making use of (5) standard inferences. Thus, the inference

If a man exists then an animal exists; but if an animal exists then a substance exists; therefore if a man exists then a substance exists

if used in the context of a logical discussion, can be understood as representing the law of the hypothetical syllogism. Note that standard examples that illustrate logical principles are propositions, whereas standard inferences are sets of propositions. In an inference one proposition, usually indicated by the preceding "therefore," is the conclusion; the remaining propositions are called premises. These five different ways of referring to logical principles must be kept in mind when tracing their gradual emergence in the course of the development of logic.

HISTORY OF LOGIC

Ancient logic

PRECURSORS OF ANCIENT LOGIC

Though in theory the logic of propositions is logically prior to the logic of noun expressions, it was the latter that historically preceded the former. The first system of the logic of noun expressions was worked out by Aristotle in the 4th century BC. Aristotle was right when he claimed that in creating syllogistic, which is a part of the logic of noun expressions that can be formulated in terms of categorical propositions, he was a pioneer—though preparatory work of considerable extent and significance had been done for him already by earlier Greek mathematicians, rhetoricians, and philosophers. The Greeks were, in fact, almost unique in their practice, cultivation, and enjoyment of argument and proof; and without interest in the art of argumentation one cannot have an intellectual climate conducive to the development of logical theories. The ability, however, to produce valid inferences in the course of an argument does not necessarily presuppose any knowledge of logic. Indeed, the preserved speeches of ancient Attic orators and the dialogues of Plato contain numerous valid inferences, but the logical laws that account for their validity were unknown at the time.

It was mentioned earlier that in its wider sense logic comprises theories of language. In this field, preparatory work had been done by the Sophists, a class of private tutors in ancient Greece, who were the first to distinguish different types of sentences—affirmations, denials, questions, answers, and injunctions—and who were also interested in the study of synonyms. Further progress in theories of language was later made in the Academy, the school that Plato founded. In the *Sophist,* written in the 4th century BC, Plato remarked that neither a series of noun expressions alone nor a series of verbs alone can produce a proposition, a principle that is accepted even by logicians today, who have generalized it to read that no series of expressions belonging to one and the same semantic category can form a meaningful expression. More important for the development of logic, however, was Plato's conception of the axiomatic method, which lays down principles (usually self-evident) and deduces theorems from them, as an ideal means of systematizing knowledge. It was not until Aristotle and Euclid (*c.* 300 BC) that Plato's program was successfully turned into concrete achievement.

Theories of language of the Sophists and Plato

In discussing the prehistory of logic, mention should be made of fallacious arguments and paradoxes. These are arguments that, from premises that appear to be true, lead, by steps that appear to be valid, to a conclusion that appears to be false. The Sophists used to teach their pupils how to argue for and also against any propounded statement. This procedure, of course, gave occasion to fallacious arguments, which called for skillful refutation. In the 5th and 4th centuries BC, the interest in fallacious arguments was widespread, as is evidenced by Plato's *Euthydemus* and Aristotle's *De sophisticis elenchis.* The most famous of the early paradoxes was perhaps the "Achilles," invented by Zeno of Elea, an ancient Rationalist. Zeno argued that Achilles, reputed to be the fastest runner of all, could not overtake a tortoise in a race once the tortoise was allowed to get in front of him. For before overtaking the tortoise, Achilles would have to cover the distance that separated him from the tortoise at the time t_1. This he could accomplish at the time t_2; but at that time the tortoise would have moved away and Achilles would still be separated from it. The argument can then be repeated for the distance between them at the time t_2 and so on, pointing apparently to the conclusion that Achilles would never overtake the tortoise.

In a system of logic some of the theses are described as definitions. In introducing definitions, there are a number of conditions that must be carefully satisfied if they are to be used without endangering the consistency of the system. Though rules for thus introducing correct definitions into systems of logic are of recent origin, the credit for being the first to practice the art of definition should, according to Aristotle, be given to Socrates.

ARISTOTLE AND THE LOGIC OF NOUN EXPRESSIONS

The collection of Aristotle's logical treatises is known as the *Organon,* a title invented probably by an ancient editor of his writings, which reflects the view that logic is a

The *Organon*

tool for sharpening thought. It consists of the *Categoriae,* the *De interpretatione,* the *Analytica priora* (two books), the *Analytica posteriora* (two books), the *Topica* (eight books), and the *De sophisticis elenchis.* The order of the treatises is supposed to be systematic; it does not seem to correspond to the chronological order, however. From the point of view of logic, the *De interpretatione* and the *Analytica priora* are by far the most important.

Structure of language. In Aristotle, the discussion of the syntactical and semantical features of language constituted a preamble to his syllogistic. Following Plato, he distinguished between nouns and verbs and regarded them as essential parts of a proposition; but he seemed to concentrate his attention on nouns and, in particular, on referential common nouns with limited generality. He did so because he believed that nonreferential or empty noun expressions, like "goat-stag," singular nouns, and the nouns of ultimate generality, such as "thing" or "object," are of lesser importance for scientific inquiries. Noun expressions, which in Aristotle included adjectives, are used for the purpose of constructing simple affirmations and denials of the following four types:

1. Singular: *e.g.,* Socrates is white—Socrates is not white;
2. Universal: *e.g.,* Every man is white—No man is white;
3. Particular: *e.g.,* Some man is white—Some man is not white;
4. Indefinite: *e.g.,* Man is white—Man is not white.

In Aristotelian syllogistic, use is made of universal and particular propositions only. Called categorical propositions, these involve the following constant vocabulary: "Every . . . is," "No . . . is," "Some . . . is," and "Some . . . is not," which form propositions that were designated by medieval logicians as A, E, I, and O, respectively.

Theory of opposition. Syllogistic comprises (1) the theory of opposition, (2) the theory of conversion, and (3) the syllogistic proper. The theory of opposition, expounded in the *De interpretatione,* considers the logical relations that obtain between the four types of propositions formed by the affirmation and denial of universal and particular propositions. It stipulates that a universal affirmation, "Every *a* is *b,*" and the corresponding particular denial, "Some *a* is not *b,*" are contradictories; *i.e.,* that one of the pair must be true whereas the other must be false. A universal denial and the corresponding particular affirmation are also contradictories. A universal affirmation, "Every *a* is *b,*" however, and the corresponding universal denial, "No *a* is *b,*" form a pair of contrary propositions, which cannot both be true though they can both be false. A particular affirmation and the corresponding particular denial cannot both be false, though Aristotle did not state this fact. Finally, either of the two universal propositions implies the corresponding particular proposition (its "subaltern"), a fact that, judging from a passage in the *Topica,* was known to Aristotle. In the *De interpretatione,* the theory of opposition was stated in metalogical terms. Using the direct method, it is now possible to express the relevant logical principles as follows:

Contradictories, contraries, and subalterns

For all *a* and *b,* every *a* is *b* if and only if it is not the case that some *a* is not *b.*
For all *a* and *b,* no *a* is *b* if and only if it is not the case that some *a* is *b.*
For all *a* and *b,* if every *a* is *b,* then it is not the case that no *a* is *b.*
For all *a* and *b,* if it is not the case that some *a* is *b,* then some *a* is not *b.*
For all *a* and *b,* if every *a* is *b,* then some *a* is *b.*
For all *a* and *b,* if no *a* is *b,* then some *a* is not *b.*

Theory of conversion. Discussed in the *Analytica priora,* the theory of conversion considers the logical relationship that holds between a categorical proposition and its converse, obtained by reversing the order of the noun expressions *a* and *b,* and states in which cases such a converse can be inferred from the original proposition. The logical principles involved are not only described in metalogical terms and illustrated with the aid of standard examples but also are expressed directly. They are:

For all *a* and *b,* if every *a* is *b,* then some *b* is *a.*
For all *a* and *b,* if some *a* is *b,* then some *b* is *a.*
For all *a* and *b,* if no *a* is *b,* then no *b* is *a.*
For all *a* and *b,* if no *a* is *b,* then some *b* is not *a.*

This last principle, however, was not mentioned by Aristotle and has been disregarded by his successors.

Syllogistic proper. The *Analytica priora* contains also a systematic exposition of syllogistic proper. By a syllogism Aristotle understood a "propositional expression [*logos*] in which certain things having been laid down, something other than what has been laid down follows of necessity from their being so." This definition, however, is far too general; and it seems convenient for the present purposes to narrow it to apply to any inference for which the premises and conclusion are all categorical propositions—*i.e.,* simple propositions, rather than "if . . . then," "either . . . or," or "both . . . and" propositions—or to a logical principle that lends validity to such an inference. In Book I, chapters 4–7, Aristotle examined syllogisms with two premises. Such syllogisms, construed as logical principles, are all of the form

If α and β, then γ

in which the Greek letters alpha, beta, and gamma stand for categorical propositions with variable noun expressions called "terms" (see below *Syllogistic*). In syllogisms of this form, Aristotle distinguished: (1) the "major" term, which is the predicate in the conclusion and is assumed to occur in one of the premises; (2) the "minor" term, which is the subject in the conclusion and is assumed to occur in the other premise; and (3) the "middle" term, which occurs in both premises but does not occur in the conclusion. Except for the definition of the middle term, which was employed by Aristotle, these definitions of the major and minor terms appear to have been introduced by John Philoponus, a 6th-century Christian theologian and Aristotelian commentator. In an Aristotelian syllogism the premises are also referred to as "the major" and "the minor," depending on which term (the major or the minor) they contain. Aristotle, in his exposition of syllogistic, was concerned to find out which pairs of categorical premises give rise to a valid syllogism. To systematize his search, he distinguished what he called the "three figures" of the syllogism, which he defined according to the placement of the middle term. First he examined those pairs of categorical propositions in which the middle term is the predicate in the minor premise and the subject in the major. Such combinations of premises are said to constitute the first syllogistic figure. Aristotle found that, in this figure, four different pairs of premises yield syllogistic principles. They are set out below together with their medieval names, the vowels of which indicate the kinds of propositions involved, whether A, E, I, or O (as earlier defined), and their sequence. The vowels of syllogism *Ferio,* for example, being EIO, indicate that the major (or first) premise is an E proposition ("No . . . is"), the minor (or second) premise is an I proposition ("Some . . . is"), and the conclusion is an O proposition ("Some . . . is not").

The three terms and three figures of the syllogism

Barbara:	For all *a, b,* and *c,* if every *b* is *c* and every *a* is *b,* then every *a* is *c.*
Celarent:	For all *a, b,* and *c,* if no *b* is *c* and every *a* is *b,* then no *a* is *c.*
Daril:	For all *a, b,* and *c,* if every *b* is *c* anc some *a* is *b,* then some *a* is *c.*
Ferio:	For all *a, b,* and *c,* if no *b* is *c* and some *a* is *b,* then some *a* is not *c.*

With the aid of appropriate counterexamples, Aristotle showed that the remaining combinations of premises in this figure do not yield valid syllogisms. The second figure is defined as the one in which the middle term occurs in the premises as the predicate. In this figure, too, four combinations of premises were found by Aristotle to give rise to syllogisms (*Cesare, Camestres, Festino, Baroco*). In the third figure, the middle term occurs in the premises as the subject. In this case six combinations of premises prove to be syllogistic (*Darapti, Disamis, Datisi, Felapton, Bocardo, Ferison*). The syllogisms of the second and the third figures are not stated here explicitly since they can easily be reconstructed from their medieval names. As was described in the case of the first figure, the combinations of premises that, in the second and the third figures, do not produce valid syllogisms were shown by Aristotle to fail by means of counterexamples.

Important features of Aristotle's syllogistic

The most important features of Aristotle's presentation of syllogistic are the following: First, he gave his syllogisms the form of direct principles of logic and not the form of inferences or inference schemata. This he could do because, for the first time in the history of logic, he had thought of making use of variables, or letters that can stand for any noun expression. Second, he deduced his syllogisms of the second and third figures from the syllogisms of the first figure, and by doing so he became the author of the first deductive system in the history of logic. In his deductions, Aristotle made explicit use of the theories of opposition and conversion. If his proofs were to be complete, however, he would have had to use a number of theses from the logic of propositions; but of this he was not aware.

In his syllogistic, Aristotle did not mention the combination of premises in which the major term is the subject and the minor is the predicate. Five such combinations yield valid syllogisms (*Bramantip, Camenes, Dimaris, Fesapo, Fresison*). They constitute a fourth figure, acknowledged by medieval logicians and the traditional logicians of modern times. To be sure, the syllogisms were known to Aristotle; but they were not included in his account of syllogistic. This fault was rectified by his pupil Theophrastus, who redefined the first figure as the one in which the middle term occurs as the subject in one premise and as the predicate in the other, thus making it possible for the five syllogisms in question to be incorporated into the system. A further expansion of syllogistic was apparently effected by Ariston of Alexandria (1st century BC), a Peripatetic (member of Aristotle's school), who is said to have included in the system the "subaltern" syllogisms; *i.e.,* the syllogisms that result from taking a syllogism of a form already recognized and replacing the universal conclusion by the corresponding particular conclusion (*Barbari, Celaront, Cesaro, Camestrop, Camenop*).

Modal logic. In addition to constructing the first system, however modest, of the logic of noun expressions, Aristotle initiated the development of modal logic, which deals with the notions of necessity, impossibility, possibility, and contingency. Not only did he study their application to syllogistic, but he also arrived at some more general principles. He seems to have abandoned his earlier view that a necessary proposition cannot be said to be possible in favour of the following principles:

Relation of possibility to necessity

1. A proposition is necessary if and only if its negation is not possible;
2. A proposition is possible if and only if its negation is not necessary;
3. A proposition is contingent if and only if it is neither necessary nor impossible.

Aristotle's modal syllogistic contains, however, a number of controversial ideas, which were widely discussed in his school, the Lyceum. Unfortunately, the treatise *On the Disagreement Concerning Mixed Moods Between Aristotle and his Friends* by the expositor of Aristotle, Alexander of Aphrodisias (AD 200), which could have thrown light on the controversy, has not survived. Even today modal logic is a subject of controversy, logicians being divided on the fundamental issue of the nature of modal principles. Some include them within the logic of propositions; others assign them a place in metalogic.

Aristotle was not aware that his syllogistic presupposes a more general logical theory, viz., the logic of propositions. To be sure, he mentioned "syllogisms from hypothesis," by which he may have meant some principles belonging to that logic; but he gave no details. As will be seen, there are firmer grounds for associating the origin of the logic of propositions with Theophrastus; with the Megarians, a 4th-century school of Socratic dialecticians and logicians; and with the Stoics.

THEOPHRASTUS AND THE LOGIC OF NOUN EXPRESSIONS

The writings of Theophrastus on logic are now lost; but evidence from secondary sources shows that he made creditable improvements in the systematization of syllogistic. He suggested, for instance, that the indefinite proposition, which asserts, for example, that "Man is white," should be treated as if it were particular (referring to "some men"), and he redefined the first figure and included in it five syllogisms originally omitted by Aristotle. Within the modal syllogistic he abandoned the notion of contingency. For the remaining modal notions, he introduced the principle of "the weaker premise," which regards a true proposition as weaker than a necessary one (since a proposition can be true, yet not necessary—but not vice versa) and a possible proposition as weaker than a true one, and demands that the conclusion in a valid syllogism should not be stronger, in respect of modality, than its weakest premise.

Prosleptic syllogisms

Theophrastus' theory of prosleptic syllogisms is perhaps his most interesting contribution to the logic of noun expressions. By prosleptic propositions he understood expressions of the form

For all x, if $F(x)$ then $G(x)$

in which both $F(x)$ and $G(x)$ stand for categorical propositions in each of which x occurs as one of the terms. (Example: "For all x, if every man is x, then every x is substance.") It should be noted that x is here a variable term bound by a universal quantifier, the expression "For all x." It was described by Theophrastus as indefinite and called the middle term. Depending on the position of the middle term, conceived in this way, prosleptic propositions can be divided into three figures. In combination with an appropriate categorical premise, a prosleptic premise yields a prosleptic syllogism, as exemplified in the following inference schema:

For all x, if every x is a, then no x is b;
but every c is a; therefore no c is b.

Some prosleptic propositions are equivalent to categorical propositions (Example: "For all x, if every x is a, then every x is b" is equivalent to "every a is b"), but this cannot be said about all of them, as Galen claimed that it could. For instance, no categorical proposition with a and b as arguments is equivalent to the prosleptic proposition which says that

For all x, if some a is not x, then every x is b.

It was Galen (died AD 199), however, one of the two greatest physicians of antiquity, who made the last major contribution to the logic of noun expressions in antiquity by developing a theory of compound syllogisms, occurring in four figures, in which two middle terms appear. Syllogisms of this sort can be illustrated as follows:

Every man is an animal; every man is rational; no irrational being is rational; therefore, some animal is not an irrational being.

FOUNDING OF THE LOGIC OF PROPOSITIONS

The logic of Theophrastus and that of the Megarians

The earliest consideration of the logic of propositions occurred in the works of Theophrastus, of the Megarians, and of the Stoics. Theophrastus, according to Alexander of Aphrodisias, worked out a theory of totally hypothetical syllogisms, which, following the example of Aristotle, he neatly set out in three figures:

1st figure,	If p then q; if q then r; therefore if p then r.
or	If p then q; if q then r; therefore if not-r then not-p.
2nd figure,	If p then q; if not-p then r; therefore if not-q then r.
or	If p then q; if not-p then r; therefore if not-r then q.
3rd figure,	If p then r; if q then not-r; therefore if p then not-q.
or	If p then r; if q then not-r; therefore if q then not-p.

In these inference schemata, the letters p, q, and r represent complete propositions; and for Theophrastus they probably represented categorical ones. But in fact, the nature of the propositions in this context is irrelevant to the validity of the inferences involved. By grouping the hypothetical syllogisms into three figures, Theophrastus tried to assimilate them to the Aristotelian categorical syllogisms of the predicate logic. He does not seem to have realized that hypothetical syllogisms belong to a different logical theory, to the logic of propositions and not to the logic of noun expressions. This fact, however, gradually became clear to the logicians in the Megarian school and in the Stoa (Stoic school).

The Megarian school was made famous by Eubulides (*c.*

4th century BC), a critic of Aristotle, by the dialectician Diodorus Cronus (died *c.* 307 BC), and by Philo of Megara, a pupil of Diodorus. To Eubulides logicians are indebted for certain paradoxes, including the celebrated "liar paradox" ("Does a man who says that he is now lying speak truly?"). Diodorus tried to connect modal notions, such as those of necessity and possibility, with the notions of past, present, and future; and Philo was the first to work out the truth-functional conception of implication; that is to say, of "if . . . then" propositions. (A compound proposition is said to be conceived truth-functionally if its truth or falsity is taken to depend exclusively on the truth or falsity of the propositions of which it is compounded.) Now, Philo argued, against Diodorus, that an implication is true if and only if its "if" proposition, known as the antecedent, is false or its "then" proposition, or consequent, is true; and that it is false if and only if its antecedent is true and its consequent is false.

Stoic logic. It is not known whether the truth-functional interpretation of the rest of the vocabulary of the logic of propositions is also attributable to Philo; but it is known that implication, conjunction ("and" propositions), exclusive disjunction ("either . . . or" propositions), and negation ("it-is-not-the-case-that" propositions) were all interpreted truth-functionally by Chrysippus of Soli (died *c.* 207 BC), the second founder of the Stoa. The truth-functional vocabulary was used by Chrysippus to formulate various logical principles. As a rule, these principles were represented by means of standard inferences or by inference schemata. The following schemata are said to have been set out by Chrysippus as indemonstrable:

The inference schemata of Chrysippus

1. If the first then the second; but the first; therefore the second;
2. If the first then the second; but not the second; therefore not the first;
3. Not both the first and the second; but the first; therefore not the second;
4. Either the first or the second; but the first; therefore not the second;
5. Either the first or the second; but not the second; therefore the first.

Various other inference schemata can be derived from these five in accordance with certain rules or directives. Two such rules, which can be reconstructed from preserved fragments, can be stated in terms of variables as follows:

A. For all *a, b, c,* and *d,* if *b* together with *c* implies *d* and *a* imply *b* then *a* together with *c* implies *d.*
B. For all *a, b,* and *c,* if *a* together with *b* implies *c,* then *a* together with the negation of *c* implies the negation of *b.*

By applying rule A—which, incidentally, was known to the Peripatetics and used by them for the purpose of constructing compound syllogisms—the Stoics could derive the following inference schemata:

6. If the first then if the first then the second; but the first; therefore the second;
7. If the first and the second then the third; however, not the third; but the first; therefore not the second.

It is clear from a moment's inspection that, in an appropriate context, schema 7 could replace rule B; and it is interesting to note that, according to Alexander of Aphrodisias, rule B results from dividing a rule that, in fact, was the same as rule A. From the modern point of view, rules A and B belong to a branch of metalogic.

Comparison of the Lyceum and the Stoa

In matters of logic there was little collaboration between the Lyceum and the Stoa. There was even some effort on the part of both schools to create quite artificial differences; they developed, for instance, two different logical terminologies. For the Stoics, logic was a part of philosophy, whereas for the Peripatetics it was a tool. Aristotle presented his syllogisms in the form of direct logical principles, whereas the Stoics preferred to make use of inference schemata. The Stoics knew, however, that inferences owe their validity to direct logical principles, and they had a simple method of converting schemata into principles. Aristotle and his followers used letters as variables; the Stoics used numerals. In their logical practice, the Peripatetics allowed for a considerable amount of linguistic flexibility, whereas the Stoics became notorious for the formalistic approach, which demanded that expressions should have the same form if they were to have the same meaning. The most important difference, however, between the logics of the two schools is that the Peripatetic syllogistic is part of the logic of noun expressions, whereas the schemata of the Stoics represent principles that belong to the logic of propositions. In the time of Galen both of these logics became part of the syllabus for general philosophical education.

Commentaries and compendiums. After Theophrastus and Chrysippus, little creative work in logic was done in ancient times. To be sure, Galen wrote critical commentaries on the logical writings of Aristotle, Theophrastus, and Chrysippus; but those commentaries have not survived. After Galen there were only certain compendiums of logic, on the one hand, and scholarly commentaries on Aristotle's logical treatises, on the other. Much of this literature has been lost, but the works of Alexander of Aphrodisias; of Porphyry (died *c.* 301), a pupil of Plotinus, the founder of Neoplatonism; of Ammonius Hermiae (5th century AD), a Neoplatonic mathematician and astronomer; and of the latter's pupil Simplicius have been preserved at least in part. Although these authors made, on occasion, improvements in the minutiae of the Aristotelian and Stoic logic, they can hardly be credited with any major contribution. Logicians are indebted to them, however, for salvaging numerous fragments from the lost writings of earlier logicians. The same can be said of certain Roman authors: of Cicero (died 43 BC); of Martianus Capella (5th century AD) and the senator Cassiodorus (died *c.* 582)—both of whom prepared compendiums of ancient learning; of the translator and commentator Boethius (died AD 524/5), whose influence dominated the early Middle Ages; and of others. These scholars worked out the Latin terminology for the field of logic and transmitted the achievements of the Greek logicians to the logicians of the Middle Ages.

Medieval logic

DEVELOPMENT OF MEDIEVAL LOGIC

During the five centuries that followed the end of antiquity, little or nothing of any significance happened in the field of logic. As part of the basic curriculum (the trivium), logic was treated as a subsidiary subject providing useful training for students of law and theology. For the most part it was limited to discussing Porphyry's *Isagoge* and Aristotle's *Categoriae* and *De interpretatione* in a Latin translation by Boethius. Boethius' own logical writings were also read, along with some treatises by Gaius Marius Victorinus, a 4th-century African rhetorician, and the sections on logic to be found in the compendiums by Martianus Capella and Cassiodorus, and in the *Etymologiae* by the influential Spanish encyclopaedist Isidore of Seville (died *c.* 636). This sterility eventually began to change, however, because of the work of St. Anselm of Canterbury (died 1109), known for his ontological proof of the existence of God, and of Peter Abelard (died probably 1144), an eminent though controversial theologian and dialectician. During the 12th century, Latin translations of the remaining treatises of Aristotle's *Organon* were made, adding to the "old logic" a wealth of topics, then unfamiliar, that were given the name of the "new logic." In the 13th century logicians split: while some adhered to the Aristotelian orthodoxy and advocated the *logica antiqua,* others took a more liberal view and, in the faculties of arts of the newly established universities, propounded the *logica moderna.* The English logician William of Sherwood and his student Peter of Spain, later Pope John XXI, author of the most used textbook on logic for the next 300 years, seem to have been the most outstanding representatives of this new trend in logic. Among the logicians of the 14th century, mention should at least be made of the foremost philosopher of the century, William of Ockham; of Jean Buridan, also a scientist, who was the first to derive the laws of deduction deductively; and of his student and colleague Albert of Saxony. In the century that followed, Paul of Venice, an Augustinian theologian, produced a very comprehensive logic textbook under the descriptive title *Logica magna.* In the Greek world, the

The *logica antiqua* and the *logica moderna*

tradition of paraphrasing Aristotle's logical treatises and writing commentaries to them was continued by John Philoponus, by Stephanus of Alexandria, a 7th-century Neoplatonist, who transferred the tradition to Byzantium, and by others; and, in the 11th and 13th centuries, several logical compendiums were produced that deserve mention.

Arabic logic

Logical studies were also cultivated by the Arabs. By the middle of the 9th century, Porphyry's *Isagoge* and Aristotle's *Categoriae, De interpretatione,* and *Analytica priora* had been translated first into Syriac and then into Arabic. The remaining parts of the *Organon* were similarly translated during the second half of the century. The works of Galen, too, attracted the attention of Arabic scholars, and it is through this channel that the logic of the Stoics percolated to Arab logicians. A more original work on the Aristotelian *Organon* was initiated in a school of logicians that flourished at Baghdad in the 10th century, of which the most famous representatives were the Nestorian Abū Bishr Mattā ibn Yūnus; Abū Naṣr al-Fārābī, who acquainted the Arabs with Plato and Aristotle; and a pupil of both of the foregoing, the Jacobite Yaḥyā ibn ʿAdī. Unfortunately, their commentaries on Aristotle's logical writings and their own logical treatises, except for some commentaries by al-Fārābī, have been lost. In the view of the outstanding Persian philosopher Avicenna (died 1037), the school of Baghdad was too slavishly adherent to the Aristotelian tradition. He himself advocated a more independent line and gave expression to his concept of logic in his *Kitāb ash-shifāʾ* ("The Book of Healing"). The tradition of the school of Baghdad, which in Islām was regarded as Western, was revived in Muslim Spain during the 11th and 12th centuries. It reached its climax in the works of Ibn Rushd, known as Averroës, whose commentaries on the *Organon* played an important role in reviving the interest of Western scholars in Aristotle. In the 14th and 15th centuries, the logical studies by Arab scholars degenerated into the compiling of handbooks and the writing of explanatory notes to the handbooks bequeathed by earlier logicians.

The value of the Arabic contribution to the development of logic is not very great. Arab logicians kept interest in the *Organon* alive at a time when, in the West, only a small part of it was known to philosophers. They also preserved the Stoic tradition as transmitted by Galen; but so far, hardly any fresh light has been thrown on the logic of the Stoics from Arab sources. Of the three branches of medieval logic—the Byzantine, the Arabic, and the Scholastic—the latter appears to have been by far the most fruitful. Neither the Byzantine nor the Arab logicians seem to have achieved any lasting results. Even the hope that the writings of the Byzantine and Arab logicians might reveal important fragments of the logic of the ancient Greeks is not likely to be fulfilled.

THEORIES OF LANGUAGE

As in antiquity, so also in the Middle Ages, the study of language played a significant part in the development of logic. Following Aristotle, the Scholastic logicians of the High Middle Ages distinguished between nouns and verbs; but this distinction was often replaced by the distinction between the subject and the predicate in a proposition.

Categoremata and *syncategoremata;* theory of supposition

Expressions that could be used as subjects or as predicates were called *categoremata* (from Greek *katēgorein,* "to predicate") and distinguished from *syncategoremata* (from Greek *synkatēgorein,* "to predicate jointly"); *i.e.,* from such parts of propositions as "every," "all," "no," "some," "only," and the like. Both *categoremata* and *syncategoremata* were meaningful expressions; but whereas a categorematic expression designated something outside of the language (*i.e.,* had a referent), a syncategorematic expression was not capable of designation in the usual sense. By and large, syncategorematic expressions coincide with what is today described as logical vocabulary, and, indeed, some Scholastic texts implicitly suggested that logic should be construed as a theory of *syncategoremata.* According to Scholastic logicians, the function of a syncategorematic expression consists in changing or modifying the designation of categorematic expressions in a proposition—a problem connected with that of ambiguity in the use of *categoremata.* The Scholastics attempted to solve both of these problems with the aid of their theory of supposition, which is regarded as one of the original achievements of medieval logic. This theory appears to have originated in the 12th century; but it was Ockham and Albert of Saxony who gave it its traditional form. More than half a dozen different kinds of supposition can be distinguished. Supposition is a property of a term; *i.e.,* of a *categorema,* as used in a proposition. Thus, for instance, in the proposition "Man is mortal," the term man is supposed to have formal supposition; but the same term is said to have material supposition in the proposition "Man is a noun." The terminology could perhaps be more illuminating, but the distinction at issue clearly corresponds to the present distinction between use and mention. In the proposition "Every man is mortal," the term man designates several men; hence, its supposition is common. But in the proposition "This man is running," the term man designates exactly one man, and its supposition is described as discrete. This sort of distinction has special significance in the case of languages that, like Latin, have no articles. In such propositions as "Man is a species" and "Every animal apart from man is irrational," the term man has simple supposition. In application to the categorical propositions, the doctrine of supposition can be summarized, following Albert of Saxony, as follows: The subject in a singular proposition has discrete supposition, but in all indefinite or particular propositions it has determinate supposition. The supposition of the subject in a universal proposition is confused and distributive and so is the supposition of the predicate in a negative proposition. The supposition of the predicate in a particular affirmative proposition would be described, presumably, as determinate. The theory of supposition survives in traditional syllogistic as the theory of distributed terms, which asserts that in a given true proposition a term *a* is "distributed" if and only if, *a* having been replaced by a term *b* whose designata form a subclass of the designata of *a,* the proposition remains true.

FORMAL LOGIC

Versions of the logics of noun expressions, of propositions, and of modal expressions can all be distinguished in the history of medieval logic.

Logic of noun expressions. In the Middle Ages the logic of noun expressions continued in the form of syllogistic. Medieval logicians abandoned the use of variables. Consequently, they had to illustrate logical principles with the aid of standard examples or describe them in metalogical terms; and they employed examples of inferences rather than of logical laws. One of the more important improvements in their syllogistic was the recognition of the fourth figure (in which the middle term is predicated of the major and the minor of the middle). To be sure, syllogisms constituting the fourth figure had been known to the Scholastics. But this figure was not recognized as such until the Averroistic Jewish philosopher Isaac Albalag did so in the 13th century. A great deal of ingenuity was devoted to the task of devising various mnemonics for the purpose of memorizing valid syllogisms and remembering the ways in which they can be "reduced" to syllogisms of the first figure (in which the middle term is predicated of the minor and the major of the middle).

The fourth figure and reduction to the first figure

Logic of propositions. The logic of propositions as developed by Theophrastus and the Stoics was transmitted to the Scholastics by Boethius. It seems not to have stimulated further development, however; it seems, instead, to have ossified in the early medieval compendiums; hence it had to be rediscovered. Its revival in the 13th century took the form of the theory of "consequences." At first its application was limited to syllogistic; but gradually it reached the level of abstraction at which the specific contents of the propositions involved lost their relevance, and the theory acquired the generality that it had possessed in the logic of the Stoics. The Scholastic consequences involved the use of conjunctions, inclusive disjunctions (*i.e.,* alternations), and implications. All of these kinds of compound propositions were interpreted by the Scholastics truth-functionally. According to

Pseudo-Scotus, the author of *In universam logicam quaestiones* ("Questions on Universal Logic") formerly attributed to John Duns Scotus, a consequence is a hypothetical proposition composed of an antecedent and a consequent joined by means of a conditional connective ("if") or, more often, by one expressing a reason ("therefore," "hence") and thus forming an inference. These, however, can be divided into material and formal inferences. A consequence that is valid for any terms, provided that their arrangement is preserved, is formal; and a consequence that does not fulfill this condition is material. A material consequence is valid if by adjoining a true proposition as premise it can be reduced to a formal consequence. (Example: "Socrates is a man; therefore, Socrates is mortal" is a valid material consequence because "Every man is mortal; Socrates is a man; therefore, Socrates is mortal" is a formal consequence.)

In the medieval logic of propositions, logical principles were either described in metalogical terms or else they were illustrated with the aid of standard examples. Among the principles that were known to the Scholastics, but not known to the logicians of antiquity, the following can be mentioned:

New logical principles of the Scholastics

1. An alternation follows from each of its parts; *i.e.*,
   For all p and q, if p then (p or q)
   and
   For all p and q, if q then (p or q).
2. From a conjunction of contradictories there follows any other proposition; *i.e.*,
   For all p and q, if p and not-p then q.
3. From a false proposition there follows any other proposition; *i.e.*,
   For all p and q, if not-p then (if p then q).
4. Every true proposition follows from any other proposition; *i.e.*,
   For all p and q, if p then (if q then p).
5. An alternation of which the parts are contradictories of the corresponding parts of a conjunction is itself a contradictory of the conjunction; *i.e.*,
   For all p and q, (p or q) if and only if it is not the case that (not-p and not-q)
   and
   For all p and q, (p and q) if and only if it is not the case that (not-p or not-q).
6. If the opposite of the consequent is consistent with the antecedent, the consequence is not valid; *i.e.*,
   For all p and q, if p and not-q then it is not the case that if p then q.
7. If the opposite of the consequent is not consistent with the antecedent, the consequence is valid; *i.e.*,
   For all p and q, if it is not the case that (p and not-q) then if p then q.

Logic of modal expressions. In the Middle Ages, modal logic continued to be applied to syllogistic, but there is also some evidence of a more general development. Thus, the notions of truth, falsity, knowledge, opinion, doubt, appearance, will, and preference were added to the list of the traditional modal notions, and a number of modal theses were set up by Pseudo-Scotus, Ockham, Albert of Saxony, and Walter Burley, a 14th-century English Scholastic. The following theses can serve as examples of their modal principles:

1. The contingent does not follow from the necessary.
2. The impossible does not follow from the possible.
3. Anything whatsoever follows from the impossible.
4. The necessary follows from anything whatsoever.
5. Actuality (truth) can be validly inferred from necessity.
6. Possibility can be validly inferred from actuality (truth).
7. The possibility of negation can be validly inferred from the negation of necessity, and conversely.
8. The necessity of negation can be validly inferred from the negation of possibility, and conversely.

Fallacies and paradoxes. Aristotle's work *De sophisticis elenchis* attracted a great deal of attention from medieval logicians after it became available in the 12th century. Some of them regarded the study of sophisms as an essential part of an education in logic. As an exercise in clarifying the ambiguities of ordinary language, the inventing and solving of linguistic puzzles was not entirely fruitless; but of greater significance for logic was the discussion of the *insolubilia*. According to the polymath Scholastic St. Albertus Magnus (died 1280), teacher of St. Thomas, insoluble propositions are such that, whichever side of the contradictory is granted, the opposite also follows. (Example: If someone swears that he swears falsely, then, if he swears falsely, what he swears is not true—*i.e.*, he does not swear falsely; but if he does not swear falsely, then what he swears is true—*i.e.*, he swears falsely.) This is one of the many different and sometimes quite complicated versions of the ancient paradox of the liar. Various solutions of this paradox were discussed by logicians of the 14th and 15th centuries; in fact, Paul of Venice gave a list of 15 different attempts at explaining it. But the most interesting suggestions are to be found in an anonymous manuscript dating from the beginning of the 14th century. According to one doctrine, whoever says "What I am saying now is false" is in fact saying nothing; his utterance is simply meaningless. Another way of dealing with the difficulty consists in pointing out that the word false occurring in that statement cannot refer to the proposition of which it is a part. Both solutions have found advocates among contemporary logicians and philosophers.

The liar paradox

Logic in the East

Judging from the outline of the development of logic given so far, it would appear that logic has been an exclusive product of Western culture. Some historians of the subject, however, have found this view parochial and sought to identify traces of logic in Indian and Chinese thought. But research in these two fields is beset with tremendous difficulties: most of the texts remain unpublished or untranslated; some of the monographs are unreliable; and scholars well trained both in logic and philology are extremely rare. Thus, a fair evaluation is, as yet, impossible.

INDIAN LOGIC

Indian logic, which arose in the last centuries BC, when certain grammarians and religious thinkers became interested in methods of philosophical discussion, involved the study of texts called *sūtra*s and of numerous commentaries on them written over centuries. Around each of these texts a school developed, which cultivated and defended its doctrines, of which the Mīmāṃsā, Vaiśeṣika, Nyāya, and New Nyāya schools of Hindu logic are of particular interest. Although Pāṇini and Patañjali, Indian grammarians, were active as early as the 5th and 2nd centuries BC, the first millennium AD was the classical period in the development of Indian logic. Among the most important Hindu logicians of this period were Kumārila (7th–8th century) of the Mīmāṃsā school, Praśastapāda (5th–6th century) of the Vaiśeṣika school, and Vātsyāyana (5th–6th century) of the Nyāya school. Several Buddhist logicians also flourished during the classical period, most notably Dignāga (5th century), Dharmakīrti (7th century), and Dharmottara (8th–9th century). The second half of the second millennium AD constituted a new chapter in Indian logic, represented by the New Nyāya, among the followers of which were Raghunātha (16th century), and Mathurānātha, Jagadīśa, and Annam Bhaṭṭa (all 16th and 17th centuries).

Leading Hindu and Buddhist logicians

As the grammarians devised a special terminology for describing their language and formulating grammatical rules, they encountered problems similar to those of the modern distinction between language and metalanguage, as well as problems of definition, clarity, and economy in stating the rules of grammar.

In the *Vaiśeṣika-sūtra* (1st century AD), a doctrine was expounded that resembled Aristotle's theory of categories; *i.e.*, the theory of the kinds of entities that constitute the universe or of the generic terms that refer to them. The author distinguished (1) substances: earth, water, fire, air, ether, time, space, self, and mind; (2) attributes: colour, taste, smell, and touch, numbers, measures, separateness, conjunction and disjunction, priority and posteriority, understandings, pleasure and pain, desire and aversion, and volitions; and (3) actions: throwing upward, throwing downward, contraction, expansion, and motion. The remaining categories were (4) genus; (5) ultimate difference; and (6) inherence. To these, later texts added the category of (7) absence. It was no more clear, however, than it was

in the case of Aristotle whether the list was meant to classify the meanings of words or the things denoted by words.

The *Nyāya-sūtra* and the five-member syllogism

The *Nyāya-sūtra,* which was edited in the 2nd century AD, is regarded as the fundamental text of Indian logic. Its so-called five-member syllogism is usually illustrated by the following argument:

Proposition:	There is fire on the mountain.
Reason:	Because there is smoke on the mountain.
Example:	As in a kitchen—not as in a lake.
Application:	It is so (*i.e.,* it is so in the present case).
Conclusion:	There is fire on the mountain.

The interpretation of the argument is not simple. In particular, the role of the example is obscure. If it is meant to remind one of instances in which smoke and fire are present and of instances in which they are absent, then the syllogism is an argument by analogy and has no validity in logic. On the other hand, the reason would imply the conclusion if the example could be interpreted as standing for a universal statement: "Where there is smoke there is fire." Development in this direction seems, in fact, to have taken place in Buddhist logic. Thus, according to Dignāga, the following three-membered rule holds: if (1) the presence of a definite mark in the subject has been stated, and if it is remembered that (2) the same mark is certainly in everything similar to the subject, but (3) absolutely absent in everything unlike it, then the result of the inference is certainly valid. Dharmakīrti, a commentator on Dignāga, used in fact a universal premise as the example. In the final form of the syllogism, as found in a textbook by Annam Bhaṭṭa of the New Nyāya school, the author distinguished between inference for oneself and inference for another. The inference for oneself consists of three steps such as:

1. Everywhere where there is smoke, there is fire.
2. There is smoke on this mountain.

From the combination there arises the knowledge:

3. There is fire on the mountain.

Inference for another involves five members and reads:

> The mountain is fiery—that is the proposition. Because smoky—that is the reason. All that is smoky is fiery—that is the example. So here—that is the application. Therefore it is so—that is the conclusion.

The member that is described as the application raises a problem. Although modern commentators seem to regard it as mere restatement of the reason, which is the second member, it is tempting to view it as the application of the universal statement—*i.e.,* the third member—to the particular instance under consideration. If this were correct, then the expressions "It is so" and "So here" could be interpreted as abbreviations standing for "If there is smoke in the mountain then there is fire in the mountain" or "If the mountain is smoky then the mountain is fiery." Be that as it may, the five-member syllogism bears a close resemblance to the prosleptic syllogism of Theophrastus (see above *Theophrastus and the logic of noun expressions*).

Nascent inference schemata

An argument reminiscent of the Stoic inference schemata has been found in a Buddhist work, the *Kathāvatthu,* which reads:

> If the soul be known in the sense of a real and ultimate fact, then indeed, good sir, you should also say, the soul is known in the same way as any other real and ultimate fact is known. (2) That which you say here is wrong, namely, that we ought to say, "The soul is known in the sense of a real and ultimate fact," but we ought not to say, the soul is known in the same way as any other real and ultimate fact is known. (3) If the latter statement cannot be admitted, then indeed the former statement should not be admitted. (4) In affirming the former statement, while denying the latter, you are wrong.

It is easy to see that this argument is an instance of the following schema:

	(1) If p then q;
therefore,	(2) Not (p and not-q);
therefore,	(3) If not-q then not-p;
therefore,	(4) Not-(p and not-q).

All of these inferences are valid; but although another argument of similar form, with "not p" instead of "p" and "not q" instead of "q"," is found in the text, it would be too generous to credit the Buddhist logician with the discovery of the logic of propositions. In the *Kathāvatthu,* the arguments are not used as standard examples to represent any argument whatsoever of the same form. They occur in a discussion on the knowability of the soul and belong to psychology rather than to logic. Discussions of this kind do appear, however, to follow very strict rules based probably on practice and tradition.

The notion of negation seems to have been of particular interest to Indian logicians. Thus, the followers of the Mīmāṃsā explored the ways in which an injunction could be negated. They distinguished "He shall-not look," "He shall not-look," and "Not-he shall look" as possible negations of "He shall look." Mathurānātha, of the New Nyāya school, wrote, "Wherever one perceives a pot, there one does not perceive constant absence of a pot . . . . Accordingly, absence of constant absence of pot is essentially identical with pot." This statement looks like a double negation law expressed with the aid of a standard example. In the texts of the New Nyāya, the notion of negation often occurs in connection with abstract terms. Thus, instead of the simple proposition "The mountain is smoky," for instance, the cumbersome one "Not-mountain-ness qualifies the locus of not-being-smoky" is employed. This type of language invites comparison with the highly artificial Latin of the medieval logicians.

Evaluation of Indian logic

Compared with the logic of the ancient Greeks, Indian logic is not very impressive. It must be emphasized, however, that—unlike the logic of the Arabs—it developed independently of Greek thought. While it may be granted that the logic of propositions may have been anticipated by some Buddhist logicians, it does not seem that much progress was made. The logic of noun expressions asserted itself more firmly, as in the five-member syllogism and its variants, but it never reached the level of Aristotle's syllogistic. The development of Indian logic was severely handicapped by the failure of its logicians to make use of variables. As a result, no logical principles could be stated directly; they had to be illustrated by standard examples or described metalinguistically (*i.e.,* in talk about the language in which they might have been stated). Finally, in Indian thought, logical topics were not always separated from metaphysical and epistemological topics (on the nature of being and knowledge, respectively). It must be remembered, however, that present knowledge of the development of logic in India is incomplete and that it may have to be revised in the light of future research.

CHINESE LOGIC

For the most part, Chinese philosophy is concerned with practical and moral problems on the one hand and with mystical interpretations of life on the other. It has little room for the study of logic, which has remained neglected since the establishment of Neo-Confucianism in the 11th century AD. In the classical period of Chinese philosophy (5th to 3rd century BC), the controversies between the major philosophies of Confucianism, Taoism, and Moism gave rise to a kind of dialectic, which eventually was systematized to some extent by the Moists. They turned it into a weapon against the skepticism of the Taoists, who maintained that one point of view is as good as another because each is based upon different assumptions regarding the relations between names and objects. According to the Taoists, assumptions of this kind are entirely arbitrary. As dialecticians, the Moists were primarily interested in the study of names in relation to objects. They distinguished "unrestricted" names, which apply to every object; "private" names, which apply to one object each; and names like "horse," which apply to every object of a kind. (As noted earlier, a comparable classification of noun expressions had been worked out by Aristotle; see above *Aristotle and the logic of noun expressions: Structure of language.*) The Moists were also interested in defining the terms used in philosophical discussions and in listing various senses of ambiguous terms. Their definitions of "all" and "some" resemble the laws of obversion for the universal affirmative and particular affirmative propositions. They distinguished between absolute identity and relative identity (sameness in a specified respect). Gradually, the

The Moist dialectic of names and objects

Moists developed an interest in the forms of propositions and in inferences from propositions of the same form; but, whereas Western logicians searched for the similarity of form in the structures of propositions with a view to establishing logical principles or schemata of valid inference, the Moists considered instances in which the similarity of form does not warrant any inference. "A white horse is a horse" implies "Riding white horses is riding horses"; but one may ask whether "A robber is a man" implies "Killing robbers is killing men." The Moists' answer was in the negative; and they supported it by pointing out that, although a robber is a man, abounding in robbers is not abounding in men, and disliking the abundance of robbers is not disliking the abundance of men. The purpose of this argument, however, does not seem to have been to elucidate any logical or linguistic puzzle but to show that, contrary to the opponents of Moism, killing robbers was compatible with love for all men; it was not a point of logic that was at issue but a point of morals. Discussions of this sort could have led, in due course, to the investigation of logical problems, but as it happened, they did not do so. In developing logic, the Chinese thinkers did not advance beyond the stage of preliminaries, a stage that was reached in Greece by the Sophists in the 5th century BC.

Modern logic

LOGIC IN THE RENAISSANCE AND IN HUMANISM

Returning to the West, the logical tradition of the Middle Ages survived for about three centuries after it had reached its maturity in the 14th century. The intellectual climate, however, that established itself in the West with the advent of the Renaissance and Humanism did not enhance logical studies. The Latin of the Roman writers Lucretius, Cicero, and Quintilian proved more attractive than the language of Boethius and the Scholastics; and rhetoric took precedence over logic. At the same time, the authority of Aristotle was gradually replaced by that of Plato and the Neoplatonist Plotinus. The rise of the natural sciences also contributed to the neglect of logic, which, as a deductive discipline, gave way to methodological investigations. These developments called for a new approach to logic; and in the mid-16th century such an approach was in fact taken by Petrus Ramus, an anti-Aristotelian logician and educational reformer. Ramus described logic as an "art of discussing" and distinguished it from grammar and rhetoric, which in his view were concerned with matters of style. Logic, according to Ramus, should treat of concepts, judgments, inferences, and proofs—in that order. Among inferences he counted syllogisms, both categorical and hypothetical. To the former he added syllogisms involving the use of singular propositions. In accordance with the spirit of the time, he used illustrations from classical literature. The divisions of logic suggested by Ramus were adopted by the authors of *La Logique: Ou l'art de penser* (1662; Eng. trans. in *The Port-Royal Logic,* 1851), which consisted of four parts. The discussion of concepts and judgments in parts I and II, respectively, seems to contain a few original observations, which belong, however, to epistemology rather than to logic. In part III, on reasoning, the validity of syllogisms is determined by giving rules for each of the figures and proving corollaries. Although some of these rules were known to the Scholastics, *The Port-Royal Logic* gives a systematic presentation of what later writers of textbooks developed into a metalogical theory of syllogistic. In part IV, on method, Euclid's *Elements* is recommended as a model of scientific method. Following René Descartes, the father of modern philosophy, they insisted that, in any scientific inquiry, obscure or equivocal terms should be defined; that only perfectly known terms should be used in definitions; that only self-evident truths should be used as axioms; and that all propositions that are not self-evident should be proved with the aid of axioms, definitions, and already proved propositions. Despite the competition of an entirely new conception of logic afforded by Gottfried Wilhelm Leibniz, a German Rationalist of widespread interests, the ideas expounded in *The Port-Royal Logic* continued to be popular to the mid-19th century.

Period of neglect and the new approach of Ramus

The Port-Royal Logic

RISE OF MATHEMATICAL LOGIC DURING THE ENLIGHTENMENT

Modern logic began in the 17th century with Leibniz and developed in partnership with mathematics.

Leibniz' characteristica universalis. Leibniz influenced his contemporaries and successors through his ambitious program for logic, which for him had ceased to be "a scholar's diversion" and began to take the form of a "universal mathematics." His program envisaged a universal language based on an alphabet of thought, or *characteristica universalis,* a general calculus of reasoning, and a general methodology. In Leibniz' view, the universal language should be like algebra or like a version of the Chinese ideograms and consist of basic signs standing for simple unanalyzable notions. Complex notions would then be signified by appropriate constructions involving the basic signs, which would thus reflect the structure of complex notions and, in the final analysis, that of reality. The use of numerals to represent unanalyzable notions would hopefully enable the truths of any science, when formulated in terms of the universal language, to be computed by arithmetical operations. Neither the idea of an artificial language nor that of reducing reasoning to computing was entirely new; it was owing to Leibniz' vision and confidence, however, that the two ideas came eventually to fruition—as witnessed by the arithmetization of logical languages by Kurt Gödel, a Moravian-U.S. mathematical logician, on the one hand, and by the emergence of contemporary computer science, on the other. Leibniz' more concrete contributions to the development of logic appear to be twofold: he successfully applied mathematical methods to the interpretation of the Aristotelian syllogistic, and in proposing a "calculus of real addition" he showed that parts of algebra are open to nonarithmetical interpretation.

In discussing Aristotle's syllogistic, Leibniz rightly insisted that each of the four categorical propositions has existential import (*i.e.,* its terms have instances), and that, consequently, there are six valid syllogisms in each of the four figures. He also correctly suggested that singular propositions could be treated as universal. But his most important finding consisted in a very ingenious arithmetical interpretation of syllogistic, which rests on the assumption that variables take ordered pairs of relatively prime numbers as their values. Leibniz showed how, on this assumption, the four categorical propositions should be interpreted, and he demonstrated that, when so interpreted, all valid syllogisms are verified and the rejected syllogisms falsified. Morever, his interpretation has proved to be adequate even in relation to a very intricate rule of rejection proposed for syllogistic in the 1940s by the Polish logician Jerzy Słupecki of the Warsaw school of logic, later professor at the University of Wrocław (formerly Breslau). The problem of defining the four propositions in terms of identity, difference, logical product of noun expressions, and logical complement long exercised Leibniz' mind. Anticipating the use of quantifiers, Leibniz suggested the following definitions (slightly streamlined):

Arithmetical interpretation of syllogistic

1. For all *a* and *b,* every *a* is *b* if and only if for all *c,* if *c* is *a* then for some *d, d* is *b* and *c* is the same object as *d.*
2. For all *a* and *b,* some *a* is *b* if and only if for some *c, c* is *a* and for some *d, d* is *b* and *c* is the same object as *d.*
3. For all *a* and *b,* no *a* is *b* if and only if for all *c,* if *c* is *a* then for all *d,* if *d* is *b* then *c* is different from *d.*
4. For all *a* and *b,* some *a* is not *b* if and only if for some *c, c* is *a* and for all *d,* if *d* is *b* then *c* is different from *d.*

Leibniz claimed that, given these definitions, one can prove all of the valid syllogisms with the aid of two principles: one of them states that "For all *a, b,* and *c,* if *a* is the same object as *b* and *b* is the same object as *c,* then *a* is the same object as *c*"; and the other states that "For all *a, b,* and *c,* if *a* is the same object as *b* and *b* is different from *c,* then *a* is different from *c.*" Actually, Leibniz was right, provided that one is entitled to assume that "For all *a* and *b, a* is different from *b* if and only if it is not the case that *a* is the same object as *b,*" and that "For all *a,* for some *b, b* is *a.*"

Leibniz' treatment of Aristotelian syllogistic reflects his interest in the notions of identity and difference. In one of his papers he stated that those terms are "the same"

Calculus of identity

or "coincident" of which either can be substituted for the other wherever one pleases without loss of truth—*e.g.*, "triangular" and "trilateral." This informal definition expresses an idea that today would be expressed directly as the law of extensionality for noun expressions: "For all a and b, $a = b$ if and only if [for all F, $F(a)$ if and only if $F(b)$]," F standing for any predicate (verblike expression). Here "$a = b$" means that "a's are the same as b's." Leibniz expanded his calculus of identity through the notions of "complement" and "real addition": In his symbolism, the "real sum" of the terms a and b is expressed as "$a \oplus b$," but the complement of a is simply expressed as "not-a." Leibniz' calculus of real addition has the form of a deductive system based on axioms and definitions. Given the theory of identity, the axiomatic foundations of this system can be condensed into the following simple proposition: "For all a, b, and c, $a = b \oplus c$ if and only if ($a = c \oplus b$ or it is not the case that $a = a \oplus a$)." If one wishes to incorporate the notion of complement, however, then the single axiom would have to be replaced by the following: "For all a, b, and c, $a = b \oplus c$ if and only if it is not the case that [a = not-($c \oplus b$) and $a = a \oplus a$]." Leibniz' logical papers remained for the most part unknown until *La Logique de Leibniz* was published by Louis Couturat, a French philosopher of mathematics, in 1901. Had these papers been published during the lifetime of Leibniz, the revival of logic, which in fact occurred in the mid-19th century, would have happened much earlier.

Search for clarity and use of diagrams. The new approach to syllogistic and the search for an algebra capable of nonarithmetical interpretation characterized the development of logic after Leibniz. At the end of the 17th century and during the 18th century, these two lines of research were followed mainly by mathematicians. Girolamo Saccheri, an Italian Jesuit, influenced by the Cartesian methodology, proceeded to apply to logic the standards of clarity expected in geometrical proofs. He was dissatisfied with Aristotle's use of extra-logical counterexamples in rejecting invalid syllogisms. Consequently, he proposed a different method of rejection, suggested by a certain kind of proof found in Euclid. It was the method of *reductio ad absurdum*, but a *reductio* with a difference. Saccheri used, for example, the premises

Saccheri's *reductio ad absurdum*

1. Every syllogism of the first figure with a universal major and an affirmative minor premise is a valid syllogism;

and

2. No AEE-syllogism (*i.e.*, with A and E premises and an E conclusion) of the first figure is a syllogism with a universal major and an affirmative minor premise;

and concluded from them that

3. No AEE-syllogism of the first figure is a valid syllogism.

Actually, though the premises and the conclusion are true, the argument as it stands is not valid. A further premise, which Saccheri tacitly assumed, is needed, viz.,

4. If some AEE-syllogism of the first figure is a valid syllogism, then for all a, b, and c, if every b is c and no a is b then no a is c.

With this addition, given the syllogistic laws of contradiction, the premises (1), (2), and (4) now imply (3) as a conclusion and enable the usual *reductio* to be applied; for if, contrary to the conclusion, it is additionally assumed that "Some AEE-syllogism of the first figure is a valid syllogism," then it can be easily shown that, even on this assumption, the conclusion holds—which means that the assumption is redundant. No extra-logical vocabulary has here been used, provided that logic is understood to include metalogic, and in particular metasyllogistic.

The practice of illustrating categorical propositions by geometrical diagrams was already in evidence in the 16th century. Though it was later used by Leibniz, and by Johann Heinrich Lambert and Gottfried Ploucquet—authors of two early formulations of logical calculi—it has come to be associated with the name of Leonhard Euler, an 18th-century Swiss mathematician.

If one wants to go further than Euler and distinguish between singular, general, and nonreferential noun expressions, then properties and relations that belong to or hold between extensions of singular, general, and nonreferential noun expressions can be determined with reference to diagrams tabulated as shown in Figure 1. A small shaded circle represents the extension of a singular noun expression and an unshaded circle that of a general noun expression. No circle is used in the case of a nonreferential noun expression. The diagrams I.1, I.2, and I.3 (Figure 1) illustrate the following propositions: "There exists exactly one a," "There exist more a's than one," and "There exist no a's." The spatial relations between the circles in diagrams II.1–II.16 correspond to the relations between the extensions of noun expressions represented by a and b. Some, but not all, of these relations are fairly familiar. Thus, for instance, diagrams II.1, II.2, and II.3 illustrate the following propositions: "a is the same object as b," "a is different from b," and "a is one of many b's." And the proposition "No a's are b's" is illustrated by diagram II.13.

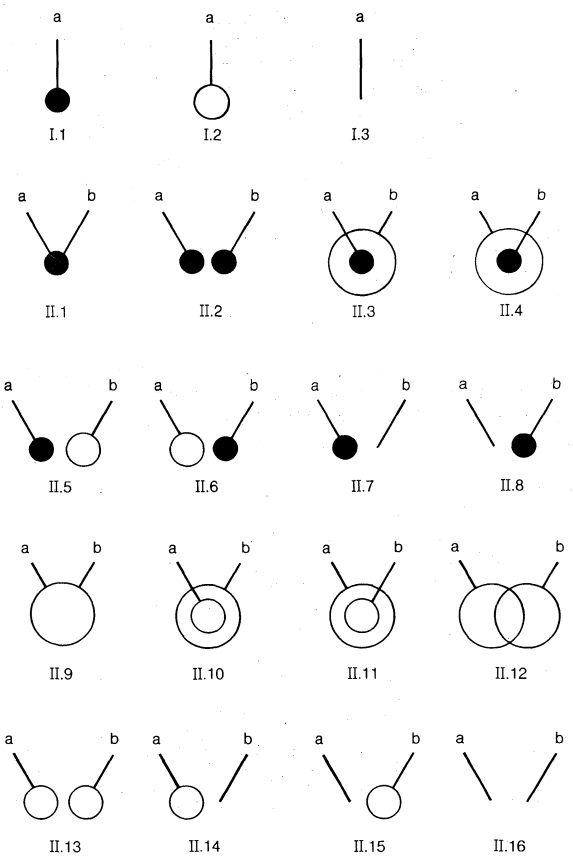


Figure 1: Properties and relations belonging to or holding between extensions of singular, general, and nonreferential noun expressions (see text).
From *Ratio* (1958); Blackwell, Oxford

Diagrams for relations between kinds of noun expressions

In his *Lettres à une princesse d'Allemagne* (3 vol., 1768–72), Euler used diagrams II.10 and II.13 to illustrate the universal affirmative and the universal negative propositions and used II.12 to illustrate the two particular propositions. If the interpretation suggested by his diagrams is adhered to rigorously, however, it is found to fail for the syllogistic laws of identity, contradiction, subalternation, and subcontrariety, as well as for the syllogisms that depend on these laws.

Use of the diagrams by Euler

DEVELOPMENT OF MATHEMATICAL LOGIC IN THE 19TH CENTURY

During the 19th century, and especially during its last half, the foundations were being laid for the most remarkable developments in the history of logic.

Expansions of syllogistic by Gergonne and Bentham. More than 100 years after Euler, his idea of geometrical diagrams was elaborated systematically by the Frenchman Joseph Diez Gergonne, a mathematician and editor of a mathematical journal. Following the Aristotelian tradition,

Gergonne disregarded both singular and nonreferential noun expressions, but his treatment of general noun expressions can be regarded as complete. He needed only five diagrams to determine the five simple relations between the extensions of two general referential noun expressions, viz., $a \mid b$, $a \subset b$, $a \supset b$, $a \times b$, and $a \mathrel{H} b$ (in Figure 1: II.9, II.10, II.11, II.12, and II.13). These simple relations can now be used to define the categorical propositions of syllogistic, as set out below:

For all a and b, every a is b if and only if $a \mid b$ or $a \subset b$;
For all a and b, no a is b if and only if $a \mathrel{H} b$;
For all a and b, some a is b if and only if $a \mid b$ or $a \subset b$ or $a \supset b$ or $a \times b$;
For all a and b, some a is not b if and only if $a \supset b$ or $a \times b$ or $a \mathrel{H} b$.

Gergonne's relations can be defined, conversely, in terms of the syllogistic constants. Thus, Gergonne's innovation consisted, in fact, in expanding the vocabulary of syllogistic and proposing new kinds of inference based upon the expansion. He did not, however, axiomatize his system. This was done, instead, in the 20th century by John Acheson Faris, a Belfast logician.

From the point of view of logic, "every . . . is" (or "all . . . is") is treated as a single syntactically unanalyzable term, which in concatenation with two noun expressions forms a proposition. Traditionally, however, it had been held that "every" (or "all") modifies the way in which the subject should be construed. This suggested to logicians a similar modification of the predicate. The idea was not new, however; it had been discussed by Aristotle in the *De interpretatione,* only to be abandoned later. Some of Aristotle's ancient commentators determined that a modification of this sort would generate 16 different propositions. But the authority of Aristotle prevented them from enquiring more deeply into the problem. Because of the writings of Sir William Hamilton, an influential Scottish metaphysician, "the doctrine of the quantified predicate" became topical in the first half of the 19th century and led to a bitter controversy between him and Augustus De Morgan, an English mathematician and logician. Hamilton distinguished eight different forms of proposition that involved quantification of the predicate, but apparently he did not know that the same distinctions had already been made by George Bentham, an English botanist and systematist. Bentham was concerned with the relations of identity and exclusion (or outsidedness) as holding between both total and partial extensions of two noun expressions. In his system of notation, "$tX = pY$" meant that the total extension of X is identical with a proper part of the extension of Y, and "$pX \parallel pY$" meant that a proper part of the extension of X is entirely outside a proper part of the extension of Y. Hamilton used a symbolism of his own to express his eight forms of proposition, but he also expressed them in ordinary language. The table shown below gives, on the left-hand side, propositions distinguished by Bentham, and, on the right-hand side, their Hamiltonian versions:

The work of Hamilton and Bentham

a.	$tX = pY \ldots$	All X is some Y.
b.	$tX \parallel pY \ldots$	Any X is not some Y.
c.	$tX = tY \ldots$	All X is all Y.
d.	$tX \parallel tY \ldots$	Any X is not any Y.
e.	$pX = pY \ldots$	Some X is some Y.
f.	$pX \parallel pY \ldots$	Some X is not some Y.
g.	$pX = tY \ldots$	Some X is all Y.
h.	$pX \parallel tY \ldots$	Some X is not any Y.

Obviously, Bentham arrived at his list of propositions by permutating his symbols, but he was not very clear about their interpretation. Thus, he claimed, for instance, that propositions (b) and (d) both represent the E proposition. He thought that proposition (d) was rather useless and that the remaining propositions could be reduced to the four standard types. Hamilton, too, was somewhat confused about the meanings of his eight propositions. He identified propositions (a), (d), (e), and (h) with the A, E, I, and O propositions, but, following Euler, he illustrated each of these propositions with the aid of one set of circles. This means, in fact, that each of these propositions exhibits the use of a Gergonne constant. For the purpose of such an interpretation, however, "some" had to be understood as meaning "some but not all." On these assumptions, propositions (a), (c), (d), (e), and (g) can be regarded as exemplifying the Gergonne constants $\subset$, $\mid$, H, $\times$, and $\supset$, respectively. Proposition (f) then turns out to be just another version of (e), whereas (b) and (h) can be taken to express compound relations between extensions: proposition (b) can be construed as equivalent to "All X is some Y or some X is some Y" and proposition (h) as equivalent to "Some X is all Y or some X is some Y." Alternatively, one can accept Hamilton's identification of propositions (a), (d), (e), and (h) with the four categorical propositions and treat propositions (b), (c), (f), and (g) as expanding the vocabulary of syllogistic. This, in fact, is how Hamilton viewed his contribution, though he also claimed, wrongly, that his scheme was an improvement on Aristotelian logic. Using the new vocabulary, Hamilton and his followers worked out a great number of additional syllogisms and tried to work out a simple criterion for testing their validity. At the root of the difficulties experienced by Hamilton was his pardonable failure to distinguish between "all" and "some" as integral parts, on the one hand, of the constants "all . . . are" and "some . . . are," and, as used, on the other hand, in "For all $a, \ldots$" and "For some $a, \ldots$," which are quantifiers in the sense accepted today. When expressed in a more sophisticated logical idiom, the conceptions of Bentham and Hamilton reveal an affinity to the Leibnizian definitions involving identity and difference.

Expansion of syllogistic by De Morgan. A more successful attempt at expanding syllogistic was made by Augustus De Morgan, who developed a logic of noun expressions in *Formal Logic* (1847), in his *Syllabus of a Proposed System of Logic* (1860), and also in a number of articles. Like Bentham and Hamilton, De Morgan was concerned with comparing the extensions of terms. In his symbolism, a parenthesis following the subject—as X)—or preceding the predicate—as (Y—indicates that the total extensions of X and Y are under consideration. If only part of the extension of a term is involved, the parenthesis is reversed—as $X($ or $)Y$. By placing the predicate after the subject together with the respective parentheses, a proposition is formed that states the identity of the extensions as specified by the parentheses. For instance, "$X))Y$" states that the total extension of X is identical with a part of the extension of Y—*i.e.,* that "All X is Y"; and "$X()Y$" states that a part of the extension of X is identical with a part of the extension of Y—*i.e.,* that "Some X is Y." To express exclusion between extensions instead of identity, a dot is placed between the parenthesis belonging to the subject and that belonging to the predicate. Thus, "$X).(Y$" means that "No X is Y." De Morgan also made use of negated terms, introducing the notion of a logical complement. In his system negated terms are counted as primitive. They are symbolized by small letters; thus, "x" means the same as "non-X," "x" and "X" forming a pair of contradictory terms. Given the use of negative terms, De Morgan examined systematically all of the forms of affirmative categorical propositions. They are:

Relations between the extensions of terms

a.	$X))Y \ldots$	Every X is Y.
b.	$x))y \ldots$	Every non-X is non-Y.
c.	$X))y \ldots$	Every X is non-Y.
d.	$x))Y \ldots$	Every non-X is Y.
e.	$X()Y \ldots$	Some X is Y.
f.	$x()y \ldots$	Some non-X is non-Y.
g.	$X()y \ldots$	Some X is non-Y.
h.	$x()Y \ldots$	Some non-X is Y.

De Morgan formulated a simple rule for generating propositions equivalent to those above; viz., replace one of the terms by its contradictory, turn its bracket the other way, and put a dot between the brackets—or if there is a dot already present, remove it. Thus, from each of the eight propositions one derives three other equivalent propositions. One now finds that to each proposition involving a negated term there is an equivalent in which both terms are positive. Thus, the original list can now be replaced by the following:

a′.	$X))Y \ldots$	Every X is Y.
b′.	$X((Y \ldots$	Every Y is X.
c′.	$X).(Y \ldots$	No X is Y.

d′. $X(.)Y\ldots$ Everything is either X or Y.
e′. $X()Y\ldots$ Some X is Y.
f′. $X)(Y\ldots$ Something is neither X nor Y.
g′. $X(.(Y\ldots$ Some X is not Y.
h′. $X).)Y\ldots$ Some Y is not X.

The meaning of (d′) and that of its contradictory (f′), however, does not quite harmonize with the way in which the symbolism was to be interpreted originally. One would expect, for instance, that (f′) should mean the same as "Only every X is Y," thus expressing an identity of the extension of X and that of Y. But this would not be compatible with the meaning of (f).

New kinds of inference and the theory of relations

De Morgan achieved a further extension of his syllogistic vocabulary with the aid of definitions. This extension gave rise to new kinds of inference, both direct (involving one premise) and indirect (involving two premises). Through this symbolism he was able to work out purely structural rules for transforming a premise or a pair of premises into a valid conclusion.

Inferences that appeared to illustrate principles belonging to the logic of noun expressions, but which could not be accommodated in syllogistic, had been known to logicians before De Morgan. It is to him, however, that credit should be given for initiating and developing a theory of relations. He invented a symbolism in which such notions as the contradictory, the converse, and the transitivity of a relation could be expressed, as well as the union of two relations (by which he meant their relative product). He also stated some logical principles, among them:

Contradictories of converses are themselves converses;
Converses of contradictories are contradictories;
The contradictory of the converse is the converse of the contradictory;
If a first relation is contained in a second, then the converse of the first is contained in the converse of the second, and the contradictory of the second is contained in the contradictory of the first;
A transitive relation has a transitive converse;
The contradictory of a transitive relation is not necessarily transitive.

De Morgan also set out, in four figures, various inference schemata analogous, in a sense, to those of syllogistic.

Boole's algebra of logic. Leibniz' pioneering attempts at constructing an algebraical calculus capable of nonarithmetical interpretation were continued by Ploucquet, Lambert, Jean Castillon, an Italian-born mathematician, and others, but the results achieved by them were not of lasting value. Real progress was made by George Boole (died 1864), a self-taught English mathematician who presented his new ideas on logic in the *Mathematical Analysis of Logic* (1847) and in *An Investigation into the Laws of Thought, on Which are Founded the Mathematical Theories of Logic and Probabilities* (1854).

Notion of a "true calculus" and its laws of combination

In Boole's view, the notion of a "true calculus" rests on the employment of symbols in accordance with some general but well-determined laws of combination. Although the combining of symbols in this way should admit of a consistent interpretation, the interpretation should not prejudge the validity of what Boole called the process of analysis; *i.e.*, the process whereby interpretable combinations of symbols are obtained. Whereas Leibniz sought a calculus that admitted of a nonarithmetical interpretation, Boole was advocating a formal calculus that could accommodate a number of different interpretations—in particular, two: one relevant to relations among things, and the other to relations among facts as expressed by propositions. In other words, Boole appears to have been aware of the distinction between the logic of noun expressions and that of propositions, respectively. He did not, however, realize that the logic of noun expressions presupposes the logic of propositions.

Boole's calculus was not a deductive system but it is clear from the way in which it was developed that the following presuppositions were acceptable to him:

1. For all x and y, $xy = yx$.
2. For all x and y, $x + y = y + x$.
3. For all x, y, and z, $x(y + z) = xy + xz$.
4. For all x, y, and z, $x(y - z) = xy - xz$.
5. For all x, y, and z, if $x = y$ then $xz = yz$.
6. For all x, y, and z, if $x = y$ then $x + z = y + z$.
7. For all x, y, and z, if $x = y$ then $x - z = y - z$.
8. For all x, $x(1 - x) = 0$.

Except for the last formula, the system of these presuppositions lends itself to ordinary numerical interpretations. If the values of the variables are restricted to 0 and 1, then the eight presuppositions become interpretable in such a restricted algebra. "The laws, the axioms, and the processes of such an algebra," Boole claimed, "will be identical in their whole extent with the laws, the axioms, and the processes of an algebra of logic." Boole had in mind the interpretation relevant to relations among things; *i.e.*, the class interpretation within the logic of noun expressions. Thus, "$x = y$" means that the classes x and y have the same members; "xy" represents the intersection of the classes x and y (*i.e.*, their common part); "$x + y$" represents the class consisting of the members of x and the members of y when they have no members in common; "$x - y$" represents the class of those members x that are not members of y; and finally, "1" represents the universal class, of which everything is a member, and "0" the null class, of which nothing is a member.

Boole's calculus can also be interpreted within the logic of propositions. In this interpretation, the variables x, y, and z become propositional variables, "xy" becomes a conjunction of two propositions (read "x and y"), "$x + y$" is exclusive disjunction "Either x or y but not both," and "$-x$" becomes "It is not the case that x." The symbols 1 and 0 are now construed as the standard true and the standard false propositions, respectively. The sign of identity "=" becomes that of equivalence "if and only if." Interpreted in this way, the calculus can be extended by adding to it

9. For all x, either $x = 1$ or $x = 0$,

which does not hold for the class interpretation.

Refinements by Jevons and Peirce

Refinements of the calculus. Boolean algebra was subsequently improved by various researchers. William Stanley Jevons, a British economist and logician, rightly urged that all intermediate formulas involved in the process of algebraical transformations should lend themselves to the same interpretation as the initial and final formulas. A further improvement was the replacement of the exclusive by the inclusive disjunction: thus, within the logic of noun expressions, "$x + y$" denoted all those objects each of which was either x but not y, or y, *but not* x, or both x and y. Consequently, it was true to say that "For all x, $x + x = x$," and also that "For all x, $x + -x = 1$." Neither of these formulas obtains, however, in the arithmetical algebra.

For the purpose of computing the validity (or invalidity) of inferences, Jevons constructed a "logical piano," a forerunner of the modern computer.

Charles Sanders Peirce, a U.S. logician, engineer, and Pragmatist philosopher, also held that inclusive disjunction is more suitable for an algebra of logic than exclusive disjunction. In addition, he employed the notion of inclusion, for which he devised a special symbol: —<. Peirce interpreted "a—<b" as meaning "Whatever is a is b" or "If a then b." Thus believing, like Boole, that the logic of noun expressions and that of propositions are merely two different interpretations of one and the same algebra of logic, he assimilated propositions to terms expressing relations. Usually, such terms had been thought of as involving two arguments. According to Peirce, however, there is no reason for not subsuming under the notion of relations terms that require three or more arguments (a, b, c, . . .) as well as terms that require one argument (predicates), or no argument (propositions).

Peirce also set up, independently of Frege, an axiom system for the logic of propositions based on "if . . . then" and on the notion of a proposition that implies any proposition; and he systematically developed the truth-table method of defining constant terms within the algebra of logic. He calculated that, given two values T and F (obviously standing for "true" and "false"), there can be four different functors (or operators) for one argument, 16 for two arguments, and 256 for three. Apparently Peirce considered introducing more values than two and thus anticipated the 20th-century development of many-valued

logics. Peirce introduced the notion of quantifiers, "For all $a, \ldots$" and "For some $a, \ldots,$" needed in elaborating his theory of relations. And finally, he was responsible for the idea of treating relations as classes of ordered pairs or, more generally, as classes of ordered n-tuples.

Refinements by Schröder

Boole's algebra of logic was given the form of a deductive system by Ernst Schröder, a German logician and mathematician, in his *Operationskreis des Logikkalkuls* (1877; "The Operational Sphere of the Logical Calculus") and, more systematically, in his *Vorlesungen über die Algebra der Logik* (1890–1905; "Lectures on the Algebra of Logic"), which includes the theories of identity and of relations. The system that he had constructed, like that of Peirce, made use of the notion of inclusion, the meaning of which can be determined by reference to diagrams II.1, II.3, II.8, II.9, II.10, II.15, and II.16 in Figure 1. If inclusion is symbolized thus: $\subset$, then its axioms are:

1. For all a, $a \subset a$.
2. For all a, b, and c, if $a \subset b$ and $b \subset c$ then $a \subset c$.
3. For all a, $0 \subset a$/
4. For all a, $a \subset 1$.
5. For all a, b, and c, $c \subset ab$ if and only if ($c \subset a$ and $c \subset b$).
6. For all a, b, and c, $a + b \subset c$ if and only if ($a \subset c$ and $b \subset c$).
7. For all a, b and c, $a(b + c) = ab + ac$.
8. For all a, $a - a \subset 0$.
9. For all a, $1 \subset a + -a$.

Schröder's interpretation of + was in harmony with that of Jevons and Peirce. At the time of its publication, Schröder's *Vorlesungen* gave, and still gives, the fullest account of an algebra of logic and of its development.

One of the notions investigated by Schröder was that of being an individual, which he defined as follows:

10. For all a, a is an individual if and only if (it is not the case that $a \subset 0$ and for all x, $a \subset x$ or $a \subset -x$).

In addition, he characterized the new notion with the aid of an axiom that states that: "A nonempty class contains at least one individual," or more precisely

11. For all a, if it is not the case that $a \subset 0$, then for some x, $x \subset a$ and x is an individual.

Whereas Schröder's conception of an algebra of logic emphasized the logical aspect of the theory, the inquiries of Alfred North Whitehead, who later became a leading process metaphysician, as presented in *A Treatise on Universal Algebra* (1898), were aimed at integrating the algebra of logic with other branches of mathematics. Extensive studies of the axiomatic foundations of the algebra of logic were successfully undertaken by Edward Vermilye Huntington, a Harvard University mathematician, and by Alfred Tarski, the founder of semantics as part of metalogic.

Logic and the foundations of mathematics. As mentioned earlier, Boolean algebra has interpretations within both the logic of propositions and the logic of noun expressions. Neither Boole nor Schröder, however, was concerned with the relationship between the two interpretations or the two logics. But this was not true of Gottlob

Frege's reduction of arithmetic to logic

Frege, an outstanding German mathematician and an original philosopher. Frege set for himself a new problem, the elucidation of the notion of numbers. He felt that the notion of natural number could be reduced to logical notions and that, consequently, arithmetic could be shown to be a part of logic. Frege therefore devised a symbolic language that, in proving logical or arithmetical theses, dispensed with ordinary usage. By applying axiomatic methods, he hoped to determine the meaning of the vocabulary of the logic of propositions. Thus, for Frege, the problem was to develop a calculus or an algebra for a given interpretation. This he did for "if . . . then" and "it is not the case that" (or simply "not," construed as the negation of a proposition). In terms of ordinary usage, the axioms required for Frege's logic of propositions read as follows:

1. For all p and q, if p then if q then p.
2. For all p, q, and r, if (if p then if q then r) then if (if p then q) then (if p then r).
3. For all p and q, if (if p then q) then (if not-q then not-p).
4. For all p, if not-(not-p) then p.
5. For all p, if p then not-(not-p).

In his subsequently published *Begriffsschrift: Eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* (1879; "Concept-Script: A Formal Language of Pure Thought on the Pattern of Arithmetic"), Frege supplemented the foregoing axiom system with the following propositions:

6. For all a, b, and F, if a is the same object as b then if $F(a)$ then $F(b)$.
7. For all a, a is the same object as a.
8. For all a and F, if for all x, $F(x)$ then $F(a)$.

Axioms 6–8 constitute the foundations of a system of the logic of noun expressions or, more precisely, of the logic of predicates.

In *Die Grundlagen der Arithmetik* (1884; Eng. trans., *The Foundations of Arithmetic,* 2nd rev. ed., 1953), Frege discussed, thoroughly but not formally, the notion of number. He first introduced the notion of equality or equinumerosity between concepts by positing that "The concept F is equal to concept G if and only if for some R, R correlates one to one the objects falling under F with the objects falling under G." By the number that belongs to the concept F, Frege understands an extension of the concept "equal to the concept of F." Thus, the following definition of number is formulated: "For all n, n is a number if and only if for some F, n is the number that belongs to F."

A formal reduction of arithmetic to logic was given in Frege's *Grundgesetze der Arithmetik: Begriffsschriftlich abgeleitet* (1893 and 1903; "Basic Laws of Arithmetic: Conceptually Derived"). The axioms and rules of inference given in the *Grundgesetze* differ slightly from those laid down in the *Bergriffsschrift,* and he also added rules for determining the correctness of definitions. Frege also introduced the device of binding variables by the universal quantifier. Finally, Frege made important contributions to the theory of language. His analyses of the notions of sense and reference as well as of object, concept, and function have contributed greatly to the interest in semantics and in theories of meaning.

Cantor's development of set theory

Another trend in the study of logic and the foundations of mathematics began with the establishment, by the German mathematician and philosopher Georg Cantor (died 1918), of the theory of manifolds or sets as a new mathematical discipline. By a set, Cantor understood "a collection into one whole of definite, distinct objects of our perception or our thought, which are called the elements of the set." Thus, the notion of set appears to have much in common with that of class. A set is uniquely defined by its elements. Sets with identical elements are themselves identical, and identical sets have identical elements. It would appear that set theory is just another version of the logic of noun expressions, similar to one of the interpretations of the algebra of logic. And indeed, a number of set-theoretical notions have their analogues in the algebra of logic and in systems such as that of Frege. Thus, whereas Frege talked about the number belonging to a concept, Cantor made use of the notion of the power or cardinal number of a set, a measure of its order of magnitude. He was especially interested in the cardinal numbers of nonfinite sets. His general theorem, to the effect that the cardinal number of a set is always smaller than that of the set of its subsets, gives rise to a series of transfinite numbers, each of which is the cardinal number of a nonfinite set. The smallest such number is the cardinal number of the set of natural numbers. Preoccupation with the notion of numerical infinity led Cantor to the idea of developing a general arithmetic of transfinite numbers within the framework of his theory of sets. Frege highly appreciated Cantor's *Grundlagen einer allgemeinen Mannigfaltigkeitslehre* (1883; "Foundations of a General Theory of Aggregates"), which on many counts differed from his own system, mainly in terminology. But, whereas Frege's findings were presented as a deductive system based on axioms and developed in accordance with explicit rules, Cantor's theory of sets was still in a pre-axiomatic stage.

Discovery of fundamental paradoxes

Cantor eventually discovered a paradox in the theory of sets. If one considers the set U that comprises all sets, then its subsets must be among its elements; hence, the cardinal number of U cannot be smaller than that of the set of the subsets of U; yet, in accordance with the general

theorem mentioned above, the cardinal number of the set of the subsets of U should be greater than the cardinal number of U. The problem became acute when in 1902 Bertrand Russell, author (with A.N. Whitehead) of the *Principia Mathematica,* discovered another paradox, this time in the system of Frege, involving the notion of the class C of all of those classes that are not elements of themselves. If the class C is an element of itself, then, like all of the elements of C, it is not an element of itself; but if C is not an element of itself, then it has the property that qualifies it for the membership of C; *i.e.,* it is an element of itself—hence the contradiction. Frege could not find a fault in Russell's reasoning. In a postscript to volume II of his *Grundgesetze,* Frege tried to remedy the situation by altering one of his axioms. It is now known, however, that the corrected system remains contradictory, and Frege himself was not at all confident that he had found the right solution. Nonetheless, the collapse of Frege's system did not prevent Russell from holding, unreservedly, that arithmetic can be derived from logic.

20th-century logic

LOGICISM, INTUITIONISM, AND FORMALISM

When, early in the 20th century, Bertrand Russell set out to show that arithmetic is an extension of logic, he had the benefit of the earlier research of Giuseppe Peano, an Italian mathematician and logician. Peano had not been convinced that the primitive notions of arithmetic—of 0, of number, and of the successor of a number—could be reduced to the vocabulary of logic. He thought it desirable, however, to base arithmetic on axiomatic foundations. A proposal to this effect was first published in his *Arithmetices Principia Nova Methodo Exposita* (1889). In his *Formulaire de mathématiques,* 5 vol. in 3 (1895–1908), Peano made use of the following axiom system for arithmetic:

1. 0 is a number.
2. The successor of any number is a number.
3. No two numbers have the same successor.
4. 0 is not the successor of any number.
5. Any property that belongs to 0, and also to the successor of every number that has the property, belongs to all numbers.

Supplemented with appropriate definitions, these axioms yield all of the characteristic laws of arithmetic; but they fail to single out the progression of natural numbers from other progressions. It is this defect that Russell was anxious to remove by deriving arithmetic from logic. He first approached the problem informally in *The Principles of Mathematics* (1903). Then, in collaboration with Whitehead, he produced the monumental *Principia Mathematica,* 3 vol. (1910–13), which has become a classic of logic. The *Principia* marked a climax of the researches in logic and the foundations of mathematics that had been flourishing since the time of Leibniz. At the same time, it provided a starting point for the development of logic in the 20th century.

In addition to Cantor's paradox in his theory of sets and Russell's proof that Frege's system is contradictory, several other paradoxes were discovered during the 15 years prior to the *Principia* that cast doubt upon men's logical intuitions. Among them were those known as Burali-Forti's paradox of the greatest ordinal number, Berry's paradox of the least integer not nameable in fewer than 19 syllables, Richard's paradox of the class of all decimal numbers definable in a finite number of words, and Grelling's paradox of autological and heterological adjectives. The liar paradox of Eubulides, already known among the ancient Megarian logicians, can also be included in this list.

Russell's logicism and the theory of types

In the *Principia,* where the paradoxes are discussed at considerable length, Russell argued that they result from a "vicious circle" that consists in assuming illegitimate totalities. A totality is illegitimate when it is supposed to involve *all* of a collection but is itself *one* of the same collection. In the *Principia,* such totalities cannot be generated because of Russell's adoption of the theory of logical types. The theory of logical types demands that a functor or an operator should be regarded as belonging to a higher logical type than any of its arguments, or that a class belongs to a higher logical type than that to which its elements belong. (Similarly, a predicate belongs to a higher logical type than the object of which it is predicated.) Consequently, to say that a class is an element of itself is neither true nor false but simply meaningless. Although the theory of logical types obviated the paradox of the class of all classes not members of themselves, it raised certain problems of its own. It was by no means clear whether the theory was a kind of ontology that classified extra-linguistic entities or a kind of grammar that classified expressions of a logical language. Moreover, some critics charged that it was an ad hoc palliative against paradoxes, and that the ramifications of its hierarchy of types were unduly complicated. Far-reaching suggestions for simplifying the theory were offered by Leon Chwistek, a Polish logician, and the short-lived Cambridge philosopher Frank Ramsey. And, clearly, the paradoxes called for more individual treatment.

Another theory proffered by Russell, of particular interest to philosophers, was his theory of definite descriptions. The *Principia* is a work of impressive scope that, in addition to purely logical topics such as the logic of propositions, and the theories of quantification, of classes, and of relations, includes cardinal arithmetic and the theories of ordinals and of relation-numbers in general, of series, and of measurement.

The view that mathematics is a continuation of logic, with no sharp demarcation between them, has been called "logicism." Opposed to it are intuitionism and formalism. "Intuitionism" is associated with the names of L.E.J. Brouwer, a Dutch mathematician, and his pupil Arend Heyting, according to whom mathematics presupposes mental constructions, which, in turn, presuppose the set of natural numbers. The progression of natural numbers is somehow derived from the intuition of time and the awareness of separable moments of human experience. Thus, intuitionistic mathematics is an activity of thought, and the statement of universal logical laws is not necessary. In particular, intuitionists reject "the thoughtless application of the logical theorem of tertium exclusum," which states that there can be no middle (third) between two contradictories; and by doing so, they believe that they are rejecting, in fact, the principle that every mathematical problem is soluble. They also reject the elimination of double negation, which, in their view, allows its proponents erroneously to infer the provability of a proposition from the unprovability of its negation. Existential statements of the form "There exists an x such that . . ." are prohibited unless they are implied by statements of the form "a is such that . . . ," a restriction that harmonizes with the intuitionistic insistence on constructivism, the view that mathematical proofs must be constrained by the possibilities of finitistic mental constructions. Finally, intuitionists reject the notion of complete or actual infinity; there are no infinite sets other than denumerable sets, such as the set of natural numbers, which exemplifies potential infinity.

Brouwer's intuitionism

Whereas Frege regarded the truth of the axioms of a theory as evidence that the theory is consistent, a Göttingen mathematician, David Hilbert, the founder of modern formalism, maintained that freedom from contradiction in an arbitrarily posited axiom system is the guarantor of the truth of the axioms. Although the problem of whether the fundamental notions of mathematics could be defined in terms of logical notions did not interest Hilbert, he would have answered the question in the negative. He conceived of mathematics as a stock of formulas, which constitute the subject matter of a new mathematics, a "metamathematics," the task of which is to examine and secure the foundations of the former. Hilbert, like Frege, thought very highly of axiomatic methods; but, unlike Frege, he was less concerned with the meaning of the axioms. Within mathematics proper, he advocated a complete formalization with purely formal inferences without any reference to a possible interpretation of the formulas involved. In metamathematics, however, inferences are to have regard to the subject matter. The principal problem of metamathematics is the search for mathematical theories that can be proved to be consistent and complete.

Hilbert's formalism

As a result of the extensive studies undertaken by Gödel, however, it is known that most interesting mathematical theories are incomplete because every theory that includes the arithmetic of natural numbers is incomplete; and in order to prove the consistency of such theories, the consistency of stronger theories must be assumed.

LOGIC OF PROPOSITIONS

As logics became more specialized, it was the logic of propositions that, largely because of its simplicity, attracted the attention of those logicians who favoured specialization. Propositional connectives other than "if . . . then," "or," and "it is not the case that" were examined for their suitability as primitive (or basic) terms in systems of propositional calculus. The method of truth tables that display the T–F values of compound propositions for all combinations of the T–F values of the component propositions was developed by Peirce, by the Polish logician Jan Łukasiewicz, by the U.S. logician Emil Post, a specialist in recursive unsolvability, by the Austrian-English logician and philosopher Ludwig Wittgenstein, and by other logicians to determine the meanings of propositional connectives and to evaluate expressions consisting of connectives and variables. The problem of the interdefinability of the connectives was examined and solved by Henry M. Sheffer, a Harvard logician, who proved that the joint denial "neither . . .nor" and the alternative denial "either it is not the case that . . . or it is not the case that . . ." are each so fundamental that all of the remaining propositional connectives can be defined in terms of it—a finding that had been anticipated by Peirce. J.-G.-P. Nicod, a French logician, found that a system of the calculus of propositions equivalent to that of the *Principia* could be based on a single axiom with the alternative denial as the only primitive notion. On the problem of simplifying the axiomatic foundations of this logic, considerable work was done by Łukasiewicz (died 1956) and by his collaborators in Warsaw and later in Dublin. Efforts to develop systems of the logic of propositions that dispense with certain dubious principles have been made, as in Hilbert's positive calculus, in the minimal calculus of the mathematician Ingebright Johansson of the University of Oslo, and in Heyting's intuitionistic system. Here, too, belong Wilhelm Ackermann, one of Hilbert's collaborators, and his system of rigorous implication, and the system of entailment constructed by Alan Ross Anderson of the University of Pittsburgh, a philosopher of mathematics and science, and N.D. Belnap, his associate.

Leśniewski's prototheticand Lewis' systems of strict implication

Another line of development consisted in strengthening systems of the logic of propositions by employing internal quantifiers to bind propositional variables. A system extended in this way was set up by Łukasiewicz and Tarski. A still stronger system was envisaged by Stanisław Leśniewski (died 1939), one of the founders of the Warsaw school, who called it a system of "prototheticc." It allowed for the use of functorial variables, such as those for which propositional connectives can be substituted. Axiomatic foundations can be obtained for prototheticc by subjoining to the axioms of a classical system the law of extensionality for propositions; *i.e.,* that "For all p and q (p if and only if q) if and only if for all f, $f(p)$ if and only if $f(q)$." In addition to employing the usual rules of inference, systems of prototheticc are equipped with rules of quantification, definition, and extensionality. The rule of extensionality enables laws of extensionality to be introduced for any semantic category that may be brought into the system by means of definitions. Systems of prototheticc are the richest as regards their vocabulary and syntactical variety, and the strongest as regards their deductive power.

In classical logic, a proposition of the form "If p then q" is considered to be true if and only if it is not the case that both p is true and q is false. This Philonian interpretation of "if . . . then," which was challenged in antiquity, was challenged again, soon after the publication of *Principia,* by Clarence Irving Lewis, a conceptual Pragmatist and author (with C.H. Langford) of *Symbolic Logic* (1932), a classic of modal logic. In this and an earlier work (1918), Lewis distinguished between the material implication of classical logic and strict implication, which he found preferable. Construed as an instance of strict implication, a proposition of the form "If p then q" is true if and only if it is not possible that p should be true and q false. Lewis's treatment of implication involves the modal notion of possibility and thus marks the beginning of new developments. Lewis suggested several different systems of strict implication (S_1–S_5), and many more have been constructed since. Among them, Gödel's system with necessity as a primitive notion deserves special mention, and so does the system with the notion of contingency, constructed by E.J. Lemmon, an Oxford logician. More recently, the logic of propositions has been enlarged to include deontic logic, the aim of which is to exhibit the use of such expressions as "it ought to be the case that," "it is permissible that," and "it is forbidden that" and to interrelate them with modal notions (see below *Practical logic*).

Łukasiewicz' multivalued logic

A different approach to extending the logic of propositions was taken by Łukasiewicz. His analysis of the modal notions in Aristotle and his resolve to justify indeterminism suggested the idea of a three-valued logic and of many-valued logics in general (see below *Formal logic*). In these systems, Łukasiewicz employed the notions of implication, alternation, conjunction, and negation, but the meanings of the corresponding constants—"if . . . then," "or," "and," and "it is not the case that"—differ from their meanings in two-valued logic. Łukasiewicz determined their new meanings with the aid of many-valued truth tables, which, in addition to truth and falsity, allow for intermediate truth-values—such as possibility, for instance. Although Łukasiewicz suggested axiomatic foundations for three-valued and infinitely-many-valued logics, his systems were not complete. Nor was it possible to define, within their framework, all of the constants for which the meaning could be determined with the aid of appropriate truth tables. Many-valued systems that have no such defects have been developed by Bolesław Sobociński, a logician formerly at Warsaw, later at Notre Dame, and by the Polish logician Jerzy Słupecki. Purely formal systems of many-valued logics have been constructed independently by Emil Post.

The most recent and, philosophically, by far the most exciting way of extending the logic of propositions consists in constructing systems of tense logic with a view to clarifying the conception of time (see below *Logics of physical application*). According to the British logician A.N. Prior, the characteristic constants of tense logic are "it has been the case that," "it is now the case that," and "it will be the case that." Several other constants can be introduced by means of appropriate definitions.

Modal logic, deontic logic, tense logic, and, of course, many-valued logics presuppose the abandonment of two-valued logic. In many cases, they also presuppose the rejection of the principle of extensionality. Some logicians, however, do not approve of these serious sacrifices and believe that alternative ways of solving these problems should be explored.

LOGIC OF NOUN EXPRESSIONS

At the end of the 19th century, the logic of noun expressions had developed the following branches: Aristotle's syllogistic, Leibniz' theory of identity, Schröder's algebra of logic interpreted as a theory of classes and relations, Frege's theory of quantification, and Cantor's set theory. Although in the 20th century some of these branches have been growing on their own, the ideal of comprehensiveness, achieved by the authors of *Principia,* has not been abandoned altogether. As a result of Łukasiewicz' research, syllogistic has been given the form of a modern deductive theory. It has been axiomatized and proved to be consistent and decidable in the sense that every syllogistic or polysyllogistic expression can be either derived from the axioms or shown to be rejected. Other fragments of the logic of noun expressions have been distinguished by logicians and made objects of special studies—such as, for instance, the functional calculus of first order, functional calculi of higher orders, and the functional calculus of order omega (ω), often referred to as the simple theory of types.

The tendency toward comprehensiveness found expres-

sion in the system of the logic of noun expressions constructed by Leśniewski, whose interest was aroused by Russell's paradox of the class of all those classes not elements of themselves. Although Leśniewski accepted the need for a theory of logical types, he did not believe that the paradox resulted from confusing types (or "semantic categories" as he called them) but from the failure to realize that class expressions often are used ambiguously.

Leśniewski's ontology and mereology

In order to explain this ambiguity, he constructed two theories, called ontology and mereology. Ontology, which is a system of the logic of noun expressions, is obtained by subjoining some ontological axioms to the protothetic, adapting its rules to them, and setting up a rule of ontological definition and a rule of ontological extensionality. In standard systems of ontology, the copula "is" is the only primitive ontological term. Diagrams II.1 and II.3 (Figure 1) determine its meaning. In Schröder's algebra, it can be defined as follows: "For all a and b, a is b if and only if both $a \subset b$ and a is an individual." In ontology, it is characterized by the following single axiom: "For all a and b, a is b if and only if (*i*) for some c, c is a, (*ii*) for all c and d, if c is a and d is a then c is d, and (*iii*) for all c, if c is a then c is b." As regards its contents, ontology comprises syllogistic, the theory of identity, a counterpart of the theory of classes, and counterparts of functional calculi of any orders. If required, ontology can be supplemented with the axioms of infinity and of choice. Whereas the logical language of the *Principia* presupposes the concept of proper name, that of ontology is based on common noun expressions, to which proper names are assimilated. The mere use of the language of ontology implies no existential assertions, which is not the case with the language of the *Principia*. Instead of a theory of logical types, Leśniewski had developed a theory of semantic categories, which was inspired by certain ideas of Edmund Husserl, the founder of Phenomenology, and was concerned with the grammar of logical language and not with any extralinguistic entities.

Mereology, which is the theory of part–whole relations, is obtained by subjoining mereological axioms to ontology and adapting the rules of ontology to them. Whereas ontology was meant to provide the foundations of arithmetic, mereology was to serve as the basis of geometry. Mereology, and *a fortiori* ontology, is consistent if protothetic is consistent. Theories of part–whole relations have also been worked out independently of Leśniewski's mereology by J.H. Woodger, author of an axiomatized biology, by the U.S. logician Henry Leonard, and Nelson Goodman, a U.S. philosopher.

The development of Cantor's theory of sets was greatly affected by the discovery of paradoxes. For, it soon became clear that more solid foundations were required if paradoxes were to be eliminated. With this aim in view, the German mathematician Ernst Zermelo axiomatized the theory and succeeded in avoiding the paradoxes without resort to the theory of types. The Israeli mathematician Abraham Fraenkel has informally set out Zermelo's axioms:

Zermelo's axiomatization of set theory

1. *Axiom of Extensionality:* Sets containing the same members are equal.
2. *Axiom of Pairing:* For any two different sets a and b, there exists a set c that contains just a and b.
3. *Axiom of Union:* For any set a that contains at least one member, there exists the set the members of which are just the members of the members of a.
4. *Axiom of Power-set:* For any set a there exists the set the members of which are just subsets of a.
5. *Axiom Schema of Singling-out:* For any set a and any monadic predicate P, there exists the set that contains just those members x of a that fulfill the condition $P(x)$.
6. *Axiom of Choice:* If a is a disjointed set that does not contain the null-set, the Cartesian product of the members of a is not empty (the Cartesian product of two sets is the set of all pairs having a member from each set).
7. *Axiom of Infinity:* There exists a set a such that the empty set is a member of a and for all b, if b is a member of a then the set of which the only member is b is also a member of a.

This axiom system is deductively strong enough to yield a substantial part of Cantor's theory, but it does not imply the existence of certain special sets presupposed by Cantor. To accommodate these sets, Fraenkel suggested

8. *Axiom of Replacement:* If the domain of a single-valued function is a set, its counter-domain is also a set.

And in order to eliminate certain extraordinary sets, the Hungarian-U.S. mathematician John von Neumann, a pioneer in computer mathematics, proposed:

9. *Axiom of Foundation:* Every non-empty set a contains a member b such that a and b have no members in common.

Von Neumann also distinguished between sets, which can be members of other sets, and classes, which have no such property. Both sets and classes can have members, but the notion of membership in the case of sets is not the same as membership in the cases of classes.

More recently, the Swiss logician and mathematician Paul Bernays, Hilbert's collaborator, has made important contributions to the study of the axiomatic foundations of set theory.

So far, the theory based on the above-listed axioms has proved to be free from contradictions, but the axioms have scarcely made the notion of a set any clearer. If the axioms describe some aspects of reality, it is not known what these aspects are. Thus, the theory is becoming detached from any interpretation.

According to Cantor's "continuum hypothesis," if the cardinal number of a denumerable set is denoted by $\aleph_0$ ("aleph-null") and that of the power-set (the set of all subsets) of a denumerable set by $2^{\aleph_0}$, then there is no cardinal number between $\aleph_0$ and $2^{\aleph_0}$—which means that $2^{\aleph_0} = \aleph_1$. In its generalized form, the hypothesis says that "For any A if a is the cardinal number of an infinite well-ordered set then there is no cardinal number between a and 2^a."

In the systems of the logic of noun expressions developed by W.V.O. Quine, a Harvard University logician, the theory of logical types is replaced by the much simpler theory of stratification. A propositional formula is said to be stratified if it is possible to put numerals for the variables in such a way that $\in$ occurs only in contexts of the form "$n \in (n+1)$." Now, the rule of abstraction, which corresponds to the axiom schema of singling out, stipulates that the formula from which abstraction is made should be stratified. This sort of restriction enabled Quine to abandon the theory of types. His system, first published in 1937 and known as system **NF**, was subsequently developed into system **ML**, published in 1940 and revised in 1951.

Quine's theory of stratification

METALOGICAL STUDIES

Once systems of logic had been axiomatized, their development seemed to have slowed down. Instead of proving new theses within the framework of axiomatized systems, logicians began to examine the systems themselves and prove theses about them. This was the beginning of metalogic (see below *Metalogic*), the development of which owes a great deal to Hilbert's conception of metamathematics and to his formalist philosophy of mathematics.

Problems of consistency, completeness, and decidability

The problems that are the concern of metalogic are many and various. In the first place, there are the problems of the consistency and completeness of deductive systems and their decidability (since for some systems one can construct intrinsically undecidable propositions) as well as the problem of the independence of the axioms and that of the definability of primitive terms within an axiom system.

For systems of the logic of propositions, methods have been developed for solving these problems. Thus, the consistency of the calculus of propositions was proved by Post and independently by Łukasiewicz. Methods of proving the completeness of the calculus were worked out by Łukasiewicz and others. Truth tables provided means of determining the decidability of systems and the mutual independence of their axioms. The Italian mathematician Alessandro Padoa (died 1938), a collaborator of Peano, found a method of proving the independence of primitive notions within an axiomatized theory.

Systems of the logic of noun expressions are more difficult objects for metalogical investigation, but here too there are several findings to be recorded. Because of Łukasiewicz' and Słupecki's researches, it is known that syllogistic, as

axiomatized by Łukasiewicz, is consistent, decidable, and based upon independent axioms. The completeness of the first order functional calculus was proved by Gödel and its consistency by Hilbert. As Alonzo Church, an eminent U.S. logician, has shown, however, the decision problem for a pure functional calculus of first order is unsolvable, although there is a decision procedure for that part of the calculus that contains monadic functions only. The most interesting finding in this field of study was obtained by Gödel, who succeeded in proving that a system of logic that is deductively strong enough to yield the arithmetic of natural numbers, viz., the system of Russell's *Principia,* is not complete and cannot be made complete by extending its axiomatic foundations. Furthermore, Gödel proved that the consistency of systems that include the arithmetic of natural numbers can only be proved if one assumes the consistency of stronger systems. Other metalogical findings obtained by Gödel concern intuitionistic logic and the theory of sets. In the latter field, Gödel established the fact that the *axiom of choice* and the *generalized continuum hypothesis* are consistent with the remaining axioms provided that these axioms are consistent themselves. The independence of the *axiom of choice* and that of the *generalized continuum hypothesis* were proved, in 1963, by Paul J. Cohen, a Stanford University mathematician, thus opening a new chapter in the history of set theory. It has been proved that Quine's system **ML,** as revised, is consistent if system **NF** is consistent.

Another branch of metalogical studies is syntax, which is concerned with the language of logic and describes the purely formal aspects of that language. One of its tasks is to formulate the directives of a logical system; *i.e.,* the rules of inference, of extensionality, and of definition. Among logicians whose contributions to syntax have been of lasting value are the German-U.S. philosopher Rudolf Carnap, once a member of the Vienna Circle, who defined a number of fundamental syntactical concepts applicable to languages of any form; Leśniewski, who stated the directives of his systems, including the rule of definition, in purely syntactical terms; Kazimierz Ajdukiewicz, a Polish methodologist and philosopher of science, who developed a system of categorial indices to identify syntactical categories of a language; and Gödel, who devised a method of arithmetizing syntax—*i.e.,* of uniquely assigning a number to an expression of symbolic language.

Carnap's studies in philosophical syntax and Tarski's in semantics

Syntax must be distinguished from semantics, which is another metalogical discipline. Speaking generally, semantics studies the relationship between language and reality. Among semantical concepts are those of truth, denotation, satisfaction, and definition or rather determination. Semantics as a metalogical discipline was established by Tarski, whose paper "Der Wahrheitsbegriff in den formalisierten Sprachen" ("The Concept of Truth in the Formalized Languages"), in the *Studia Philosophica,* vol. I (1935), has become a classic. Tarski also laid the foundations for the theory of logical consequence, which provides a conceptual framework for the study of deductive systems in abstraction from the language in which they are formulated. Independently of Tarski's researches, a thorough study of the concept of logical deducibility was made by Karl Popper, an Austrian-British authority in methodology of natural sciences.

An alternative to the method of expressing logical principles directly with the aid of variables consists in setting up schemata of valid inference. The method was developed by the German mathematician Gerhard Gentzen, who was for several years an assistant to Hilbert, and independently by Stanisław Jaśkowski, a logician of the Warsaw school. Both logicians tried to fashion their systems of logic to reflect the actual procedures of practicing mathematicians; they have therefore come to be known as systems of natural deduction. These systems are widely employed by logicians today.

Foundational function of logic in education

Role of logic as an element within a culture

Both in the West and in the East, the origin of logic is associated with an interest in the grammar of language and in the methodology of argument and discussion, be it in the context of law, religion, or philosophy. More than interest is needed, however, for the successful development of logic: it appears that logic can thrive only in a culture that upholds the conviction that controversies should be settled by the force of reason rather than by the orthodoxy of a dogma or the tradition of a prejudice. For this reason logic has made much greater progress in the West than in the East.

It has been acknowledged since the time of Aristotle that logic has a part to play in general education. Before embarking on the study of any science, one should, as Aristotle thought, receive some training in logic. Today logic is taught at almost every university, and for many courses it is a compulsory subject; it is gradually taking over the role that the study of classical languages used to play in general instruction. Indeed, for better understanding the working of one's native language, the study of logic is as helpful as the study of Latin grammar.

According to Plato, the axiomatic method is the best method to use in presenting and codifying knowledge, and for centuries Euclid's *Elements* has served as the paradigm of deductive theory. In the 17th century, when, because of the work of René Descartes, the ideals of clarity and precision were vindicated, it became the ambition of many scholars to present their disciplines *more geometrico;* today, however, it is logic and not geometry that sets the example of precision.

Bertrand Russell once wrote that there are two ways of researching in mathematics: one aims at expansion, the other at exploring the foundations. And the same is true of any other discipline. But it is in exploring the foundations of a science that one again encounters logic; for every science that claims to describe some aspect of reality and allows for proof makes use of logical vocabulary. This means that logic lies at the head of a ramified hierarchy of sciences and can be conceived of as the most abstract and most general description of reality. Viewed in this way, logic has a part to play in philosophy and, more precisely, in metaphysics. (Cz.L./Ed.)

LOGIC SYSTEMS

Formal logic

As noted earlier, the discipline known as formal logic takes as its main subject matter propositions (or statements, or assertively used sentences) and deductive arguments, and it abstracts from their content the structures or logical forms that they embody. The logician customarily uses a symbolic notation to express these structures clearly and unambiguously and to enable manipulations and tests of validity to be more easily applied. Although this discussion freely employs the technical notation of modern symbolic logic, its symbols are introduced gradually and with accompanying explanations so that the serious and attentive general reader should be able to follow the development of ideas.

Formal logic is an a priori, and not an empirical, study. In this respect it contrasts with the natural sciences and with all other disciplines that depend on observation for their data. Its nearest analogy is with pure mathematics; indeed, many logicians and pure mathematicians would regard their respective subjects as indistinguishable, or as merely two stages of the same unified discipline. Formal logic, therefore, is not to be confused with the empirical study of the processes of reasoning, which belongs to psychology. It must also be distinguished from the art of correct reasoning, which is the practical skill of applying logical principles to particular cases; and, even more sharply, it must be distinguished from the art of persuasion, in which invalid arguments are sometimes more effective than valid ones.

GENERAL OBSERVATIONS

Probably the most natural approach to formal logic is through the idea of the soundness or validity of an argument of the kind known as deductive. A deductive argument can be roughly characterized as one in which the claim is made that some proposition (the *conclusion*) follows with strict necessity from some other proposition or propositions (the *premises*)—*i.e.,* that it would be inconsistent or self-contradictory to assert the premises but deny the conclusion.

If a deductive argument is to succeed in establishing the truth of its conclusion, two quite distinct conditions must be met: first, the conclusion must in fact follow from the premises—*i.e.,* the deduction of the conclusion from the premises must be logically sound; and, second, the premises themselves must be true. Of these two conditions, the logician as such is concerned only with the first; the second, the determination of the truth or falsity of the premises, is the task of some special discipline or of common observation, etc., appropriate to the subject matter of the argument. When the conclusion of an argument is correctly deducible from its premises, the *inference* from the premises to the conclusion is said to be (deductively) *valid,* irrespective of whether the premises are true or false. Other ways of expressing the fact that an inference is deductively valid are to say that the truth of the premises gives (or would give) an absolute guarantee of the truth of the conclusion or that it would involve a logical inconsistency (as distinct from a mere mistake of fact) to suppose that the premises were true but the conclusion false.

The deductive inferences with which formal logic is concerned are, as the name suggests, those for which validity depends not on any features of their subject matter but on their form or structure. Thus the two inferences

(A) "Every dog is a mammal"; "Some quadrupeds are dogs"; therefore, "Some quadrupeds are mammals"

and

(B) "Every anarchist is a believer in free love"; "Some members of the Government party are anarchists"; therefore, "Some members of the Government party are believers in free love"

differ in subject matter and hence require different procedures to check the truth or falsity of their premises. But their validity is ensured by what they have in common, viz., that the argument in each is of the form:

Inference forms

(C) "Every X is a Y"; "Some Z's are X's"; therefore "Some Z's are Y's."

Line (C) above may be called an *inference form,* and (A) and (B) are then *instances* of that inference form. The letters—X, Y, and Z—in (C) mark the places into which expressions of a certain type may be inserted. Symbols used for this purpose are known as *variables;* their use is analogous to that of the x in algebra, which marks the place into which a numeral can be inserted. An instance of an inference form is produced by replacing all of the variables in it by appropriate expressions—*i.e.,* ones that make sense in the context—and by doing so uniformly—*i.e.,* by substituting the same (or some synonymous) expression wherever the same variable recurs. The feature of (C) that guarantees that every instance of it will be valid is its construction in such a manner that every uniform way of replacing its variables to make the premises true automatically makes the conclusion true, too; or in other words that no instance of it can have true premises but a false conclusion. In virtue of this feature, the form (C) is called a *valid* inference form. In contrast,

(D) "Every X is a Y"; "Some Z's are Y's"; therefore, "Some Z's are X's"

is not a valid inference form, for although instances of it can be produced in which premises and conclusion are all true, instances of it can also be produced in which the premises are true but the conclusion is false—*e.g.,*

(E) "Every dog is a mammal"; "Some winged creatures are mammals"; therefore, "Some winged creatures are dogs."

Formal logic as a study is concerned with inference forms rather than with particular instances of them; one of its tasks is to discriminate between valid and invalid inference forms and to explore and systematize the relations that hold among valid ones.

Valid proposition forms

Closely related to the idea of a valid inference form is that of a valid proposition form. A proposition form is an expression of which the instances (produced as before by appropriate and uniform replacements for variables) are not inferences from several propositions to a conclusion but propositions taken individually; and a *valid* proposition form is one for which all of the instances are true propositions. A simple example is

(F) "Nothing is both an X and a non-X."

An invalid proposition form is one that has at least some false instances. Formal logic is concerned with proposition forms as well as with inference forms. The study of proposition forms can, in fact, be made to include that of inference forms in the following way: let the premises of any given inference form (taken together) be abbreviated by alpha (α) and its conclusion by beta (β). Then the condition stated above for the validity of the inference form "α, therefore β" amounts to saying that no instance of the proposition form "α and not-β" is true—*i.e.,* that every instance of the proposition form

(G) "Not both: α and not-β"

is true—or that line (G), fully spelled out, of course—is a valid proposition form. The study of proposition forms, however, cannot be similarly accommodated under the study of inference forms, and so for reasons of comprehensiveness it is usual to regard formal logic as the study of proposition forms. Because a logician's handling of proposition forms is in many ways analogous to a mathematician's handling of numerical formulas, the systems he constructs are often called *calculi.*

Much of the work of a logician proceeds at a more abstract level than that of the foregoing discussion. Even a formula such as (C) above, though not referring to any specific subject matter, contains expressions like "every" and "is a," which are thought of as having a definite meaning; and the variables are intended to mark the places for expressions of one particular kind (roughly, common nouns or class names). It is possible, however—and for some purposes it is essential—to study formulas without attaching even this degree of meaningfulness to them. The construction of a system of logic, in fact, involves two distinguishable processes: one consists in setting up a symbolic apparatus—a set of symbols, rules for stringing these together into formulas, and rules for manipulating these formulas; the second consists in attaching certain meanings to these symbols and formulas. If only the former is done, the system is said to be *uninterpreted,* or *purely formal;* if the latter is done as well, the system is said to be *interpreted.* This distinction is important because systems of logic turn out to have certain properties quite independently of any interpretations that may be placed upon them. An axiomatic system of logic can be taken as an example—*i.e.,* a system in which certain unproved formulas, known as axioms, are taken as starting points, and further formulas (theorems) are proved on the strength of these. As will appear later (*Axiomatization of PC*), the question whether a proof of a given theorem in an axiomatic system is a sound one or not depends solely on which formulas are taken as axioms and on what the rules are for deriving theorems from axioms, and not at all on what the theorems or axioms mean. Moreover, a given uninterpreted system is in general capable of being interpreted equally well in a number of different ways; hence in studying an uninterpreted system, one is studying the structure that is common to a variety of interpreted systems. Normally a logician who constructs a purely formal system does have a particular interpretation in mind, and his motive for constructing it is the belief that when this interpretation is given to it the formulas of the system will be able to express sound or true principles in some field of thought; but, for the above reasons among others, he will usually take care to describe the formulas and state the rules of the system without reference to interpretation and to indicate as a separate matter the interpretation that he has in mind.

Many of the ideas used in the exposition of formal logic,

Problems in the philosophy of logic

including some that have already occurred above, raise problems that belong to philosophy rather than to logic itself (see PHILOSOPHY). Examples are: What is the correct analysis of the notion of truth? What is a proposition, and how is it related to the sentence in which it is expressed? Are there some kinds of sound reasoning that are neither deductive nor inductive? Fortunately, it is possible to learn to do formal logic without having satisfactory answers to such questions, just as it is possible to do mathematics without answering questions belonging to the philosophy of mathematics such as: Are numbers real objects or mental constructs?

THE PROPOSITIONAL CALCULUS

Basic features of PC. The simplest and most basic branch of logic is the propositional calculus, hereafter called PC, named from the fact that it deals only with complete, unanalyzed propositions and certain combinations into which they enter. Various notations for PC are used in the literature. In that used here the symbols employed in PC comprise first, variables (for which the letters *p, q, r,* . . . are used, with or without numerical subscripts); second, operators (for which the symbols $\sim$, $\cdot$, $\vee$, $\supset$, $\equiv$ are employed); and third, brackets or parentheses. The rules for constructing formulas are deferred to *Formation rules for PC;* and the intended interpretations—*i.e.,* the meanings to be given to these symbols—are indicated here immediately: The variables are to be viewed as representing unspecified propositions, or as marking the places in formulas into which propositions, and only propositions, may be inserted. (This is sometimes expressed by saying that they *range over* propositions, or that they take propositions as their *values.*) Hence they are often called *propositional variables.* It is assumed that every proposition is either true or false and that no proposition is both true and false. Truth and falsity are said to be the *truth values* of propositions. The function of an operator is to form a new proposition from one or more given propositions, called the *arguments* of the operator. The operators $\sim$, $\cdot$, $\vee$, $\supset$, and $\equiv$ correspond respectively to the English expressions "not," "and," "or," "if . . . then" (or "implies"), and "is equivalent to," when these are used in the following senses:

Fundamental definitions

(i) Given a proposition p, then $\sim p$ ("not p") is to count as false when p is true and true when p is false; $\sim$ (when thus interpreted) is known as the *negation sign* and $\sim p$ as the *negation* of p.

(ii) Given any two propositions, p and q, then $p \cdot q$ ("p and q") is to count as true when p and q are both true and as false in all other cases (viz., when p is true and q false, when p is false and q true, and when p and q are both false); $p \cdot q$ is said to be the *conjunction* of p and q; $\cdot$ is known as the *conjunction sign* and its arguments (p,q) as *conjuncts.*

(iii) $p \vee q$ ("p or q") is to count as false when p and q are both false and true in all other cases; thus it represents the assertion that at least one of p and q is true. $p \vee q$ is known as the *disjunction* of p and q; $\vee$ is the *disjunction sign* and its arguments (p,q) are known as *disjuncts.*

(iv) $p \supset q$ ("if p [then] q" or "p [materially] implies q") is to count as false when p is true and q is false, and as true in all other cases; hence it has the same meaning as "either not-p or q," or as "not both; p and not-q." The symbol $\supset$ is known as the (*material*) *implication sign,* the first argument as the *antecedent,* and the second as the *consequent;* $q \supset p$ is known as the *converse* of $p \supset q$.

(v) Finally, $p \equiv q$ ("p is [materially] equivalent to q" or "p if and only if q") is to count as true when p and q have the same truth value (*i.e.,* either when both are true or when both are false), and false when they have different truth values; the arguments of $\equiv$ (the [*material*] *equivalence sign*) are called *equivalents.*

Brackets are used to indicate grouping; they make it possible to distinguish, for example, between $p \cdot (q \vee r)$ ("Both p and either-q-or-r") and $(p \cdot q) \vee r$ ("Either both-p-and-q or r"). Precise rules for bracketing are given below.

The above account is meant solely as an explanation of the meaning to be attached to the PC operators in logic; it is not intended as an explanation or theory of the meaning in everyday speech of the English expressions by which they have here been roughly translated. Reading $p \supset q$ as "If p,q," for example, is made convenient by the fact that "if" is often used in English in a "material-implication" sense (someone who asserted "If it isn't raining, it's snowing" would usually be held to have spoken truly *unless* the antecedent were true but the consequent false; *i.e.,* unless it was neither raining nor snowing); but no claim is thereby being made that this is the only standard use of "if." Other uses of "if" have in fact led to the construction of various other systems of logic (see below *Alternative systems of modal logic* and also *Applied logic*) and have also raised important questions in the philosophy of logic; but these are irrelevant to the understanding of the relation expressed by $\supset$ in PC, which is one of fundamental importance for logic.

All PC operators take propositions as their arguments, and the result of applying them is also in each case a proposition. For this reason they are sometimes called *proposition-forming operators on propositions,* or more briefly *propositional connectives.* An operator that, like $\sim$, requires only a single argument is known as a *monadic* operator; operators that, like all the others listed, require two arguments are known as *dyadic.*

Truth functionality

All PC operators also have the following important characteristic: given the truth values of the arguments, the truth value of the proposition formed by them and the operator is determined in every case. An operator that has this characteristic is known as a *truth-functional* operator, and a proposition formed by such an operator is called a *truth function* of the operator's argument(s). The truth functionality of the PC operators is clearly brought out by summarizing the above account of them in Table 1. In

Table 1: Truth Table for Most Common Operators

monadic operator		dyadic operators					
p	$\sim p$	p	q	$p \cdot q$	$p \vee q$	$p \supset q$	$p \equiv q$
1	0	1	1	1	1	1	1
0	1	1	0	0	1	0	0
		0	1	0	1	1	0
		0	0	0	0	1	1

it, "true" is abbreviated by "1" and "false" by "0," and to the left of the vertical line are tabulated all possible combinations of truth values of the operators' arguments. The columns of 1s and 0s under the various truth functions indicate their truth values for each of the cases; these columns are known as the *truth tables* of the relevant operators. It should be noted that any column of four 1s and/or 0s or both will specify a dyadic truth-functional operator. Because there are precisely 2^4 (*i.e.,* 16) ways of forming a string of four symbols each of which is to be either 1 or 0 (1111, 1110, 1101, . . . , 0000), there are 16 such operators in all; the four that are listed here are only the four most generally useful ones.

Well-formed formulas (*wff*'s)

Formation rules for PC. In any system of logic it is necessary to specify which sequences of symbols are to count as acceptable formulas—or, as they are usually called, *well-formed formulas* (*wff*'s). Rules that specify this are called *formation rules.* From an intuitive point of view it is desirable that the *wff*'s of PC be just those sequences of PC symbols that, in terms of the interpretation given above, make sense and are unambiguous; and this can be ensured by stipulating that the *wff*'s of PC are to be all of those expressions constructed in accordance with the following PC formation rules, and only these:

FR1. A variable standing alone is a *wff*
FR2. If α is a *wff,* so is $\sim\alpha$
FR3. If α and β are *wff*'s, $(\alpha \cdot \beta)$, $(\alpha \vee \beta)$, $(\alpha \supset \beta)$, and $(\alpha \equiv \beta)$ are *wff*'s.

(In these rules α and β are variables representing arbitrary formulas of PC. They are not themselves symbols of PC but are used in discoursing about PC. Such variables are known as *metalogical* variables. For further explanation, see below *Metalogic.*) It should be noted that the rules, though designed to ensure unambiguous sense for the *wff*'s of PC under the intended interpretation, are themselves stated without any reference to interpretation and in such a way that there is an effective procedure for determining, again without any reference to interpretation, whether any

arbitrary string of symbols is a *wff* or not. (An *effective* procedure is one that is "mechanical" in nature and can always be relied on to give a definite result in a finite number of steps. The notion of effectiveness plays an important role in formal logic.)

Examples of *wff*'s are: p; $\sim q$; $\sim(p \cdot q)$, "not both p and q"; $[\sim p \vee (q \equiv p)]$, "Either not p or else q is equivalent to p."

For greater ease in writing or reading formulas the formation rules are often relaxed in various ways. The following relaxations are common:

(a) Brackets enclosing a complete formula may be omitted

(b) The typographical style of brackets may be varied within a formula, to make the pairing of brackets more evident to the eye

(c) Conjunctions and disjunctions may be allowed to have more than two arguments; *e.g.*, $p \cdot (q \supset r) \cdot \sim r$ may be written instead of $[p \cdot (q \supset r)] \cdot \sim r$. (The conjunction $p \cdot q \cdot r$ is then interpreted to mean that p, q, and r are all true, $p \vee q \vee r$ to mean that at least one of p, q, and r is true, and so forth.)

Validity in PC. Given the standard interpretation, a *wff* of PC becomes a proposition, true or false, when all of its variables are replaced by actual propositions. Such a *wff* is therefore a proposition form in the sense explained above and hence is valid if and only if all its instances are true propositions. A *wff* of which all instances are false is said to be *unsatisfiable* or *inconsistent;* and one with some true and some false instances is said to be *contingent.*

Decision procedure with truth tables

An important problem for any logical system is the *decision problem* for the class of valid *wff*'s of that system (sometimes simply called the decision problem for the system). This is the problem of finding an effective procedure, in the sense explained in *Formation rules for PC,* for testing the validity of any *wff* whatsoever of the system. Such a procedure is called a *decision procedure.* For some systems a decision procedure can be found; the decision problem for that system is then said to be *solvable,* and the system is said to be a *decidable* one. For other systems it can be proved that no decision procedure is possible; the decision problem for that system is then said to be *unsolvable,* and the system is said to be an *undecidable* one.

PC is a decidable system. In fact, several decision procedures for it are known. Of these the simplest and most important theoretically (though not always the easiest to apply in practice) is the method of *truth tables,* which will now be explained briefly. Since all of the operators in a *wff* of PC are truth-functional, in order to discover the truth value of any instance of such a *wff,* it is unnecessary to consider anything but the truth values of the propositions replacing the variables. In other words, the assignment of a truth value to each of the variables in a *wff* uniquely determines a truth value for the whole *wff.* Since there are only two truth values and each *wff* contains only a finite number of variables, there are only a finite number of truth-value assignments to the variables to be considered (if there are n distinct variables in the *wff,* there are 2^n such assignments); these can easily be systematically tabulated. For each of these assignments the truth tables for the operators then enable one to calculate the resulting truth value of the whole *wff;* and if and only if this truth value is *truth* in each case, is the *wff* valid. As an example, $[(p \supset q) \cdot r] \supset [(\sim r \vee p) \supset q]$ may be tested for validity. This formula states that "If one proposition implies a second one, and a certain third proposition is true, then if either that third proposition is false or the first is true, the second is true."

The calculation is shown in Table 2. As before, 1 represents truth and 0 falsity. Since the *wff* contains three variables, there are 2^3 (*i.e.,* 8) different assignments to the variables to be considered, which therefore generate the eight lines of the table. These assignments are tabulated to the left of the vertical line. The numbers in parentheses at the foot indicate the order in which the steps (from 1 through 6) are to be taken in determining the truth values (1 or 0) to be entered in the table. Thus column (1), falling under the symbol $\supset$, sets down the values of $p \supset q$ for each assignment, obtained from the columns under p and q by the truth table for $\supset$; column (2), for $(p \supset q) \cdot r$, is then obtained by employing the values in column (1) together with those in the column under r by use of the truth table for $\cdot$; . . . until finally column (6), which gives the values for the whole *wff,* is obtained from columns (2) and (5). This column is called the truth table of the whole *wff.* Since it consists entirely of 1s, it shows that the *wff* is true for every assignment given to the variables and is therefore valid. A *wff* for which the truth table consists entirely of 0s is inconsistent, and a *wff* for which the truth table contains at least one 1 and at least one 0 is contingent. It follows from the formation rules and from the fact that an initial truth table has been specified for each operator that a truth table can be constructed for any given *wff* of PC.

Table 2: Test for Validity by Truth Table

p	q	r	$[(p \supset q) \cdot r] \supset [(\sim r \vee p) \supset q]$					
1	1	1	1	1	1	0	1	1
1	1	0	1	0	1	1	1	1
1	0	1	0	0	1	0	1	0
1	0	0	0	0	1	1	1	0
0	1	1	1	1	1	0	0	1
0	1	0	1	0	1	1	1	1
0	0	1	1	1	1	0	0	1
0	0	0	1	0	1	1	1	0
			(1)	(2)	(6)	(3)	(4)	(5)

Among the more important valid *wff*'s of PC are those of Table 3:

Table 3: Some Valid Formulas of Propositional Calculus

law	formula
Law of Identity	$p \equiv p$
Law of Double Negation	$p \equiv \sim\sim p$
Law of Excluded Middle	$p \vee \sim p$
Law of Noncontradiction	$\sim(p \cdot \sim p)$
De Morgan Laws	$(p \cdot q) \equiv \sim(\sim p \vee \sim q)$ $(p \vee q) \equiv \sim(\sim p \cdot \sim q)$
Commutative Laws	$(p \vee q) \equiv (q \vee p)$ $(p \cdot q) \equiv (q \cdot p)$
Associative Laws	$[(p \vee q) \vee r] \equiv [p \vee (q \vee r)]$ $[(p \cdot q) \cdot r] \equiv [p \cdot (q \cdot r)]$
Law of Transposition	$(p \supset q) \equiv (\sim q \supset \sim p)$
Distributive Laws	$[p \cdot (q \vee r)] \equiv [(p \cdot q) \vee (p \cdot r)]$ $[p \vee (q \cdot r)] \equiv [(p \vee q) \cdot (p \vee r)]$
Law of Permutation	$[p \supset (q \supset r)] \equiv [q \supset (p \supset r)]$
Law of Syllogism	$(p \supset q) \supset [(q \supset r) \supset (p \supset r)]$
Law of Importation	$[p \supset (q \supset r)] \supset [(p \cdot q) \supset r]$
Law of Exportation	$[(p \cdot q) \supset r] \supset [p \supset (q \supset r)]$

Construction of valid *wff*'s

All of the above *wff*'s can be shown to be valid by a mechanical application of the truth-table method. They can also be seen to express intuitively sound general principles about propositions. For instance, because "not (. . . or—)" can be rephrased as "neither . . . nor—," the first De Morgan law can be read as: "Both p and q if and only if neither not-p nor not-q"; thus it expresses the principle that two propositions are jointly true if and only if neither of them is false.

Whenever, as is the case in most of the examples given, a *wff* of the form $\alpha \equiv \beta$ is valid, the corresponding *wff*'s $\alpha \supset \beta$ and $\beta \supset \alpha$ are also valid. For instance, because $(p \cdot q) \equiv \sim(\sim p \vee \sim q)$ is valid, so are $(p \cdot q) \supset \sim(\sim p \vee \sim q)$ and $\sim(\sim p \vee \sim q) \supset (p \cdot q)$.

Moreover, although $p \supset q$ does not mean that q can be deduced from p, yet whenever a *wff* of the form $\alpha \supset \beta$ is *valid,* the inference form "α, therefore β" is likewise valid. This fact is easily seen from the fact that $\alpha \supset \beta$ means the same as "Not both: α and not-β"; for, as was noted in the *Introduction,* whenever the latter is a valid proposition form, "α, therefore β" is a valid inference form.

Let α be any *wff.* If any variable in it is now uniformly replaced by some *wff,* the resulting *wff* is called a *substitution-instance of* α. Thus $[p \supset (q \vee \sim r)] \equiv [\sim(q \vee \sim r) \supset \sim p]$ is a substitution-instance of $(p \supset q) \equiv (\sim q \supset \sim p)$, obtained from it by replacing q uniformly by $(q \vee \sim r)$. It is an important principle that whenever a *wff* is valid so is every substitution-instance of it (the *rule of* [uniform] *substitution*).

A further important principle is the *rule of substitution of equivalents.* Two *wff*'s, α and β, are said to be *equivalents* when $\alpha \equiv \beta$ is valid. (The *wff*'s α and β are equivalents

if and only if they have identical truth tables.) The rule states that if any part of a *wff* is replaced by an equivalent of that part, the resulting *wff* and the original are also equivalents. Such replacements need not be uniform. The application of this rule is said to make an *equivalence transformation.*

Interdefinability of operators. The rules that have just been stated would enable the first De Morgan law listed in Table 3 to transform any *wff* containing any number of occurrences of $\cdot$ into an equivalent *wff* in which $\cdot$ does not appear at all but in place of it certain complexes of $\sim$ and $\vee$ arise. Similarly, since $\sim p \vee q$ has the same truth table as $p \supset q$, $(p \supset q) \equiv (\sim p \vee q)$ is valid, and any *wff* containing $\supset$ can therefore be transformed into an equivalent *wff* containing $\sim$ and $\vee$ but not $\supset$. And since $(p \equiv q) \equiv [(p \supset q) \cdot (q \supset p)]$ is valid, any *wff* containing $\equiv$ can be transformed into an equivalent containing $\supset$ and $\cdot$ but not $\equiv$ —and thus in turn by the previous steps it can be further transformed into one containing $\sim$ and $\vee$ but neither $\equiv$ nor $\supset$ nor $\cdot$. Thus for every wff of PC there is an equivalent *wff,* expressing precisely the same truth function, in which the only operators are $\sim$ and $\vee$, though the meaning of this *wff* will usually be much less clear than that of the original.

Alternative primitives

An alternative way of presenting PC, therefore, is to begin with the operators $\sim$ and $\vee$ only and to define the others in terms of these. The operators $\sim$ and $\vee$ are then said to be *primitive.* If "$=_{Df}$" is used to mean "is defined as," then the relevant definitions can be set down as follows:

$$(\alpha \cdot \beta) =_{Df} \sim(\sim\alpha \vee \sim\beta)$$
$$(\alpha \supset \beta) =_{Df} (\sim\alpha \vee \beta)$$
$$(\alpha \equiv \beta) =_{Df} [(\alpha \supset \beta) \cdot (\beta \supset \alpha)]$$

in which α and β are any *wff*'s of PC. These definitions are not themselves *wff*'s of PC, nor is $=_{Df}$ a symbol of PC; they are metalogical statements about PC, used to introduce the new symbols $\cdot$, $\supset$, and $\equiv$ into the system. If PC is regarded as a purely uninterpreted system, the expression on the left in a definition is simply a convenient abbreviation of the expression on the right. If, however, PC is thought of as having its standard interpretation, the meanings of $\sim$ and $\vee$ will first of all have been stipulated by truth tables, and then the definitions will lay it down that the expression on the left is to be understood as having the same meaning (*i.e.,* the same truth table) as the expression on the right. It is easy to check that the truth tables obtained in this way for $\cdot$, $\supset$, and $\equiv$ are precisely the ones that were originally stipulated for them.

An alternative to taking $\sim$ and $\vee$ as primitive is to take $\sim$ and $\cdot$ as primitive and to define $(\alpha \vee \beta)$ as $\sim(\sim\alpha \cdot \sim\beta)$, to define $(\alpha \supset \beta)$ as $\sim(\alpha \cdot \sim\beta)$, and to define $(\alpha \equiv \beta)$ as before. Yet another possibility is to take $\sim$ and $\supset$ as primitive and to define $(\alpha \vee \beta)$ as $(\sim\alpha \supset \beta)$, $(\alpha \cdot \beta)$ as $\sim(\alpha \supset \sim\beta)$, and $(\alpha \equiv \beta)$ as before. In each case, precisely the same *wff*'s that were valid in the original presentation of the system are still valid.

Axiomatization of PC. The basic idea of constructing an *axiomatic system* is that of choosing certain *wff*'s (known as *axioms*) as starting points and giving rules for deriving further *wff*'s (known as *theorems*) from them. Such rules are called *transformation rules.* Sometimes the word "theorem" is used to cover axioms as well as theorems; the word "thesis" is also used for this purpose.

Axiomatic basis and interpretation

An *axiomatic basis* consists of:

(1) A list of primitive symbols, together with any definitions that may be thought convenient;
(2) A set of formation rules, specifying which sequences of symbols are to count as *wff*'s;
(3) A list of *wff*'s selected as axioms;
(4) A set of (one or more) transformation rules, which enable new *wff*'s (theorems) to be obtained by performing certain specified operations on axioms or previously obtained theorems.

Definitions, where they occur, can function as additional transformation rules, to the effect that, if in any theorem any expression of the form occurring on one side of a definition is replaced by the corresponding expression of the form occurring on the other side, the result is also to count as a theorem. A *proof* or *derivation* of a *wff* α in an axiomatic system S is a sequence of *wff*'s of which the last is α itself and each *wff* in the sequence is either an axiom of S or is derived from some axiom(s) or some already-derived theorem(s) or both by one of the transformation rules of S. A *wff* is a theorem of S if and only if there is a proof of it in S.

Care is usually taken, in setting out an axiomatic basis, to avoid all reference to interpretation. It must be possible to tell purely from the construction of a *wff* whether it is an axiom or not. Moreover, the transformation rules must be so formulated that there is an effective way of telling whether any purported application of them is a correct application or not and hence whether a purported proof of a theorem really is a proof or not. An axiomatic system will then be a purely formal structure, on which any one of a number of interpretations, or none at all, may be imposed without affecting the question of which *wff*'s are theorems. Normally, however, an axiomatic system is constructed with a certain interpretation in mind; the transformation rules are so formulated that under that interpretation they are *validity-preserving* (*i.e.,* the results of applying them to valid *wff*'s are always themselves valid *wff*'s); and the chosen axioms either are valid *wff*'s or are expressions of rinciples of which it is desired to explore the consequences.

Properties of an axiomatic system: PM

Probably the best known axiomatic system for PC is the following one, which, since it is derived from *Principia Mathematica,* by Alfred North Whitehead and Bertrand Russell (published in 1910), is often called PM:

Primitive symbols: $\sim$, $\vee$, (,), and an infinite set of variables, *p, q, r,* . . . (with or without numerical subscripts). *Definitions* of $\cdot$, $\supset$, $\equiv$ as above (*Interdefinability of operators*). *Formation rules* as above (*Formation rules for PC*), except that Rule 3 can be abbreviated to "If α and β are *wff*'s, $(\alpha \vee \beta)$ is a *wff,*" since $\cdot$, $\supset$, and $\equiv$ are not primitive. *Axioms:*

(1) $(p \vee p) \supset p$;
(2) $q \supset (p \vee q)$;
(3) $(p \vee q) \supset (q \vee p)$;
(4) $(q \supset r) \supset [(p \vee q) \supset (p \vee r)]$.

Axiom (4) can be read "If q implies r, then, if either p or q, either p or r." *Transformation rules:* (1) The result of uniformly replacing any variable in a theorem by any *wff* is a theorem (*rule of substitution*); and (2) if α and $(\alpha \supset \beta)$ are theorems, then β is a theorem (*rule of detachment,* or *modus ponens*).

Relative to a given criterion of validity, an axiomatic system is *sound* if every theorem is valid, and it is *complete* (or, more specifically, *weakly complete*) if every valid *wff* is a theorem. The axiomatic system PM can be shown to be both sound and complete relative to the criterion of validity given in *Validity in PC.*

An axiomatic system is *consistent* if, whenever a *wff* α is a theorem, $\sim\alpha$ is not a theorem. (In terms of the standard interpretation, this means that no pair of theorems can ever be derived one of which is the negation of the other.) It is *strongly complete* if the addition to it (as an extra axiom) of any *wff* whatever that is not already a theorem would make the system inconsistent. Finally, an axiom or transformation rule is *independent* (in a given axiomatic system) if it cannot be derived from the remainder of the basis (or—what comes to the same thing—if its omission from the basis would make the derivation of certain theorems impossible). It can, moreover, be shown that PM is consistent and strongly complete and that each of its axioms and transformation rules is independent.

A considerable number of other axiomatic bases for PC, each having all of the above properties, are known. The task of proving that they have these properties belongs to metalogic.

Axiom schemata

In some standard expositions of formal logic, the place of axioms is taken by *axiom schemata,* which, instead of presenting some particular *wff* as an axiom, lay it down that any *wff* of a certain form is an axiom. For example, in place of Axiom (1) in PM one might have the axiom schema "Every *wff* of the form $(\alpha \vee \alpha) \supset \alpha$ is an axiom"; and analogous schemata can be substituted for the other axioms. The number of axioms would then become infinite; but, on the other hand, the *rule of substitution* would no longer be needed and *modus ponens* could be the only

transformation rule. This method makes no difference to the theorems that can be derived; but in some branches of logic (though not in PC), it is simpler to work with axiom schemata rather than with particular axioms and substitution rules. Having an infinite number of axioms causes no trouble provided that there is an effective way of telling whether a *wff* is an axiom or not.

Special systems of PC. *Partial systems of PC.* Various propositional calculi have been devised to express a narrower range of truth functions than those of PC as expounded above. Of these, the one that has been most fully studied is the Pure Implicational Calculus (PIC), in which the only operator is $\supset$, and the *wff*'s are precisely those *wff*'s of PC that can be built up from variables, $\supset$, and brackets alone. Formation rules 2 and 3 of *Formation rules for PC* are therefore replaced by the rule that if α and β are *wff*'s, $(\alpha \supset \beta)$ is a *wff*. As in ordinary PC, $p \supset q$ is interpreted as "p materially implies q"; *i.e.*, as true except when p is true but q false. The truth-table test of validity can then be straightforwardly applied to *wff*'s of PIC.

The task of axiomatizing PIC is that of finding a set of valid *wff*'s, preferably few in number and relatively simple in structure, from which all other valid *wff*'s of the system can be derived by straightforward transformation rules. The best-known basis, which was formulated in 1930, has the transformation rules of *substitution* and *modus ponens* (as in PM), and the axioms:

(1) $p \supset (q \supset p)$
(2) $[(p \supset q) \supset p] \supset p$
(3) $(p \supset q) \supset [(q \supset r) \supset (p \supset r)]$

Axioms (1) and (3) are closely related to axioms (2) and (4) of PM respectively (see above *Axiomatization of PC*). It can be shown that the basis is complete and that each axiom is independent.

Under the standard interpretation the above axioms can be thought of as expressing the following principles: (1) "If a proposition p is true, then if some arbitrary proposition q is true, p is (still) true." (2) "If the fact that a certain proposition p implies some arbitrary proposition q implies that p itself is true, then p is (indeed) true." (3) "If one proposition (p) implies a second (q), then if that second proposition implies a third (r), the first implies the third." The completeness of the basis is, however, a formal matter, not dependent on these or any other readings of the formulas.

An even more economical complete basis for PIC contains the same transformation rules but the sole axiom

$$[(p \supset q) \supset r] \supset [(r \supset p) \supset (s \supset p)].$$

It has been proved that this is the shortest possible single axiom that will give a complete basis for PIC with these transformation rules.

Since PIC contains no negation sign, the previous account of consistency is not significantly applicable to it. Alternative accounts of consistency have, however, been proposed, according to which a system is consistent (a) if no *wff* consisting of a single variable is a theorem, or (b) if not every *wff* is a theorem. The bases stated are consistent in these senses.

Nonstandard versions of PC. Qualms have sometimes been expressed about the intuitive soundness of some formulas that are valid in "orthodox" PC, and these qualms have led some logicians to construct a number of propositional calculi that deviate in various ways from PC as expounded above.

The true-false dichotomy: its problems

Underlying ordinary PC is the intuitive idea that every proposition is either true or false, an idea that finds its formal expression in the stipulation that variables shall have two possible values only, viz., 1 and 0. (For this reason the system is often called the *two-valued* propositional calculus.) This idea has been challenged on various grounds. Following a suggestion made by Aristotle, some logicians have maintained that propositions about those events in the future that may or may not come to pass are neither true nor false but "neuter" in truth value. Aristotle's own example, which has been discussed frequently down to the present day, is "There will be a sea-battle tomorrow." It has also been maintained that for propositions with subjects that do not have anything actual corresponding to them—such as "The present King of France is wise" (assuming that France has no king) or "All John's children are asleep" (assuming that John has no children)—the question of truth or falsity "does not arise." Yet another view is that a third truth value (say "half-truth") ought to be recognized as existing intermediate between truth and falsity; thus it has been advanced that certain familiar states of the weather make the proposition "It is raining" neither definitely true nor definitely false but something in between the two.

Three-valued logics

The issues raised by the above examples no doubt differ significantly; but they all suggest a threefold rather than a twofold division of propositions and hence the possibility of a logic in which the variables may take any of three values (say 1, ½, and 0), with a consequent revision of the standard PC account of validity. Several such three-valued logics have been constructed and investigated; a brief account will be given here of one of them, in which the most natural interpretation of the additional value (½) is as "half-true," with 1 and 0 representing truth and falsity as before. The formation rules are as they were for orthodox PC, but the meaning of the operators is extended to cover cases in which at least one argument has the value ½ by

Table 4: Truth Values for Common Operators in a Three-Valued Logic

		negation ($\sim p$)	conjunction ($p \cdot q$) q			disjunction ($p \vee q$) q			implication ($p \supset q$) q			equivalence ($p \equiv q$) q		
			1	½	0	1	½	0	1	½	0	1	½	0
	1	0	1	½	0	1	1	1	1	½	0	1	½	0
p	½	½	½	½	0	1	½	½	1	1	½	½	1	½
	0	1	0	0	0	1	½	0	1	1	1	0	½	1

the following five entries in Table 4. (Adopting one of the three values of the first argument, p, given in the leftmost column [1, ½, or 0] and, for the dyadic operators, one of the three values of the second, q, in the top row—above the line—one then finds the value of the whole formula by reading across for p and down for q.) It will be seen that these tables, which are from Łukasiewicz, are the same as the ordinary two-valued ones when the arguments have the values 1 and 0. The other values are intended to be intuitively plausible extensions of the principles underlying the two-valued calculus to cover the cases involving half-true arguments. Clearly, these tables enable a person to calculate a determinate value (1, ½, or 0) for any *wff*, given the values assigned to the variables in it; and a *wff* is valid in this calculus if it has the value 1 for every assignment to its variables. Since the values of formulas when the variables are assigned only the values 1 and 0 are the same as in ordinary PC, every *wff* that is valid in the present calculus is also valid in PC. Some *wff*'s that are valid in PC are, however, now no longer valid. An example is $(p \vee \sim p)$, which, when p has the value ½, also has the value ½. This reflects the idea that if one admits the possibility of a proposition's being half-true, he can no longer hold of every proposition without restriction that either it or its negation is true.

Validity of *wff*'s in three-valued logic

Given the above tables for the operators, it is possible to take $\sim$ and $\supset$ as primitive and to define $(\alpha \vee \beta)$ as $[(\alpha \supset \beta) \supset \beta]$—though not as $(\sim\alpha \supset \beta)$ as in ordinary PC; $(\alpha \cdot \beta)$ as $\sim(\sim\alpha \vee \sim\beta)$; and $(\alpha \equiv \beta)$ as $[(\alpha \supset \beta) \cdot (\beta \supset \alpha)]$. With these definitions as given, all valid *wff*'s constructed from variables and $\sim$, $\cdot$, $\vee$, $\supset$, and $\equiv$ can be derived by *substitution* and *modus ponens* from the four axioms:

(1) $p \supset (q \supset p)$;
(2) $(p \supset q) \supset [(q \supset r) \supset (p \supset r)]$;
(3) $[(p \supset \sim p) \supset p] \supset p$;
(4) $(\sim p \supset \sim q) \supset (q \supset p)$.

Other three-valued logics can easily be constructed. For example, the above tables might be modified so that $\sim$½, ½ $\supset$ 0, ½ $\equiv$ 0, and 0 $\equiv$ ½ all have the value 0 instead of ½ as before, leaving everything else unchanged. The same definitions are then still possible, but the list of valid formulas is different; *e.g.*, $\sim\sim p \supset p$, which was previously

valid, now has the value ½ when p has the value ½. This system can also be successfully axiomatized. Other calculi with more than three values can also be constructed along analogous lines.

Intuitionistic calculus

Other nonstandard calculi have been constructed by beginning with an axiomatization instead of a definition of validity. Of these the best known is the *intuitionistic calculus,* devised by Arend Heyting, one of the chief representatives of the Intuitionist school of mathematicians, a group of theorists that denies the validity of certain types of proof used in classical mathematics. At least in certain contexts, members of this school regard the demonstration of the falsity of the negation of a proposition (a proof by *reductio ad absurdum*) as insufficient to establish the truth of the proposition in question. Thus they regard $\sim\sim p$ as an inadequate premise from which to deduce p and hence do not accept the validity of the *law of double negation* in the form $\sim\sim p \supset p$. They do, however, regard a demonstration that p is true as showing that the negation of p is false and hence accept $p \supset \sim\sim p$ as valid. For somewhat similar reasons, these mathematicians also refuse to accept the validity of arguments based on the *law of excluded middle* $(p \vee \sim p)$. The intuitionistic calculus aims at presenting in axiomatic form those and only those principles of propositional logic that are accepted as sound in intuitionist mathematics. In this calculus, $\sim$, $\cdot$, $\vee$, and $\supset$ are all primitive, the transformation rules as before are *substitution* and *modus ponens,* and the axioms are the following:

(1) $p \supset (p \cdot p)$
(2) $(p \cdot q) \supset (q \cdot p)$
(3) $(p \supset q) \supset [(p \cdot r) \supset (q \cdot r)]$
(4) $[(p \supset q) \cdot (q \supset r)] \supset (p \supset r)$
(5) $p \supset (q \supset p)$
(6) $[p \cdot (p \supset q)] \supset q$
(7) $p \supset (p \vee q)$
(8) $(p \vee q) \supset (q \vee p)$
(9) $[(p \supset r) \cdot (q \supset r)] \supset [(p \vee q) \supset r]$
(10) $\sim p \supset (p \supset q)$
(11) $[(p \supset q) \cdot (p \supset \sim q)] \supset \sim p$

From this basis neither $p \vee \sim p$ nor $\sim\sim p \supset p$ can be derived, though $p \supset \sim\sim p$ can. In this respect this calculus resembles the second of the three-valued logics described above. It is, however, not possible to give a truth-table account of validity—no matter how many values are used—that will bring out as valid precisely those *wff*'s that are theorems of the intuitionistic calculus and no others.

Natural deduction method in PC. PC is often presented by what is known as the method of *natural deduction.* Essentially this consists of a set of rules for drawing conclusions from hypotheses (assumptions, premises) represented by *wff*'s of PC and thus for constructing valid inference forms; but it also provides a method of deriving from these inference forms valid proposition forms, and in this way it is analogous to the derivation of theorems in an axiomatic system. One such set of rules is presented in Table 5 (and there are various other sets that yield the same results):

Table 5: Sample Set of Rules for the Natural Deduction Method in Propositional Calculus

rule	given	one may then conclude
1. Modus ponens	α and $\alpha \supset \beta$	β
2. Modus tollens	$\sim\beta$ and $\alpha \supset \beta$	$\sim\alpha$
3. Double negation	α	$\sim\sim\alpha$
	$\sim\sim\alpha$	α
4. Conjunction introduction	α and β	$\alpha \cdot \beta$
5. Conjunction elimination	$\alpha \cdot \beta$	α and also β
6. Disjunction introduction	either α or β separately	$\alpha \vee \beta$
7. Disjunction elimination	$\alpha \vee \beta$, a derivation of γ from α, and a derivation of γ from β	γ
8. Conditional proof	a derivation of β from the hypothesis α (perhaps with the help of other hypotheses)	$\alpha \supset \beta$ as a conclusion from these other hypotheses (if any)
9. Reductio ad absurdum	a derivation of $\beta \cdot \sim\beta$ from the hypothesis α (perhaps with the help of other hypotheses)	$\sim\alpha$ as a conclusion from these other hypotheses (if any)

Rules of inference and their application

A natural deduction proof is a sequence of *wff*'s beginning with one or more *wff*'s as hypotheses; fresh hypotheses may also be added at any point in the course of a proof. The rules may be applied to any *wff* or group of *wff*'s, as appropriate, that have already occurred in the sequence. In the case of rules 1–7, the conclusion is said to *depend on* all of those hypotheses that have been used in the series of applications of the rules that have led to this conclusion; *i.e.,* it is claimed simply that the conclusion follows from these hypotheses, not that it holds in its own right. An application of rule 8 or rule 9, however, reduces by one the number of hypotheses on which the conclusion depends; and a hypothesis so eliminated is said to be a *discharged* hypothesis. In this way a *wff* may be reached that depends on no hypotheses at all. Such a *wff* is a *theorem.* It can be shown that the theorems derivable by the rules stated above—together with the definition of $\alpha \equiv \beta$ as $(\alpha \supset \beta) \cdot (\beta \supset \alpha)$—are precisely the valid *wff*'s of PC. A set of natural deduction rules yielding as theorems all of the valid *wff*'s of a system is *complete* (with respect to that system) in a sense obviously analogous to that in which an axiomatic basis was said to be complete in *Axiomatization of PC.*

Sample proof of a theorem

As an illustration, the formula $[(p \supset q) \cdot (p \supset r)] \supset [p \supset (q \cdot r)]$ will be derived as a theorem by the natural deduction method. (The sense of this formula is that, if a proposition $[p]$ implies each of two other propositions $[q,r]$, then it implies their conjunction.) Explanatory comments follow the proof.

1	(1)	$(p \supset q) \cdot (p \supset r)$	hypothesis
2	(2)	p	hypothesis
1	(3)	$p \supset q$	1, conjunction elimination
1	(4)	$p \supset r$	1, conjunction elimination
1,2	(5)	q	2,3, *modus ponens*
1,2	(6)	r	2,4, *modus ponens*
1,2	(7)	$q \cdot r$	5,6, conjunction introduction
1	(8)	$p \supset (q \cdot r)$	2,7,conditional proof
	(9)	$[(p \supset q) \cdot (p \supset r)] \supset [p \supset (q \cdot r)]$	1,8, conditional proof

The figures in parentheses immediately preceding the *wff*'s are simply for reference. To the right is indicated either that the *wff* is a hypothesis or that it is derived from the *wff*'s indicated by the rules stated. On the left are noted the hypotheses on which the *wff* in question depends (either the first or the second line of the derivation, or both). Note that since (8) is derived by *conditional proof* from hypothesis (2) and from (7), which is itself derived from hypotheses (1) and (2), (8) depends only on hypothesis (1), and hypothesis (2) is discharged. Similarly (9) depends on no hypotheses and is therefore a theorem.

By varying the above rules it is possible to obtain natural deduction systems corresponding to other versions of PC. For example, if the second part of the *double negation* rule is omitted and the rule is added that, given $\alpha \cdot \sim\alpha$, one may then conclude: β, it can be shown that the theorems then derivable are precisely the theorems of the Intuitionistic Calculus.

THE PREDICATE CALCULUS

Fundamental definitions

Propositions may also be built up, not out of other propositions but out of elements that are not themselves propositions. The simplest kind to be considered here are propositions in which a certain object or individual (in a wide sense) is said to possess a certain property or characteristic; examples are: "Socrates is wise" and "The number 7 is prime." Such a proposition contains two distinguishable parts: (1) an expression that names or designates an individual and (2) an expression, called a *predicate,* that stands for the property that the individual in question is said to possess. If $x, y, z, \ldots$ are used as individual variables (replaceable by names of individuals) and the symbols φ (phi), ψ (psi), χ (chi), ... as predicate variables (replaceable by predicates), the formula φx is used to express the form of the propositions in question.

Here x is said to be the *argument* of φ; a predicate (or predicate variable) with only a single argument is said to be a *monadic,* or *one-place,* predicate (variable). Predicates with two or more arguments stand not for properties of single individuals but for relations between individuals. Thus the proposition "Tom is a son of John" is analyzable into two names of individuals ("Tom" and "John") and a dyadic or two-place predicate ("is a son of"), of which they are the arguments; and the proposition is thus of the form φxy. Analogously, ". . . is between . . . and . . ." is a three-place predicate, requiring three arguments, and so on. In general, a predicate variable followed by any number of individual variables is a *wff* of the predicate calculus. Such a *wff* is known as an *atomic* formula, and the predicate variable in it is said to be of *degree n,* if n is the number of individual variables following it. The degree of a predicate variable is sometimes indicated by a superscript—*e.g.,* φxyz may be written as φ^3xyz; φ^3xy would then be regarded as not well formed. This practice is theoretically more accurate, but the superscripts are commonly omitted for ease of reading when no confusion is likely to arise.

Atomic formulas may be combined with truth-functional operators to give formulas such as $\varphi x \vee \psi y$ [Example: "Either the customer (x) is friendly (φ) or else John (y) is disappointed (ψ)"]; $\varphi xy \supset \sim\psi x$ [Example: "If the road (x) is above (φ) the flood line (y), then the road is not wet ($\sim\psi$)"]; and so on. Formulas so formed, however, are valid when and only when they are substitution-instances of valid *wff*'s of PC and hence in a sense do not transcend PC. More interesting formulas are formed by the use, in addition, of *quantifiers.* There are two kinds of quantifiers: *universal* quantifiers, written as "$(\forall__\ )$"—or often simply as "$(__\ )$"—where the blank is filled by a variable, which may be read "For all___ "; and *existential* quantifiers, written as "$(\exists__\ )$," which may be read "For some___ ," or "There is a ___ such that." ("Some" is to be understood as meaning "at least one.") Thus $(\forall x)\varphi x$ is to mean "For all x, x is φ," or more simply, "Everything is φ"; and $(\exists x)\varphi x$ is to mean "For some x, x is φ," or more simply, "Something is φ," or "There is a φ." Slightly more complex examples are: $(\forall x)(\varphi x \supset \psi x)$ for "Whatever is φ is ψ"; $(\exists x)(\varphi x \cdot \psi x)$ for "Something is both φ and ψ"; $(\forall x)(\exists y)\varphi xy$ for "Everything bears the relation φ to at least one thing"; and $(\exists x)(\forall y)\varphi xy$ for "There is something that bears the relation φ to everything." To take a concrete case, if φxy means "x loves y," and the values of x and y are taken to be human beings, then the last two formulas mean, respectively, "Everybody loves somebody" and "Somebody loves everybody."

Quantifiers ∀ and ∃

Intuitively, the notions expressed by the words "some" and "every" are connected in the following way: to assert that something has a certain property amounts to denying that everything lacks that property (for example, to say that something is white is to say that not everything is non-white); and similarly, to assert that everything has a certain property amounts to denying that there is something that lacks it. These intuitive connections are reflected in the usual practice of taking one of the quantifiers as primitive and defining the other in terms of it. Thus ∀ may be taken as primitive, and ∃ introduced by the definition:

$$(\exists \mathbf{a})\alpha =_{\text{Df}} \sim(\forall \mathbf{a})\sim\alpha$$

in which **a** is any variable and α is any *wff;* or alternatively, ∃ may be taken as primitive, and ∀ introduced by the definition:

$$(\forall \mathbf{a})\alpha =_{\text{Df}} \sim(\exists\, \mathbf{a})\sim\alpha$$

The lower predicate calculus. A predicate calculus in which the only variables that occur in quantifiers are individual variables is known as a *lower* (or *first-order*) predicate calculus. Various lower predicate calculi have been constructed. In the most straightforward of these, to which the most attention will be devoted in this discussion and which subsequently will be referred to simply as LPC, the *wff*'s can be specified as follows: Let the primitive symbols be (a) $x, y, \ldots$ (individual variables); (b) $\varphi, \psi, \ldots$, each of some specified degree (predicate variables); and (c) the symbols $\sim$, $\vee$, ∀, (, and). An infinite number of each type of variable can now be secured as before by the use of numerical subscripts. The symbols $\cdot$, $\supset$, and $\equiv$ are defined as in PC and ∃ as explained above. The formation rules are:

1. An expression consisting of a predicate variable of degree n followed by n individual variables is a *wff.*
2. If α is a *wff,* so is $\sim\alpha$.
3. If α and β are *wff*'s, so is $(\alpha \vee \beta)$.
4. If α is a *wff* and **a** is an individual variable, then $(\forall \mathbf{a})\ \alpha$ is a *wff.* (In such a *wff,* α is said to be the *scope* of the quantifier.)

Bound and free variables

If **a** is any individual variable and α is any *wff,* every occurrence of **a** in α is said to be *bound* (by the quantifiers) when occurring in the *wff*'s $(\forall \mathbf{a})\alpha$ and $(\exists \mathbf{a})\alpha$. Any occurrence of a variable that is not bound is said to be *free.* Thus in $(\forall x)(\varphi x \vee \varphi y)$—the x in φx is bound, since it occurs within the scope of a quantifier containing x, but y is free. In the *wff*'s of a lower predicate calculus, every occurrence of a predicate variable ($\varphi, \psi, \chi, \ldots$) is free. A *wff* containing no free individual variables is said to be a *closed wff* of LPC. If a *wff* of LPC is considered as a proposition form, instances of it are obtained by replacing all free variables in it by predicates or by names of individuals, as appropriate. A bound variable, on the other hand, indicates not a point in the *wff* where a replacement is needed but a point (so to speak) at which the relevant quantifier applies.

For example, in φx, in which both variables are free, each variable must be replaced appropriately if a proposition of the form in question (such as "Socrates is white") is to be obtained; but in $(\exists x)\varphi x$, in which x is bound, it is necessary only to replace φ by a predicate in order to obtain a complete proposition (*e.g.,* replacing φ by "is white" yields the proposition "Something is white").

Validity in LPC. Intuitively, a *wff* of LPC is valid if and only if all of its instances are true—*i.e.,* if and only if every result of replacing each of its free variables appropriately and uniformly is a true proposition. A formal definition of validity in LPC to express this intuitive notion more precisely can be given as follows: for any *wff* of LPC any number of LPC *models* can be formed. An LPC model has two elements: One is a set, D, of objects, known as a *domain. D* may contain as many or as few objects as one chooses, but it must contain at least one, and the objects may be of any kind. The other element, V, is a system of value assignments satisfying the following conditions. To each individual variable there is assigned some member of D (not necessarily a different one in each case). Assignments are next made to the predicate variables in the following way: if φ is monadic, there is assigned to it some subset of D (possibly the whole of D); intuitively this subset can be viewed as the set of all the objects in D that have the property φ. If φ is dyadic, there is assigned to it some set of ordered pairs (*i.e.,* pairs of objects of which one is marked out as the first and the other as the second) drawn from D; intuitively these can be viewed as all of the pairs of objects in D in which the relation φ holds between the first object in the pair and the second. In general, if φ is of degree n, there is assigned to it some set of ordered n-tuples (groups of n objects) of members of D. It is then stipulated that an atomic formula is to have the value 1 in the model if the members of D assigned to its individual variables form, in that order, one of the n-tuples assigned to the predicate variable in it; otherwise, it is to have the value 0. Thus in the simplest case, φx will have the value 1 if the object assigned to x is one of the set of objects that was assigned to φ; and if it is not, then φx will have the value 0. The values of truth functions are determined by the values of their arguments as in PC. Finally, the value of $(\forall x)\alpha$ is to be 1 if both (a) the value of α itself is 1 and (b) α would always still have the value 1 if a different assignment were made to x but all of the other assignments were left precisely as they were; otherwise $(\forall x)\alpha$ is to have the value 0. Since ∃ can be defined in terms of ∀, these rules cover all of the *wff*'s of LPC. A given *wff* may of course have the value 1 in some LPC models but the value 0 in others. But a valid *wff* of LPC may now be defined as one that has the value 1 in every LPC model. If 1 and 0 are viewed as representing truth and falsity, respectively, then validity is defined as truth in every model.

Value assignments in the domain of a model

Although the above definition of validity in LPC is quite precise, it does not yield, as did the corresponding definition of PC validity in terms of truth tables, an effective decision procedure. It can, indeed, be shown that no generally applicable decision procedure for LPC is possible—*i.e.,* that LPC is not a decidable system. This does not mean that it is never possible to prove that a given *wff* of LPC is valid—the validity of an unlimited number of such *wff*'s can in fact be demonstrated—but it does mean that in the case of LPC, unlike that of PC, no advance assurance can be given that no matter what *wff* is produced it will always be possible to determine whether it is valid or not (see also below *Model theory*).

Laws concerning quantifiers

Logical manipulations in LPC. The intuitive connections between *some* and *every* noted earlier are reflected in the fact that the following equivalences are valid:

$$(\exists x)\varphi x \equiv \sim(\forall x)\sim\varphi x$$
$$(\forall x)\varphi x \equiv \sim(\exists x)\sim\varphi x.$$

The former of these asserts that "Something is φ" is equivalent to "Not everything fails to be φ", and the latter that "Everything is φ" is equivalent to "Nothing fails to be φ" (literally, "It is not the case that there is something which is not φ"). These equivalences remain valid when φx is replaced by any *wff,* however complex; *i.e.,* for any *wff* α whatsoever,

$$(\exists x)\alpha \equiv \sim(\forall x)\sim\alpha$$

and

$$(\forall x)\alpha \equiv \sim(\exists x)\sim\alpha$$

are valid. Because the rule of substitution of equivalents can be shown to hold in LPC, it follows that $(\exists x)$ may be replaced anywhere in a *wff* by $\sim(\forall x)\sim$, or $(\forall x)$ by $\sim(\exists x)\sim$, and the resulting *wff* will be equivalent to the original. Similarly, because the law of double negation permits the deletion of a pair of consecutive negation signs, $\sim(\exists x)$ may be replaced by $(\forall x)\sim$, and $\sim(\forall x)$ by $(\exists x)\sim$.

These principles are easily extended to more complex cases. To say that there is a pair of objects satisfying a certain condition is equivalent to denying that every pair of objects fails to satisfy that condition; and to say that every pair of objects satisfies a certain condition is equivalent to denying that there is any pair of objects that fails to satisfy that condition. These equivalences are expressed formally by the validity, again for any *wff* α, of:

$$(\exists x)(\exists y)\alpha \equiv \sim(\forall x)(\forall y)\sim\alpha$$

and

$$(\forall x)(\forall y)\alpha \equiv \sim(\exists x)(\exists y)\sim\alpha$$

—and by the resulting replaceability anywhere in a *wff* of $(\exists x)(\exists y)$ by $\sim(\forall x)(\forall y)\sim$, or of $(\forall x)(\forall y)$ by $\sim(\exists x)(\exists y)\sim$.

Analogously, $(\exists x)(\forall y)$ can be replaced by $\sim(\forall x)(\exists y)\sim$ [*e.g.,* $(\exists x)(\forall y)$ (x loves y)—"There is someone who loves everyone"—is equivalent to $\sim(\forall x)(\exists y)\sim$($x$ loves y)—"It is not true of everyone that there is someone whom he does not love"]; $(\forall x)(\exists y)$ can be replaced by $\sim(\exists x)(\forall y)\sim$; and in general the following rule, covering sequences of quantifiers of any length, holds:

(i) If a *wff* contains an unbroken sequence of quantifiers, then the *wff* that results from replacing $\forall$ by $\exists$ and vice versa throughout that sequence and inserting or deleting $\sim$ at each end of it is equivalent to the original *wff.*

This may be called the *rule of quantifier transformation.* It reflects, in a generalized form, the intuitive connections between *some* and *every* that were noted above.

The following are also valid, again where α is any *wff*:

$$(\forall x)(\forall y)\alpha \equiv (\forall y)(\forall x)\alpha$$
$$(\exists x)(\exists y)\alpha \equiv (\exists y)(\exists x)\alpha$$

The extensions of these lead to the following rule:

(ii) If a *wff* contains an unbroken sequence either of universal or of existential quantifiers, these quantifiers may be rearranged in any order and the resulting *wff* will be equivalent to the original *wff.*

This may be called the *rule of quantifier rearrangement.*

Two other important rules concern implications, not equivalences:

(iii) If a *wff* β begins with an unbroken sequence of quantifiers, and β' is obtained from β by replacing $\forall$ by $\exists$ at one or more places in the sequence, then β is *stronger* than β'—in the sense that $(\beta \supset \beta')$ is valid but $(\beta' \supset \beta)$ is in general not valid. (iv) If a *wff* β begins with an unbroken sequence of quantifiers in which some existential quantifier Q_1 precedes some universal quantifier Q_2, and if β' is obtained from β by moving Q_1 to the right of Q_2, then β is stronger than β'.

As illustrations of these rules, the following are valid for any *wff* α:

$(\forall x)(\forall y)\alpha \supset (\exists x)(\forall y)\alpha$ Rule (iii)
$(\exists x)(\forall y)(\forall z)\alpha \supset (\exists x)(\forall y)(\exists z)\alpha$ Rule (iii)
$(\exists x)(\forall y)\alpha \supset (\forall y)(\exists x)\alpha$ Rule (iv)
$(\exists x)(\exists y)(\forall z)\alpha \supset (\exists y)(\forall z)(\exists x)\alpha$ Rule (iv)

In each case the converses are not valid (though they may be valid in particular cases in which α is of some special form).

Application of the rules

Some of the uses of the above rules can be illustrated by considering a *wff* α that contains precisely two free individual variables. By prefixing to α two appropriate quantifiers and possibly one or more negation signs, it is possible to form a closed *wff* (called a *closure* of α) that will express a determinate proposition when a meaning is assigned to the predicate variables. The above rules can be used to list exhaustively the nonequivalent closures of α and the implication relations between them. The simplest example is φxy, which for illustrative purposes can be taken to mean "x loves y." Application of Rules (i) and (ii) will show that every closure of φxy is equivalent to one or another of the following 12 *wff*'s (none of which is in fact equivalent to any of the others):

(a) $(\forall x)(\forall y)\varphi xy$ ("Everybody loves everybody");
(b) $(\exists x)(\forall y)\varphi xy$ ("Somebody loves everybody");
(c) $(\exists y)(\forall x)\varphi xy$ ("There is someone whom everyone loves");
(d) $(\forall y)(\exists x)\varphi xy$ ("Each person is loved by at least one person");
(e) $(\forall x)(\exists y)\varphi xy$ ("Each person loves at least one person");
(f) $(\exists x)(\exists y)\varphi xy$ ("Somebody loves somebody");
(g–l) The respective negations of each of the above.

Rules (iii) and (iv) show that the following implications among formulas (a)–(f) are valid:

(a) $\supset$ (b); (d) $\supset$ (f); (c) $\supset$ (e);
(b) $\supset$ (d); (a) $\supset$ (c); (e) $\supset$ (f).

The implications holding among the negations of (a)–(f) follow from these by the *law of transposition; e.g.,* since (a) $\supset$ (b) is valid, so is $\sim$(b) $\supset \sim$(a). The quantification of *wff*'s containing three, four, etc., variables can be dealt with by the same rules.

Relettering of bound variables

Intuitively, $(\forall x)\varphi x$ and $(\forall y)\varphi y$ both "say the same thing," viz., that everything is φ, and $(\exists x)\varphi x$ and $(\exists y)\varphi y$ both mean simply that something is φ. Clearly, so long as the same variable occurs both in the quantifier and as the argument of φ, it does not matter what letter is chosen for this purpose. The procedure of replacing some variable in a quantifier, together with every occurrence of that variable in its scope, by some other variable that does not occur elsewhere in its scope is known as *relettering a bound variable.* If β is the result of relettering a bound variable in a *wff* α, then α and β are said to be *bound alphabetical variants* of each other; and bound alphabetical variants are always equivalent. The reason for restricting the replacement variable to one not occurring elsewhere in the scope of the quantifier can be seen from an example: If φxy is taken as before to mean "x loves y," the *wff* $(\forall x)\varphi xy$ expresses the proposition form "Everyone loves y," in which the identity of y is left unspecified, and so does its bound alphabetical variant $(\forall z)\varphi zy$; but if x were replaced by y, the closed *wff* $(\forall y)\varphi yy$ would be obtained, which expresses the proposition that everyone loves himself and is clearly not equivalent to the original.

Prenex normal form

A *wff* in which all of the quantifiers occur in an unbroken sequence at the beginning, with the scope of each extending to the end of the *wff,* is said to be in *prenex normal form* (PNF—Latin: *pre,* "in front"; and *nectere,* "to bind"). *Wff*'s that are in PNF are often more convenient to work with than those that are not. For every *wff* of LPC,

however, there is an equivalent *wff* in PNF (often simply called its PNF). One effective method for finding the PNF of any given *wff* is the following:

1. Reletter bound variables as far as is necessary to ensure (a) that each quantifier contains a distinct variable and (b) that no variable in the *wff* occurs both bound and free.
2. Use definitions or PC equivalences to eliminate all operators except $\sim$, $\cdot$, and $\vee$.
3. Use the *De Morgan laws* and the *rule of quantifier transformation* to eliminate all occurrences of $\sim$ immediately before parentheses or quantifiers.
4. Gather all of the quantifiers into a sequence at the beginning in the order in which they appear in the *wff* and take the whole of what remains as their scope.

Example: $(\forall x)\,\{[\varphi x \cdot (\exists y)\,\psi xy] \supset (\exists y)\chi_{xy}\} \supset (\exists z)\,(\varphi z \supset \psi zx)$.

Step 1 can be achieved by relettering the third and fourth occurrences of y and every occurrence of x except the last (which is free); thus:

$$(\forall w)\{[\varphi w \cdot (\exists y)\psi wy] \supset (\exists u)\chi wu\} \supset (\exists z)\,(\varphi z \supset \psi zx)$$

Step 2 now yields:

$$\sim(\forall w)\{\sim[\varphi w \cdot (\exists y)\psi wy] \vee (\exists u)\chi wu\} \vee (\exists z)\,(\sim\varphi z \vee \psi zx)$$

By step 3 this becomes:

$$(\exists w)\{[\varphi w \cdot (\exists y)\psi wy] \cdot (\forall u)\sim\chi wu\} \vee (\exists z)(\sim\varphi z \vee \psi zx)$$

Finally, step 4 yields:

$$(\exists w)(\exists y)(\forall u)(\exists z)\{[(\varphi w \cdot \psi wy) \cdot \sim\chi wu] \vee (\sim\varphi z \vee \psi zx)\}$$

which is in PNF.

Classification of dyadic relations. Consider the closed *wff*,

$$(\forall x)\,(\forall y)\,(\varphi xy \supset \varphi yx)$$

which means that, whenever the relation φ holds between one object and a second, it also holds between that second object and the first. This expression is not valid, since it is true for some relations but false for others. A relation for which it is true is called a *symmetrical* relation. (*Example:* "is parallel to.") If the relation φ is such that whenever it holds between one object and a second it fails to hold between the second and the first; *i.e.,* if φ is such that

Symmetry, transitivity, reflexivity

$$(\forall x)\,(\forall y)\,(\varphi xy \supset \sim\varphi yx)$$

then φ is said to be *asymmetrical.* (*Example:* "is greater than.") A relation that is neither symmetrical nor asymmetrical is said to be *nonsymmetrical.* Thus φ is nonsymmetrical if

$$(\exists x)\,(\exists y)\,(\varphi xy \cdot \varphi yx) \cdot (\exists x)\,(\exists y)\,(\varphi xy \cdot \sim\varphi yx)$$

(*Example:* "loves.")

Dyadic relations can also be characterized in terms of another threefold division: A relation φ is said to be *transitive* if, whenever it holds between one object and a second and also between that second object and a third, it holds between the first and the third; *i.e.,* if

$$(\forall x)\,(\forall y)\,(\forall z)\,[(\varphi xy \cdot \varphi yz) \supset \varphi xz]$$

(*Example:* "is greater than.") An *intransitive* relation is one that, whenever it holds between one object and a second and also between that second and a third, fails to hold between the first and the third; *i.e.,* φ is intransitive if

$$(\forall x)\,(\forall y)\,(\forall z)\,[(\varphi xy \cdot \varphi yz) \supset \sim\varphi xz]$$

(*Example:* "is father of.") A relation that is neither transitive nor intransitive is said to be *nontransitive.* Thus φ is nontransitive if

$$(\exists x)\,(\exists y)\,(\exists z)\,(\varphi xy \cdot \varphi yz \cdot \varphi xz) \cdot (\exists x)\,(\exists y)\,(\exists z)\,(\varphi xy \cdot \varphi yz \cdot \sim\varphi xz)$$

(*Example:* "is a first cousin of.")

A relation φ that always holds between any object and itself is said to be *reflexive; i.e.,* φ is reflexive if

$$(\forall x)\varphi xx$$

(*Example:* "is identical with.") If φ never holds between any object and itself—*i.e.,* if

$$\sim(\exists x)\varphi xx$$

—then φ is said to be irreflexive (*Example:* "is greater than.") If φ is neither reflexive nor irreflexive—*i.e.,* if

$$(\exists x)\varphi xx \cdot (\exists x)\sim\varphi xx$$

—then φ is said to be *nonreflexive* (*Example:* "admires.")

A relation such as "is of the same length as" is not strictly reflexive, since some objects do not have a length at all and therefore are not of the same length as anything, even as themselves. Nevertheless, this relation is reflexive in the weaker sense that whenever an object is of the same length as anything it is of the same length as itself. Such a relation is said to be *quasi-reflexive.* Thus φ is quasi-reflexive if

$$(\forall x)[(\exists y)\varphi xy \supset \varphi xx].$$

A reflexive relation is of course also quasi-reflexive.

For the most part, these three classifications are independent of each other; thus a symmetrical relation may be transitive (like "is equal to") or intransitive (like "is perpendicular to") or nontransitive (like "is one mile distant from"). There are, however, certain limiting principles, of which the most important are:

1. Every relation that is symmetrical and transitive is at least quasi-reflexive;
2. Every asymmetrical relation is irreflexive;
3. Every relation that is transitive and irreflexive is asymmetrical.

A relation that is reflexive, symmetrical, and transitive is called an *equivalence* relation.

Axiomatization of LPC. Rules of uniform substitution for predicate calculi, though formulable, are mostly very complicated; and, to avoid the necessity for these rules, axioms for these systems are therefore usually given by axiom schemata in the sense explained in *Axiomatization of PC.* Given the formation rules and definitions stated in the introductory paragraph of *The lower predicate calculus,* the following is presented as one standard axiomatic basis for LPC:

Axiomatic basis and undecidability of theorems

Axiom schemata:

1. Any LPC substitution-instance of any valid *wff* of PC is an axiom.
2. Any *wff* of the form $(\forall \mathbf{a})\alpha \supset \beta$ is an axiom, if β is either identical with α or differs from it only in that, wherever α has a free occurrence of **a**, β has a free occurrence of some other individual variable **b**.
3. Any *wff* of the form $(\forall \mathbf{a})\,(\alpha \supset \beta) \supset [\alpha \supset (\forall \mathbf{a})\,\beta]$ is an axiom, provided that α contains no free occurrence of **a**.

Transformation rules:

1. *Modus ponens* (as given above *Axiomatization of PC*).
2. If α is a theorem, so is $(\forall \mathbf{a})\alpha$, where **a** is any individual variable (*rule of universal generalization*).

The axiom schemata call for some explanation and comment. By an LPC substitution-instance of a *wff* of PC is meant any result of uniformly replacing every propositional variable in that *wff* by a *wff* of LPC. Thus one LPC substitution-instance of $(p \supset \sim q) \supset (q \supset \sim p)$ is $[\varphi xy \supset \sim(\forall x)\psi x] \supset [(\forall x)\psi x \supset \sim\varphi xy]$. Axiom schema 1 makes available in LPC all manipulations such as commutation, transposition, and distribution, which depend only on PC principles. Examples of *wff*'s that are axioms by axiom schema 2 are: $(\forall x)\varphi x \supset \varphi x$; $(\forall x)\varphi x \supset \varphi y$; $(\forall x)\,(\exists y)\varphi xy \supset (\exists y)\varphi zy$. To see why it is necessary for the variable that replaces **a** to be free in β, consider the last example: Here **a** is x, α is $(\exists y)\varphi xy$, in which x is free, and β is $(\exists y)\varphi zy$, in which z is free and replaces x. But had y, which would be bound by the quantifier $(\exists y)$, been chosen as a replacement, instead of z, the result would have been $(\forall x)\,(\exists y)\varphi xy \supset (\exists y)\varphi yy$, the invalidity of which can be seen intuitively by taking φxy to mean "x is a child of y": for then $(\forall x)\,(\exists y)\varphi xy$ will mean that everyone is a child of someone, which is true; but $(\exists y)\varphi yy$ will mean

that someone is a child of himself, which is false. The need for the proviso in axiom schema 3 can also be seen from an example. Defiance of the proviso would give as an axiom $(\forall x)(\varphi x \supset \psi x) \supset [\varphi x \supset (\forall x)\psi x]$; and if φx were taken to mean "x is a Spaniard," ψx to mean "x is a European," and the free occurrence of x (the first occurrence in the consequent) to stand for General Franco, then the antecedent would mean that every Spaniard is a European, but the consequent would mean that, if General Franco is a Spaniard, then everyone is a European.

It can be proved—though the proof is not an elementary one—that the theorems derivable from the above basis are precisely the *wff*'s of LPC that are valid by the definition of validity given in *Validity in LPC*. Several other bases for LPC are known that also have this property. The axiom schemata and transformation rules here given are such that any purported proof of a theorem can be effectively checked to determine whether it really is a proof or not; nevertheless, theoremhood in LPC, like validity in LPC, is not effectively decidable, in that there is no effective method of telling with regard to any arbitrary *wff* whether it is a theorem or not. In this respect axiomatic bases for LPC contrast with those for PC.

Special systems of LPC. LPC as expounded above may be modified by either restricting or extending the range of *wff*'s in various ways.

a. Partial systems of LPC. Some of the more important systems produced by restriction are here outlined:

Monadic LPC and subsystem

1. It may be required that every predicate variable be monadic, while still allowing an infinite number of individual and predicate variables. The atomic *wff*'s are then simply those consisting of a predicate variable followed by a single individual variable. Otherwise, the formation rules remain as before, and the definition of validity is also as before, though simplified in obvious ways. This system is known as the monadic LPC; it provides a logic of properties but not of relations. One important characteristic of this system is that it is decidable. (The introduction of even a single dyadic predicate variable, however, would make the system undecidable; and in fact, even the system that contains only a single dyadic predicate variable and no other predicate variables at all has been shown to be undecidable.)

2. A still simpler system can be formed by requiring (a) that every predicate variable be monadic; (b) that only a single individual variable (*e.g.*, x) be used; (c) that every occurrence of this variable be bound; (d) that no quantifier occur within the scope of any other. Examples of *wff*'s of this system are: $(\forall x)[\varphi x \supset (\psi x \cdot \chi x)]$ ("Whatever is φ is both ψ and χ"); $(\exists x)(\varphi x \cdot \sim\psi x)$ ("There is something that is φ but not ψ"); $(\forall x)(\varphi x \supset \psi x) \supset (\exists x)(\varphi x \cdot \psi x)$ ("If whatever is φ is ψ, then something is both φ and ψ"). The notation for this system can be simplified by omitting x everywhere and writing $\exists\varphi$ for "Something is φ," $\forall(\varphi \supset \psi)$ for "Whatever is φ is ψ," and so on. Although this system is more rudimentary even than the monadic LPC (of which it is a fragment), the forms of a wide range of inferences can be represented in it. It is also a decidable system, and decision procedures of an elementary kind can be given for it.

b. Extensions of LPC. More elaborate systems, in which a wider range of propositions can be expressed, have been constructed, by adding to LPC new symbols of various types. The most straightforward of such additions are:

(a) One or more *individual constants* (say *a, b, . . .*): these constants are interpreted as names of specific individuals; formally they are distinguished from individual variables by the fact that they cannot occur within quantifiers: *e.g.*, $(\forall x)$ is a quantifier but $(\forall a)$ is not.

(b) One or more *predicate constants* (say $\Phi, \Psi, \ldots$), each of some specified degree, thought of as designating specific properties or relations.

Definitions: "function," "degree," "term"

A further possible addition, which calls for somewhat fuller explanation, consists of symbols designed to stand for *functions*. The notion of a function may be sufficiently explained for present purposes as follows: There is said to be a certain function of n arguments (or, of degree n) when there is a rule that specifies a unique object (called the *value* of the function) whenever all of the arguments are specified. In the domain of human beings, for example, "the mother of—" is a monadic function (a function of one argument) since for every human being there is a unique individual who is his mother; and in the domain of the natural numbers (*i.e.*, 0, 1, 2, . . .), "the sum of—and—" is a function of two arguments since for any pair of natural numbers there is a natural number that is their sum. A function symbol can be thought of as forming a name out of other names (its arguments): thus whenever x and y name numbers, "the sum of x and y" also names a number, and similarly for other kinds of functions and arguments.

To enable functions to be expressed in LPC there may be added:

(c) One or more *function variables* (say *f, g, . . .*) or one or more *function constants* (say, *F, G, . . .*) or both, each of some specified degree. The former are interpreted as ranging over functions of the degrees specified, and the latter as designating specific functions of that degree.

When any or all of (a) – (c) are added to LPC, the formation rules listed in the first paragraph of *The lower predicate calculus* need to be modified to enable the new symbols to be incorporated into *wff*'s. This can be done as follows: A *term* is first defined as either (i) an individual variable or (ii) an individual constant or (iii) any expression formed by prefixing a function variable or function constant of degree n to any n terms (these terms—the arguments of the function symbol—are usually separated by commas and enclosed in parentheses). Formation rule 1 is then replaced by:

1′. An expression consisting of a predicate variable or predicate constant of degree n followed by n terms is a *wff*.

The axiomatic basis given in *Axiomatization of LPC* also requires the following modification: in axiom schema 2 any term is allowed to replace **a** when β is formed, provided that no variable that is free in the term becomes bound in β. The following examples will illustrate the use of the aforementioned additions to LPC: Let the values of the individual variables be the natural numbers; let the individual constants a and b stand for the numbers 2 and 3, respectively; let Φ mean "is prime"; and let F represent the dyadic function "the sum of." Then $\Phi F(a,b)$ expresses the proposition "The sum of 2 and 3 is prime" and $(\exists x)\Phi F(x,a)$ expresses the proposition "There exists a number such that the sum of it and 2 is prime."

Special axioms

The introduction of constants is normally accompanied by the addition, to the axiomatic basis, of special axioms containing those constants, designed to express principles that hold of the objects, properties, relations, or functions represented by them—though they do not hold of objects, properties, relations, or functions in general. It may be decided, for example, to use the constant Φ to represent the dyadic relation "is greater than" (so that Φxy is to mean "x is greater than y," etc.). This relation, unlike many others, is transitive; *i.e.*, if one object is greater than a second and that second is in turn greater than a third, then the first is greater than the third. Hence, the following special axiom schema might be added:

If t_1, t_2, and t_3 are any terms, then

$$(\Phi t_1 t_2 \cdot \Phi t_2 t_3) \supset \Phi t_1 t_3$$

is an axiom. By such means systems can be constructed to express the logical structures of various particular disciplines. The area in which most work of this kind has been done is that of natural-number arithmetic (see *Axiomatization of arithmetic*).

PC and LPC are sometimes combined into a single system. This may be done most simply by adding propositional variables to the list of LPC primitives, adding a formation rule to the effect that a propositional variable standing alone is a *wff*, and deleting "LPC" in axiom schema 1. This yields as *wff*'s such expressions as: $(p \vee q) \supset (\forall x)\varphi x$; $(\exists x)[p \supset (\forall y)\varphi xy]$.

c. LPC-with-identity. The word "is" is not always used in the same way. In a proposition such as (i) "Socrates is snub-nosed," the expression preceding the "is" names an individual and the expression following it stands for a property attributed to that individual. But in a proposition

such as (ii) "Socrates is the Athenian philosopher who drank hemlock," the expressions preceding and following the "is" both name individuals, and the sense of the whole proposition is that the individual named by the first is the same individual as the individual named by the second. Thus in (ii) "is" can be expanded to "is the same individual as," whereas in (i) it cannot. As used in (ii), "is" stands for a dyadic relation, viz., identity, that the proposition asserts to hold between the two individuals. An identity proposition is to be understood in this context as asserting no more than this; in particular it is not to be taken as asserting that the two naming expressions have the same meaning. A much discussed example to illustrate this last point is "The morning star is the evening star." It is false that the expressions "the morning star" and "the evening star" mean the same, but true that the object referred to by the former is the same as that referred to by the latter.

Identity propositions

To enable the forms of identity propositions to be expressed, a dyadic predicate constant is added to LPC, for which the most usual notation is = (written between, rather than before, its arguments). The intended interpretation of $x = y$ is that x is the same individual as y, and the most convenient reading is "x is identical with y." Its negation $\sim(x = y)$ is commonly abbreviated to $x \neq y$. To the definition of an LPC model in *Validity in LPC,* there is now added the rule (which accords in an obvious way with the intended interpretation) that the value of $x = y$ is to be 1 if the same member of D is assigned to both x and y and that otherwise its value is to be 0; validity can then be defined as before. The following additions (or some equivalent ones) are made to the axiomatic basis for LPC: the axiom (I1) $x = x$; and the axiom schema (I2) that where **a** and **b** are any individual variables and α and β are *wff*'s that differ only in that, at one or more places where α has a free occurrence of **a**, β has a free occurrence of **b**, $(\mathbf{a} = \mathbf{b}) \supset (\alpha \supset \beta)$ is an axiom. Such a system is known as a *lower-predicate-calculus-with-identity;* it may of course be further augmented in the other ways referred to in *Extensions of LPC,* in which case any term may be an argument of =.

Identity is an equivalence relation; *i.e.,* it is reflexive, symmetrical, and transitive. Its reflexivity is directly expressed in Axiom I1, and theorems expressing its symmetry and transitivity can easily be derived from the basis given.

Numerical quantification

Certain *wff*'s of LPC-with-identity express propositions about the number of things that possess a given property. "At least one thing is φ" could, of course, already be expressed by $(\exists x)\varphi x$; "At least two distinct (non-identical) things are φ" can now be expressed by $(\exists x)(\exists y)(\varphi x \cdot \varphi y \cdot x \neq y)$; and the sequence can be continued in an obvious way. "At most one thing is φ" (*i.e.,* "No two distinct things are both φ") can be expressed by the negation of the last mentioned *wff* or by its equivalent, $(\forall x)(\forall y)[(\varphi x \cdot \varphi y) \supset x = y]$; and the sequence can again be easily continued. A formula for "Exactly one thing is φ" may be obtained by conjoining the formulas for "At least one thing is φ" and "At most one thing is φ"; but a simpler *wff* equivalent to this conjunction is $(\exists x)[\varphi x \cdot (\forall y)(\varphi y \supset x = y)]$, which means "There is something that is φ, and anything that is φ *is* that thing." "Exactly two things are φ" can be represented by

$$(\exists x)(\exists y)\{\varphi x \cdot \varphi y \cdot x \neq y \cdot (\forall z)[\varphi z \supset (z = x \vee z = y)]\};$$

i.e., "There are two non-identical things each of which is φ, and anything that is φ is one or the other of these." Clearly, this sequence can also be extended to give a formula for "Exactly n things are φ" for every natural number n. It is convenient to abbreviate the *wff* for "Exactly one thing is φ," however, to $(\exists^1 x)\varphi x$, and in general to abbreviate the *wff* for "Exactly n things are φ" to $(\exists^n x)\varphi x$, and in this way to obtain a sequence of *numerical* quantifiers each, however, definable in terms of the primitives of LPC-with-identity.

When a certain property φ belongs to one and only one object, it is convenient to have an expression that names that object. A common notation for this purpose is $(\iota x)\varphi x$, which may be read as "the thing that is φ" or more briefly as "the φ." In general, where **a** is any individual variable and α is any *wff,* $(\iota \mathbf{a})\alpha$ then stands for the single value of **a** that makes α true. An expression of the form "the so-and-so" is called a *definite description;* and (ιx), known as a *description operator,* can be thought of as forming a name of an individual out of a proposition form. (ιx) is analogous to a quantifier in that, when prefixed to a *wff* α, it binds every free occurrence of x in α. Relettering of bound variables is also permissible: in the simplest case, $(\iota x)\varphi x$ and $(\iota y)\varphi y$ can each be read simply as "the φ."

Definite descriptions

As far as formation rules are concerned, definite descriptions can be incorporated into LPC by letting expressions of the form $(\iota \mathbf{a})\alpha$ count as terms; rule 1′ of *Extensions of LPC* will then allow them to occur in atomic formulas (including identity formulas). "The φ is (*i.e.,* has the property) ψ" can then be expressed as $\psi(\iota x)\varphi x$; "y is (the same individual as) the φ" as $y = (\iota x)\varphi x$; "The φ is (the same individual as) the ψ" as $(\iota x)\varphi x = (\iota y)\psi y$; and so forth. The correct analysis of propositions containing definite descriptions has been the subject of considerable philosophical controversy. One widely accepted account, however—substantially that presented in *Principia Mathematica* and known as *Russell's theory of descriptions,* after Bertrand Russell—holds that "The φ is ψ" is to be understood as meaning that exactly one thing is φ and that thing is also ψ. In that case it can be expressed by a *wff* of LPC-with-identity that contains no description operators, viz.,

$$(1)\ (\exists x)[\varphi x \cdot (\forall y)(\varphi y \supset x = y) \cdot \psi x].$$

Analogously "y is the φ" is analyzed as "y is φ and nothing else is φ," and hence as expressible by

$$(2)\ \varphi y \cdot (\forall x)(\varphi x \supset x = y);$$

and "The φ is the ψ" is analyzed as "Exactly one thing is φ, exactly one thing is ψ, and whatever is φ is ψ," and hence as expressible by

$$(3)\ (\exists x)[\varphi x \cdot (\forall y)(\varphi y \supset x = y)] \cdot (\exists x)[\psi x \cdot (\forall y)(\psi y \supset x = y)] \cdot (\forall x)(\varphi x \supset \psi x)$$

or more briefly, using the abbreviations introduced earlier, by

$$(3)\ (\exists^1 x)\varphi x \cdot (\exists^1 x)\psi x \cdot (\forall x)(\varphi x \supset \psi x).$$

$\psi(\iota x)\varphi x$, $y = (\iota x)\varphi x$, and $(\iota x)\varphi x = (\iota y)\psi y$ can then be regarded as abbreviations for (1), (2), and (3), respectively; and by generalizing to more complex cases, all *wff*'s that contain description operators can be regarded as abbreviations for longer *wff*'s that do not.

Negations with differing scopes

The analysis that leads to (1) as a formula for "The φ is ψ" leads to the following for "The φ is not ψ":

$$(4)\ (\exists x)[\varphi x \cdot (\forall y)(\varphi y \supset x = y) \cdot \sim\psi x].$$

It is important to note that (4) is not the negation of (1); this negation is, instead,

$$(5)\ \sim(\exists x)[\varphi x \cdot (\forall y)(\varphi y \supset x = y) \cdot \psi x].$$

The difference in meaning between (4) and (5) lies in the fact that (4) is true only when there is exactly one thing that is φ and that thing is not ψ, but (5) is true both in this case and also when nothing is φ at all and when more than one thing is φ. Neglect of the distinction between (4) and (5) can cause serious confusion of thought: in ordinary speech it is often unclear whether someone who denies that the φ is ψ is conceding that exactly one thing is φ but denying that it is ψ, or denying that exactly one thing is φ. Moreover, if (4) were mistakenly thought of as the negation of (1), the *law of excluded middle* would lead to the conclusion that at least one of them is true; and then, since each asserts *inter alia* that exactly one thing is φ and since, in addition, the argument would hold whatever property φ is taken to represent, the absurd result would be obtained that every property has exactly one instance.

The basic contention of Russell's theory of descriptions is that a proposition containing a definite description is not to be regarded as an assertion about an object of which that description is a name but rather as an existentially quantified assertion that a certain (rather complex) property has an instance. Formally, this is reflected in the rules for eliminating description operators that were outlined above. Some logicians, however, prefer instead to

introduce ι' as a new primitive symbol, accompanied by special axioms, and to treat expressions of the form $(\iota\mathbf{a})\alpha$ always as naming expressions, allowing them to replace individual variables in theorems (subject only to the restrictions about quantification mentioned in *Axiomatization of LPC*). The formula $(\forall x)\varphi x \supset \varphi(\iota y)\psi y$—"If everything is φ, then the ψ is φ"—will then, for example, count as an axiom by axiom schema 2 and as a valid *wff*; whereas by Russell's theory of descriptions it will be invalid, since it is possible for everything to be φ without there being exactly one thing that is ψ. This approach gives rise to the problem of deciding, in constructing an *LPC* model, which member of D to assign to $(\iota x)\varphi x$ in cases in which either nothing or more than one thing is φ. The solution usually adopted is to select some arbitrary member of D as the assignment in all such cases throughout the model. While this device may seem intuitively "unnatural," it proves to be technically satisfactory and convenient for many purposes.

Higher-order predicate calculi. A feature shared by LPC and all of its extensions so far mentioned is that the only variables that occur in quantifiers are individual variables. It is in virtue of this feature that they are called *lower* (or *first-order*) calculi. Various predicate calculi of higher order can be formed, however, in which quantifiers may contain other variables as well, hence binding all free occurrences of these that lie within their scope. In particular, in the *second-order* predicate calculus quantification is permitted over both individual and predicate variables; hence *wff*'s such as $(\forall\varphi)(\exists x)\varphi x$ can be formed. This last formula, since it contains no free variables of any kind, expresses a determinate proposition, viz., the proposition that every property has at least one instance. One important feature of this system is that in it identity need not be taken as primitive but can be introduced by defining $x = y$ as $(\forall\varphi)(\varphi x \equiv \varphi y)$—*i.e.*, "Every property possessed by x is also possessed by y and vice versa." Whether such a definition is acceptable as a general account of identity is a question that raises philosophical issues too complex to be discussed here; they are substantially those raised by the principle of the identity of indiscernibles, best known for its exposition in the 17th century by Leibniz.

Binding of predicate variables

(G.E.H./Ed.)

SYLLOGISTIC

Syllogistic was the earliest branch of formal logic to be developed. As explained earlier, this study of logical structures makes it possible to infer a conclusion from other given propositions (the premises).

Nature and elements of syllogistic. Since the propositions that compose all syllogisms are built up from terms and operators (signs expressing logical connections), traditional logic begins with an analysis of terms. Terms may be singular (such as "Plato" or "the author of *Hamlet*") or general (such as "man," "triangle," or "healthy"). General terms may be taken either in extension (also called denotation), which indicates the set of individuals to which a term applies, or in intension (also called connotation), which describes the set of attributes that something must have to qualify for the predication of the term. The extension, for example, of "bachelor" is the set including all men that are bachelors, and its connotation is the set of such attributes as being human, being male, being above the legal age of marriage, and being single. If the connotation of a term a includes that of b, the extension of b will include that of a. Whereas the class of bachelors is included in the class of men, the attributes of "man" are included in the class of attributes of "bachelor." The definition of a term is an analysis of its connotation.

To form a proposition of the most primitive sort, a singular term (the subject) may be joined to a general term (the predicate) by "___ is (a) ___ " or "___ is not (a) ___ ," signs called copulae. "Plato is a philosopher" is called an affirmative proposition, and "Plato is not stupid" is called a negative proposition. General terms cannot form propositions merely by being joined by copulae; in order to construct a complete proposition, a sign of quantification must be added—traditionally "every" or "some," as in "Every philosopher is a man." Instead of writing "Every man is not perfect," it is customary to write "No man is perfect," the "no" doing double duty as a sign of negation *and* quantification. A proposition of the type "Every ___ is (a) ___ " is called universal affirmative; "No ___ is (a) ___ " is called universal negative; "Some ___ is (a) ___ " is called particular affirmative; Some ___ is not (a) ___ " is called particular negative. These forms of propositions are also referred to, in that order, as *A-*, *E-*, *I-*, and *O*-propositions, the vowels being taken from the Latin words *affirmo* ("I affirm") and *nego* ("I deny"). Negative and affirmative propositions are said to differ in *quality;* universal, particular (and singular) propositions are said to differ in *quantity.*

A proposition is said to be *inferred* from other propositions if it is validly drawn as a conclusion from them. A conclusion is validly drawn if it is a logical consequence of the premises, and this is the case whenever it is imposible that the premises be true and the conclusion false. The following cases are therefore possible:

Truth and validity of inference

1. A true conclusion invalidly drawn from true (or false) premises,
2. A false conclusion invalidly drawn from true (or false) premises,
3. A true conclusion validly inferred from true (or false) premises, and
4. A false conclusion validly inferred from false premises.

The only case that cannot arise is:

5. A false conclusion validly inferred from true premises.

Therefore, to know the truth or falsity of the premises and the conclusion does not enable one to determine the validity of an inference. To judge the validity of an argument, one must examine its logical form.

Procedures of immediate inference. In expressing syllogistic propositions, logicians employ a notation introduced by Jan Łukasiewicz: lower-case Latin letters *a, b, c, ...* are employed to stand for general terms (*i.e.*, for term variables); capital letters *A, E, I,* and *O,* are then used as symbols for the syllogistic operators, which, together with the terms, make up the four kinds of propositions. Accordingly, "Every b is an a" is written "*Aba*"; "Some b is an a" is written "*Iba*"; "No b is an a" is written "*Eba*"; and "Some b is not an a" is written "*Oba.*"

Relations among propositions. In order to display the logical relations obtaining between different pairs of propositions of these four forms, they are customarily arranged in a so-called square of opposition (Figure 2). It is important to note that, in considering these different forms of opposition between propositions, the logical tradition conventionally distinguishes empty terms (such as "round square," "centaur," or "*perpetuum mobile*"), which cannot be truly applied to any individual, and excludes them from being introduced in the place of term-variables in syllogistic propositions. It is also necessary to be aware

Square of opposition

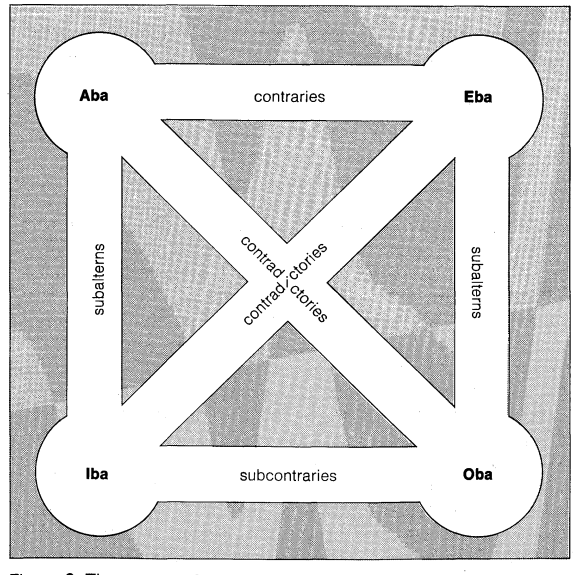


Figure 2: The square of opposition (see text).

that the operators *I* and *O* ("Some ___ is (a) ___ " and "Some ___ is not (a) ___ ") are taken in the sense of "At least one and possibly all ___ is (or is not) (a) ___ ." With these specifications made, it may be seen that "*Aba*" and "*Eba*" are contraries—*i.e.,* they cannot be jointly true; that "*Iba*" and "*Oba*" are subcontraries—*i.e.,* they cannot be jointly false; that "*Aba*" and "*Oba*" and also "*Eba*" and "*Iba,*" taken pairwise, are contradictories—*i.e.,* if "*Aba*" is true, "*Oba*" must be false and vice versa; and, finally, that if "*Aba*" is true, "*Iba*" must also be true, and if "*Eba*" is true, so is "*Oba.*" In the last cases, however, since the implication does not hold analogously in the opposite direction, it is said that "*Aba*" "subalternates" "*Iba,*" and "*Eba*" subalternates "*Oba,*" "*Aba*" being the subalternant and "*Iba*" the subalternate (and, similarly, for "*Eba*" and "*Oba*"). For any given pair of terms *a* and *b,* exactly two of these four forms of proposition are true: for every pair of terms "*a*" and "*b,*" either "*Aba*" and "*Iba*" will be true or "*Eba*" and "*Oba*" or, finally, "*Iba*" and "*Oba.*"

If the extensions of general terms are symbolized by circles, it is possible to develop a series of so-called Euler diagrams (Figure 3)—named after the Swiss mathematician Leonhard Euler, who promulgated the method (though

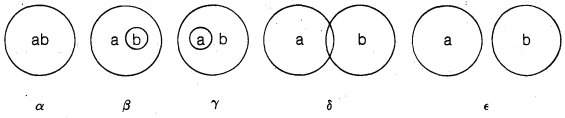


Figure 3: The Euler diagrams (see text).

he had not invented it). The diagrams show how the five possible relations of terms may be visualized: α: exactly matching; β: *b* included in *a*; γ: *a* included in *b*; δ: partially overlapping; ε: entirely distinct. As long as negated term variables—*e.g.,* "not-man" or "nonsmoker"—are not considered, the five diagrams exhaust the possibilities. These figures will make clear the logical relations between pairs of propositions that are alike in quality but have the order of the terms *a* and *b* inverted: ("*Aba,*" "*Aab*"), ("*Aba,*" "*Iab*"), ("*Iba,*" "*Aab*"), ("*Iba,*" "*Iab*"), ("*Eba,*" "*Eab*"), ("*Eba,*" "*Oab*"), ("*Oba,*" "*Eab*"), ("*Oba,*" "*Oab*"). As the former logical relations were traditionally called those of opposition, the ensuing relations are those of conversion.

Conversion

The identification of logically valid conversions is facilitated by the use of so-called truth tables (Table 6). A "t"

Table 6: Truth Table Correlating Propositional Forms with Euler Diagrams

Euler diagram	propositional form				converse propositional form			
	Aba	Eba	Iba	Oba	Aab	Eab	Iab	Oab
(α)	t	f	t	f	t	f	t	f
(β)	t	f	t	f	f	f	t	t
(γ)	f	f	t	t	t	f	t	f
(δ)	f	f	t	t	f	f	t	t
(ε)	f	t	f	t	f	t	f	t

in a given line and column signifies that the propositional form given at the top of the column will be true for the Euler diagram indicated at the left by Greek letters (referring to the different diagrams of term relations in Figure 3); an "f," on the other hand, signifies that the propositional form will be false. The columns for "*Eba*" and "*Eab*" are identical, as are those for "*Iba*" and "*Iab.*" "*Eab,*" therefore, may safely be inferred from "*Eba,*" and "*Eba*" from "*Eab*"; and "*Iab*" may also be inferred from "*Iba,*" and "*Iba,*" again, from "*Iab.*" Traditionally, this logical fact is expressed by saying that *E*- and *I*-propositions are capable of simple conversion. Generally, a syllogistic proposition *Y* may be inferred from another proposition *X* built up from the same terms whenever the column of *X* has no "t" in a line in which the column of *Y* has an "f." Following this criterion it is shown that "*Aba*" implies "*Iab*" (but not vice versa) and that "*Eba*" implies "*Oab*" (but not vice versa). In this kind of inference (traditionally called *conversio per accidens*) not only the order of the terms but also the quantity of the proposition is changed. From the columns of Table 6, it can be seen that "*Oba*" cannot imply "*Oab,*" nor any other propositional form with inverted terms, because these all have one or more "f's" in the last three lines, where "*Oba*" has only "t's." It is possible that, for particular terms *a* and *b,* both "*Oba*" and "*Oab*" may be jointly true (*e.g.,* if *a* = "German" and *b* = "blond"); but this joint truth depends on the accidental content of the particular terms, not on the logical structure of the propositions. Thus, "*Oba*" is not eligible for any conversion.

Equipollence, obversion, and contraposition

Since, in all of its lines, column "*Aba*" exhibits truth values exactly the opposite of those in column "*Oba,*" another logical connection—traditionally called equipollence (near in sense to the modern term "equivalence")—can be identified, asserting that "*Aba*" is logically equipollent to "*Not-Oba*" (or "*NOba*") and vice versa. The same relation holds between "*Eba*" and "*Iba*" (or "*Iab*"). Using "≡" as a sign for equivalence, the logician can write "*Eba* ≡ *NIba*" and "*Eba* ≡ *NIab.*" Negated term variables can be introduced for terms such as not-man or nonsmoker by employing a superposed bar, $\bar{a}$, $\bar{b}$, ..., provided only that, just as terms defining the null (empty) class were earlier excluded, all terms defining the universal class shall be disqualified—in short, that, for every *a*, $\bar{a}$ is not empty. Thus every syllogistic proposition is equivalent to the proposition that results by changing its quality (affirmative to negative or vice versa) and at the same time negating the second (predicate) term: thus, "*Aba*" is equivalent to "*Eb*$\bar{a}$," and "*Iba*" to "*Ob*$\bar{a}$," a change traditionally called obversion. Another type of immediate inference, called contraposition, rests on the fact that some syllogistic propositions imply the proposition that results from the original proposition when both of its term variables are negated and their order is reversed. Only in the case of "*Aba*" and "*Oba*" is the resulting proposition logically equivalent to the original one. It is evident, for example, that "If every member is a smoker, then every nonsmoker is a nonmember"; and that "If some member is not a smoker, then some nonsmoker is not a nonmember"; *i.e.,* "Some nonsmoker is a member," or "There are members besides smokers."

After the foregoing definitions of equipollence and obversion, it is possible to define each of the operators *A, E, I,* and *O* by negation plus any one of the other three operators. "*Aba*" may be taken as basic, for example, and "*Eba*" defined as "*Ab*$\bar{a}$," "*Iba*" as "*NAb*$\bar{a}$," and "*Oba*" as "*NAba.*" Corresponding definitions can be framed with other operators as basic. Instead of employing *A, E, I,* and *O,* for instance, modern logic uses for the same task only two operators, (∀*x*) and (∃*x*), the universal and existential quantifiers, which express only quantity ("all" and "some," respectively), leaving the quality, in the case of negative syllogistic propositions, to be expressed by obversion.

Summary of inference relations. Summing up (and using "⊃" for "implies" and "≡" for "is equivalent to"), the logical laws of Table 7 can be stated, each of which provides the basis for an immediate inference (from one syllogistic proposition to another).

Table 7: Summary of the Laws of Immediate Inference

law no.	law	law no.	law
From the square of opposition		**By conversion**	
2.1	Aba ⊃ NEba	2.11	Eba ≡ Eab
2.2	Eba ⊃ NAba	2.12	Iba ≡ Iab
2.3	NIba ⊃ Oba	2.13	Aba ⊃ Iab
2.4	NOba ⊃ Iba	2.14	Eba ⊃ Oab
2.5	Aba ⊃ Iba	**By obversion**	
2.6	Eba ⊃ Oba	2.15	Aba ≡ Eb$\bar{a}$
2.7	NIba ⊃ NAba	2.16	Iba ≡ Ob$\bar{a}$
2.8	NOba ⊃ NEba	**By contraposition**	
2.9	Aba ≡ NOba	2.17	Aba ≡ A$\bar{a}\bar{b}$
2.10	Eba ≡ NIba	2.18	Oba ≡ O$\bar{a}\bar{b}$

Rules and laws of the syllogism. *Categorical syllogistic inferences.* A categorical proposition, as previously explained, contains just one subject term, one predicate term, and one syllogistic operator to join them. A categorical syllogism is an inference of one such proposition from two others. Each premise must have one term, called the middle term, in common with the other premise and one term

Terms, moods, and figures

in common with the conclusion. The term functioning as predicate of the conclusion is called the major term, and the one functioning as subject of the conclusion is called the minor term of the syllogism. The premise containing the major term, conventionally stated first, is called the major premise, and the other is called the minor premise. In the syllogistic example:

syllogism	proposition type
No man is perfect	*E*
Every logician is a man	*A*
Therefore, no logician is perfect	*E*

"perfect" is the major, "man" the middle, and "logician" the minor term.

The quantity and quality of the three propositions—expressed, as before, in terms of the proposition types *A, E, I,* and *O*—determine what are called the moods of the syllogism. The example just given thus belongs, as shown in the right-hand column, to the mood *EAE* (reading down). Since the three propositions may take four different forms each, there are $4^3 = 64$ different moods. Another way in which the forms of syllogisms differ is in their figure, which is the pattern in which its terms *S, M,* and *P* (subject, middle, and predicate) are arranged in the premises (see Table 8). This pattern is effectively determined by the function of the middle term, which may be subject in the major and predicate in the minor premise, as in the foregoing example (first figure); predicate in both premises (second figure); subject in both premises (third figure); and, finally, predicate in the major and subject in the minor premise (fourth figure).

Table 8: The Four Figures of the Syllogism

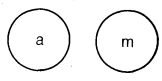
(1)	(2)	(3)	(4)
M-P	*P-M*	*M-P*	*P-M*
S-M	*S-M*	*M-S*	*M-S*
S-P	*S-P*	*S-P*	*S-P*

Analysis of the valid moods

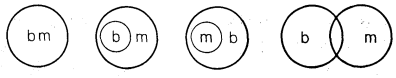
Since there are 64 moods in four figures, there can be 256 forms of the categorical syllogism. It is one of the tasks of syllogistic to discover which of these forms are valid and to show that all others are invalid. Aristotle accepted 14 valid moods officially and five unoffically (those of the fourth figure). In the Middle Ages, these 19 moods were given Latin names, such as *Barbara, Camestres,* or *Datisi,* which indicated the moods through their vowels, names that were listed in mnemonic verses. Since five of these 19 syllogisms have universal conclusions, from which corresponding particular propositions can be inferred (by passing from "all" to "some" by laws 2.5 and 2.6), there will be five additional valid syllogisms—traditionally called *Barbari, Celaront* (first figure); *Cesaro, Camestrop* (second figure); and *Camenop* (fourth figure). Upon examination, it would seem that a syllogism in *Barbara* is self-evident; and Aristotle, in fact, accepted all first-figure syllogisms as self-evident and deduced all other syllogisms from them. But Euler diagrams (see Figure 3) can also be employed. Taking as an example *EIO,* in the second figure (*Festino*), written

$$Eam \cdot Ibm \supset Oba,$$

(using · for "and"), the relation of *a* and *m* can be shown by the Euler diagram (ε):



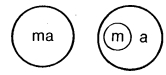
The second premise, "Some *b* is *m,*" is satisfied by each of the diagrams (α) through (δ):



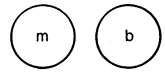
since, in each case, at least some of the "*b*'s" must be inside the extension of "*m.*" Because, by the first premise, *a* is wholly excluded from *m,* however, at least some members of *b* will be outside of *a*; and this is just what "*Oba,*" the conclusion, asserts. Similarly, it can be shown that *AEE,* in the third figure, is not valid; *i.e.,*

$$Ama \cdot Emb \supset Eba \text{ (invalid).}$$

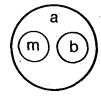
"*Ama*" is satisfied by the Euler diagrams (α) and (β),



and "*Emb*" by diagram (ε),



Clearly, the arrangement



is possible, in which case "*Aba*" would be true. Because *b* might equally well fall extensionally wholly outside of *a,* however, or *a* and *b* might merely overlap, all four propositional forms concerning *a* and *b* are compatible, and thus no conclusion follows.

Three sets of general rules for the validity of categorical syllogisms have been formulated.

Rules for testing validity

There are, first, the rules of distribution. A term is said to be distributed in a given proposition if that proposition implies all other propositions that differ from it only in having, in place of the original term, any other term whose extension is a part of that of the original term. Thus the term "student" is distributed in the propositions

Every student is a member;
No general is a student;
No student is a general;
Some boy is not a student;

inasmuch as each proposition remains true if one inserts, say, "law student" or "girl student" instead of "student." "Student" is not distributed, however, in the propositions

Every member is a student;
Some student is a boy;
Some boy is a student;
Some student is not a girl;

which may become false under a corresponding insertion for "student." In general, negative propositions distribute their predicate term, and universal propositions distribute their subject term. Therefore, in "*Eba,*" both *b* and *a* are distributed, in "*Aba*" only *b,* and in "*Oba*" only *a*; "*Iba,*" however, distributes neither of its terms.

There are two rules of distribution:

1. The middle term must be distributed at least once in the premises, and
2. No term may occur distributed in a valid conclusion unless distributed in the premises.

Thus "*Aam · Ibm*" has an undistributed middle term. In the syllogism

$$Ama \cdot Obm \supset Oba \text{ (invalid),}$$

however, the middle term is distributed; but *a,* distributed in the conclusion, is distributed in neither premise, a fault called illicit process.

The rules of quality are the following:

3. One premise, at least, must be affirmative, and
4. If one premise is negative, the conclusion must be negative; and a negative conclusion requires one negative premise.

Though rules 1–4 are sufficient to test any syllogism, three more rules facilitate the testing: one more rule of quality, viz.,

5. If both premises are affirmative, the conclusion must be affirmative,

and two rules of quantity:

6. At least one premise must be universal, and
7. If either premise is particular, the conclusion must be particular.

All syllogisms satisfying all of these rules are valid; and a syllogism violating at least one of them is invalid.

Though inferences framed in the first figure have usually

Underlying logical structures

seemed particularly natural or evident and have thus been given prominence, the logical structures lying at the roots of all syllogistic figures are the same; the different figures only bring into focus their different aspects. The logical forms can be approached from various angles:

1. Given a general law that some class has (or lacks) a certain property, one may then establish that something has (or lacks) this property by showing that it is a member of that class (first figure);
2. Given the same general law, if one wishes to establish that something is not a member of that class, he may do so by showing that it lacks (or has) that property (second figure);
3. If it has been shown that something has (or lacks) a certain property and that it is a member of a certain class, one can then infer that no general law asserting that all members of this class lack (or have) that property can be valid (third figure).

If the different functions of the syllogistic propositions are described as "the statement of a law," "the establishment of a case," and "the application of the law to the case," the special nature of the figures may then be schematically summed up as in Table 9. In all three figures, the proposition of the application step is the only one capable of assuming all four propositional forms, *A, E, I,* and *O*.

Table 9: Functional Summary of the First Three Figures of the Syllogism

figure	law	case	application	case	law
first	affirmed	affirmed	affirmed		
second	affirmed		denied	denied	
third			denied	affirmed	denied

Reduction to the first figure

As for the fourth figure, although, like the first, it conforms to the law-case-application model, it does so with special emphasis not on the class to which the property is applied but on the property itself, of which the argument shows that it has (or lacks) application to something, viz., to all or some of the members of a certain class.

Aristotle, in *Analytica priora A* 5–7, reduced all other syllogisms (called by him "imperfect") to those of the first figure, the "perfect" (*i.e.,* self-evident) syllogism. He proved their validity by deriving them as theorems from the axioms of his system—thus qualifying as the first scholar known to have developed an axiomatic system.

Aristotle used two main methods of proof: direct and indirect reduction. The first has its logical basis in an *a fortiori* argument: if a syllogism *X* is valid, he argued, the logician may also infer the validity of every syllogism *Y* such that the premises of *Y* imply those of *X,* and the conclusion of *X* implies that of *Y*. The reduction, for example, of *Cesare* (second figure),

$$Eam \cdot Abm \supset Eba$$

to *Celarent* (first figure), *Ema* · *Abm* ⊃ *Eba,* proceeds thus: "*Eam*" implies "*Ema*" (by 2.11); the second premises of *Cesare* and of *Celarent* are the same; therefore, the premises of *Cesare* imply those of *Celarent.* But *Celarent*'s conclusion, being identical, implies that of *Cesare*; therefore, *Celarent*'s validity implies that of *Cesare.* More complicated is the reduction of *Camestres* (second figure),

$$Aam \cdot Ebm \supset Eba$$

to *Celarent* (first figure). "*Aam* · *Ebm*" implies (by 2.11) "*Aam* · *Emb.*" After changing the order of premises, one obtains "*Emb* · *Aam,*" from which, by *Celarent,* "*Eab*" can be inferred. But this implies (again by 2.11) the conclusion of the original *Camestres.* Thus, *Camestres* has been reduced to *Celarent,* the one having been logically derived from the other.

Indirect reduction

Baroco (second figure) and *Bocardo* (third figure) cannot be reduced by conversion because *AO* premises (or *OA* premises) imply only *IO* (or *OI*) premises by conversion, which do not even form a valid syllogism, still less a perfect one. Therefore, Aristotle fell back on another proof, indirect reduction (or *reductio ad impossibile*), based on another logical law: that if a syllogism is valid, the negation of its conclusion (posited for the sake of argument) together with either of its premises will imply, as a conclusion, the negation of the remaining premise; *i.e.,*

1. $(p \cdot q \supset r) \supset (p \cdot \bar{r} \supset \bar{q})$

and

2. $(p \cdot q \supset r) \supset (\bar{r} \cdot q \supset \bar{p})$.

To show that *Baroco* (second figure),

$$Aam \cdot Obm \supset Oba$$

is valid, its conclusion "*Oba*" is hypothetically negated (by 2.9) to "*Aba.*" But "*Aam* · *Aba*" implies by *Barbara* (first figure) "*Abm,*" which is the contradictory of the original second premise. The hypothesis that *Baroco*'s premises may be true and its conclusion false for the same terms leads to a contradiction ("*Abm* · *Obm*"); therefore, this hypothesis must be false; *i.e., Baroco* is valid. The reduction of *Bocardo* (third figure) is strictly analogous.

Aristotle had already stated (*Analytica priora B* 11–13) that indirect reduction can be used not only for *Baroco* and *Bocardo* but for all moods of the second and third figures and that, even within the first figure, *Darii* and *Ferio* may be reduced via *Camestres* and *Festino* (second figure) to the first-figure mood *Celarent,* using one step of direct and one of indirect reduction in each case—which reduces the number of axioms, or forms commended only by their self-evidence, to just two: *Barbara* and *Celarent.* In fact, all of the 24 valid syllogisms may be arranged into groups of three, with each member derivable from the other two by indirect reduction.

Development from antilogisms

Inasmuch as the validity of a syllogism means that the truth of the premises is incompatible with the falsity of the conclusion, then the fact that, for example, *Ama* · *Abm* ⊃ *Aba* is a valid syllogism may be expressed in another way by saying that "*Ama,*" "*Abm,*" and "*Oba*" (the negation of the conclusion) cannot be jointly true, or, equivalently (using "~" to symbolize "not"), that ~(*Ama* · *Abm* · *Oba*). Thus, *AAO* is defined as a (basic) antilogism. Other antilogisms are *AAE, EAI, AIE,* and *EIA* and, for the fourth figure, *AAE* and *AEI.* Since in any conjunction the order of terms is logically indifferent, the number of antilogisms from which all 24 categorical syllogisms may be developed can be reduced to just three: *AAO, AAE,* and *AIE.*

Several other types of categorical syllogism have been described; a sorites, for example, is a polysyllogism in the form of a chain (with intermediate conclusions suppressed), with the conclusion uniting the subject of the last link with the predicate of the first, thus: *Aba* · *Acb* · *Adc* . . . *Eyx* ⊃ *Eya.* An enthymeme, as used in post-Aristotelian times, is a syllogism in which one of the premises (or even the conclusion) is suppressed, being thought too obvious to need expression. And, finally, in *Analytica priora B* 23, Aristotle introduced the inductive syllogism, which establishes a general rule by making an assuredly complete examination of all cases (one by one) and showing it to be true for every one.

Noncategorical syllogistic inferences. Besides categorical syllogisms, there are also noncategorical syllogisms of several types—hypothetical, disjunctive, and (in some treatments) copulative syllogisms. Since the fundamental difference between propositional logic, which treats unanalyzed entire propositions as its units, and predicate logic, which deals with terms in combination, was not appreciated by logicians before the rise of modern mathematical logic, there has been widespread confusion on this topic. The famous and influential *La Logique; ou l'art de penser* (1662; Eng. trans., *The Port-Royal Logic,* 1851), for example, failed to recognize the difference between implication between terms and implication between propositions and threw together arguments containing propositions such as "If the Chinese are Muslims, then they are infidels" and "If one of the elect perishes, then God is mistaken," the first being of the type expressed in modern logic by $\forall x[Fx \supset (Gx \supset Hx)]$, in which such expressions as Fx signify that "x has the property F," and the second being of the form "$p \supset q$" (in which p is called the antecedent and q the consequent). In 1887 the logician John Neville Keynes, insisting on these important formal differences, suggested

Hypothetical and disjunctive syllogisms

the name conditional for the first kind of proposition and hypothetical for the second. Noncategorical syllogistic has mainly and wisely been restricted to the important argument forms of propositional logic.

Hypothetical syllogisms are called pure if they consist of propositions all of which are of the form "$p \supset q$"; *e.g.,* (1) "If *p* then *q,* and if *q* then *r*; therefore if *p* then *r.*" Mixed hypothetical syllogisms, however, which have one hypothetical and one categorical premise and a categorical conclusion, have two valid moods: (2) "If *p* then *q,* and *p*; therefore *q*" (called *modus ponens* from the Latin *ponere,* "to affirm"), and (3) "If *p* then *q,* but not-*q*; therefore not-*p*" (called *modus tollens,* from *tollere,* "to deny"). (For relevant fallacies, see below *The critique of forms of reasoning.*)

Disjunctive syllogisms, which contain propositions built up with the "either . . . or" operator (symbolized ⊥), have two important moods: (4) "Either *p* or *q,* but not-*p*; therefore *q*"—or: " . . . but not-*q*; therefore *p*" (called *modus tollendo ponens*); (5) "Either *p* or *q,* and *p*; therefore not-*q*"—or: " . . . and *q*; therefore not-*p* (called *modus ponendo tollens*). Mood (5) is valid, however, only if "either *p* or *q*" is taken in the sense of "at least one of *p* or *q* is true, but not both"; it is invalid, on the other hand, if "*p* or *q*" is taken in the nonexclusive sense, which does not rule out the joint truth of *p* and *q*.

A more complex form of noncategorical syllogism is the dilemma. It has two moods, one constructive, the other destructive: (6) "If *q* then *p,* and if *r* then *p,* and either *q* or *r*; therefore *p*"; (7) "If *p* then *q,* and if *p* then *r,* and either not-*q* or not-*r*; therefore, not-*p.*" By an operation known as conditionalization, it is also possible—in what is essentially a reversal of the case in the dilemma—to draw hypothetical conclusions from categorical premises.

Complex and compound syllogisms

Syllogisms can also have as their terms not simple concepts like *a, b,* or *c* but terms built up from such simples by conjunction ("$a \cdot b$") or disjunction ("$a \perp b$"). Such terms, which are called complex, combine into complex propositions (example: "Every student or professor is eligible"). Compound propositions, on the other hand, are those constructed by combinations of other propositions by operators such as "If *p* then *q,*" "Either *p* or *q,*" or "Both *p* and *q.*" Compound propositions may contain simple or complex propositions or both. "Either every student is efficient or popular, or all professors and schoolteachers are frustrated," for example, is a compound proposition formed from complex propositions.

The logical analysis of inferences from complex and compound propositions has been brought to a satisfactory level only in the 20th century by the recent development of the necessary formal methods and powerful calculi by modern mathematical logicians.

Interpretation of syllogistic as an axiomatic deductive system. Aristotle's methods of reduction achieved something approaching an informal axiomatic system of syllogistic. Leibniz improved upon this method by showing how the laws of conversion and subalternation may be proved syllogistically using syllogisms of the second and third figures, and how these syllogisms, in turn, may be derived indirectly from the perfect moods. But neither of them stated rigorously the logical laws according to which these reductions and proofs were made. A formalized axiomatic system of syllogistic has, in fact, been made possible only by the methods of modern logic.

Auxiliary theory and base. For such an axiomatic representation of the syllogistic system, the following main elements are needed: As an auxiliary theory, the laws of the logic of propositions are employed. The logician may either accept as axioms (*i.e.,* without proof) the three to six axioms of the propositional calculus as variously presented by Gottlob Frege, by Bertrand Russell, or by Jan Łukasiewicz, and deduce from them all of the propositional laws that he may need; or, alternatively, he may specify certain laws (usually a dozen or more) after discovering which he will actually use.

The system must, moreover, have a base from which to derive its theorems, a list that should specify (1) primitive expression (letters or other symbols); (2) rules of formation, according to which the logician may build up other expressions from the primitive ones; (3) axioms of the system; *i.e.,* propositions that are accepted without proof; (4) definitions, which enable the logician to substitute one form of expression for another (optional); and (5) rules of inference, which state all of the steps he may take in deriving further expressions from axioms or from other expressions already derived.

Elements of the base

For the axiomatization of syllogistic, the following list (from Łukasiewicz) contains a sufficient base:

A. Primitive expressions:
1. Three term variables, *a, b,* and *m.*
2. Primitive syllogistic operators *A* and *I,* each of which may be joined to any two of the term variables.

B. Rules of formation (for well-formed formulas):
1. An expression consisting of *A* or *I* followed by any two term-variables *a, b,* or *m* is a well-formed formula or a syllogistic expression, designated by alpha (*α*), beta (*β*), and so on.
2. If *α* is a syllogistic expression, *Nα* (read "Not-*α*") is also a syllogistic expression.
3. If *α* and *β* are syllogistic expressions, $\alpha \cdot \beta$ and $\alpha \supset \beta$ are also syllogistic expressions.

C. Axioms of the system:
1. *Aaa*
2. *Iaa*
3. *Ama · Abm ⊃ Aba(Barbara)*
4. *Ama · Imb ⊃ Iba(Datisi)*

D. Definitions:
1. $Eba =_{df} NIba$
2. $Oba =_{df} NAba$

E. Rules of inference (for deriving expressions from others):
1. If *α* is an axiom or a derived expression (theorem) of the system, so is any expression that is the result of replacing one term-variable of *α* by another term-variable in all places of its occurrence in *α* (rule of substitution for term variables).
2. If *α* is a law of propositional logic, the result of replacing any propositional variable *p, q, r* . . . in *α* by a syllogistic expression may be asserted as a theorem provided that the same expression is inserted in all places of its occurence (rule of substitution for propositional variables).
3. In any derived expression, the expressions that are defined by rules D1 and D2 may be substituted for any occurrence of the expressions that define them (rule of substitution for definitions). This rule does not require, however, that the replacement cover all occurrences.
4. If *α* and $\alpha \supset \beta$ are derived expressions, *β* is a theorem (rule of detachment). This rule is an application of *modus ponendo ponens* (see above *Noncategorical syllogistic inferences*).

The rules of formation generate all of the syllogistic expressions, without regard to their validity or invalidity. To select from these expressions those that are valid (theorems of the system), the foregoing axioms and rules of inference are sufficient. To show that all 24 valid syllogisms may be derived formally, it is sufficient to show how *Barbara, Celarent,* the rules of contradictory opposition, and the rules of conversion can be derived in the system, since this is the basis used by Aristotle.

Deductive power of axioms

Properties of syllogistic as a system. It is interesting to note what deductive power is possessed by each of the axioms in the system; *i.e.,* which theorems would be lost by dispensing with a given axiom and what other expression might serve as a substitute for it. Axiom C1 ("*Aaa*") is needed to derive the theorem "*Iba ⊃ Iab*" and the other rules of conversion. Since the derivations of exactly half of the 24 moods depend on conversion, only the rules of opposition and 12 of the syllogisms would remain as theorems. In the place of C1, one might add C5, "*Iba ⊃ Iab,*" as an axiom to restore the full deductive power of the set of axioms. Without axiom C2 ("*Iaa*"), which guarantees that all of the terms are nonempty (*i.e.,* have application to something), the logician is left without the rules of subalternation and the other rules of opposition (except contradiction). Therefore, with axiom C2 would go all of the nine moods of the syllogism that depend on subalternation for their proof. Addition of the theorem C6 *Aba ⊃ Iba* instead of C2, however, would restore full deductive power to the system. Leaving out axiom C3 (*Barbara*) would cost only two syllogisms besides *Barbara,* viz., *Baroco* and *Bocardo,* each of which could thus serve as a substitute for C3. Since axiom C4 (*Datisi*) is needed in the proofs of C5 and C6, the result of leaving it out would be very damaging; it would equal the results of dropping C1 and C2 together, so that 9 + 12 = 21 syllogisms would

be blocked, leaving only *Barbara, Baroco,* and *Bocardo.* Any one of the syllogisms *Ferio, Cesare, Dimaris, Fresison,* or *Camenes,* however, could, taken as an axiom, repair the damage. If all four syllogistic operators *A, E, I,* and *O* were admitted instead of just two (*A* and *I*), the logician could also take *Barbara* and *Celarent,* as Aristotle did, as axioms together with C1 and C2. If, instead of C1 and C2, he would take C5 and C6 as axioms, the resulting set of axioms would be the base corresponding best to Aristotle's approach.

Since derivations in the system of axiomatic syllogistic depend entirely on rules for the manipulation of concatenations of signs, and not on any intuitive insights or knowledge of the meaning of these signs, it may be regarded as a formal deductive system in the modern technical sense. Thus, the formal properties of such a system can be investigated without giving the symbols any particular interpretation. One such formal property, internal consistency, is attributed to a system if it is impossible to derive a contradiction in it. Another formal question, the independence of its axioms, is at issue when the logician asks whether one of the axioms is redundant because it can be proved from the other axioms. Finally, the logician may ask whether the system is complete, in the sense that it is always possible, for any well-formed formula α, to devise a proof either that it is a theorem or that its negation is a theorem.

Consistency of a system

If a system is inconsistent, it is, for all purposes of deduction, without value, since any proposition whatsoever must then be derivable in it. That this is so is easily shown: A contradiction must be of the form "$p \cdot \bar{p}$." If "$p \cdot \bar{p}$" is a theorem, then according to the propositional law "$p \cdot \bar{p} \supset p$," one infers, by the rule of detachment, p. With p, it follows also that "$p \vee q$"—because "$p \supset (p \vee q)$" is again a law of propositional logic. If "$p \vee q$" is a theorem, however, it is possible to derive $\bar{p}$ again from "$p \cdot \bar{p}$" and to have, by an application of *modus tollendo ponens,* also q. This means that in an inconsistent system any proposition that one likes is derivable as a theorem, and the system is therefore deductively useless.

A system is sufficiently shown to be consistent if its axioms are consistent; for the rules of inference guarantee that no false proposition will be derivable from true ones. Therefore, in case of inconsistency anywhere, the axioms themselves already must have been inconsistent. Hence, a proof of the consistency of a formal system must cite an interpretation for the axioms that makes them all true propositions. For the system of syllogistic, various such interpretations, mostly from arithmetic or propositional logic, have been given.

Independence and completeness

In a similar way—and in a manner resembling Aristotle's method of invalidating moods by counterexamples—it may be shown that each of the axioms is independent of the others (*i.e.,* is not derivable from them) by producing an interpretation that makes the axiom in question false and the other axioms true. It is then obvious that this axiom cannot logically follow from the others, since its falsity would be incompatible with their joint truth. Simple proofs of independence have been devised—with interpretations taken from the propositional calculus and arithmetic—for all of the axioms except *Barbara,* for which the proof is complicated. These proofs have shown that none of the axioms can be regarded as being the one and only principle of syllogistic.

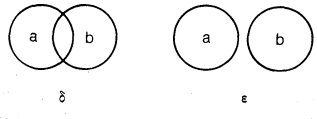
The preceding system is strong enough to derive all valid theorems of syllogistic from the stated axioms. For completeness, however, it is further required that a formal proof be provided for the negation of any invalid syllogistic expression. But, since this cannot be done, the system is not complete. The system may be strengthened by adding two axioms, viz., the negations of "*AAI*" and "*EEI*" in the second figure, and certain rules of rejection, corresponding to the rules of substitution and the rule of detachment for inference. Though with these additions to the base, it is possible to reject all invalid syllogistic moods, it is still not possible to reject such invalid syllogistic expressions as those having more than three terms and some of those having two terms. By an ingenious method, devised by the Polish logician Jerzy Słupecki, a stronger rule of rejection has been formulated that makes the system complete.

Extensions of syllogistic. *Singular and negative terms.* Expressions like "Socrates" or "the founder of psychoanalysis," called singular terms because they refer to but one unique particular, yield propositions that are formally different from general ones. All syllogistic laws, however, are formulated for general propositions only. Though Aristotle seems to have thought that singular propositions were a special kind of universal proposition, the rules of opposition fail for them and they cannot be converted or changed by contraposition, because proper names can only take the subject place in a proposition. "Socrates" in "Socrates is mortal" can be replaced, of course, by "Everything identical with Socrates," which defines a class of only one member; but this solution does not really introduce singular terms into syllogistic but keeps them out while taking care of all statements that contain them. Moreover, if proper names can only replace subject terms, the only allowable syllogisms incorporating singular premises are those in which singular terms occur twice in the subject place. One may therefore substitute singular terms instead of b in the first and second figures and in place of m in the third figure. These requirements limit the syllogisms that can have singular terms to *Barbara, Celarent, Cesare, Camestres, Darapti,* and *Felapton.*

Confinement to subject position

Aristotelian syllogistic avoids negative terms like $\bar{a}$, $\bar{b}$, and $\bar{m}$. The modern way of expressing syllogistic propositions, however, is to treat *E*- and *O*-propositions as *A*- and *I*-propositions with a negated predicate term. Because the modern functional calculus allows the negation of subject terms as well, the foreging syllogistic can only cover a fragment of the class of possible expressions of predicate logic. The question arises, therefore, whether the axiom system for syllogistic might be extended to cover propositions with negated term-variables—which would make the so-called exclusive propositions, such as "Only students are members," and exceptive propositions, such as "All but (except) heroes are frightened," also amenable to syllogistic treatment. The exclusive proposition can be translated into "Every nonstudent is a nonmember" and represented as "$A\bar{b}\bar{a}$," which is equivalent to "*Aab*" by contraposition. Moreover, the exceptive proposition, being translatable into the conjunctive proposition "Every nonhero is frightened and no hero is frightened," is of the form "$A\bar{b}a \cdot Ab\bar{a}$," which amounts to saying that a and b are disjunctive terms in the mutually exclusive sense. Because the terms may be combined in four ways, ba, $b\bar{a}$, $\bar{b}a$ and $\bar{b}\bar{a}$, each of which can take the four operators *A, E, I,* and *O,* there are 16 logically different forms of syllogistic propositions and 16 further converse forms; *i.e.,* 32 forms of propositions in this system. If one now disqualifies empty terms, he must also exclude universal terms, which apply to everything (like "self-identical" or "mentionable") since, if a is a universal term, $\bar{a}$ must be empty. A problem now arises in the interpretation of the Euler diagrams (δ) and (ε):

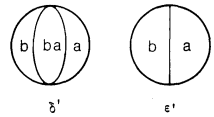
Exclusion of universal terms



δ ε

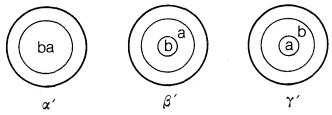
There can be objects that are $\bar{b}$ and objects that are $\bar{a}$; but it is possible that there are no objects that are both $\bar{b}$ and $\bar{a}$, because b and a might be complementary classes that between them exhaust the universe. There may then be cases of "$Ib\bar{a}$" and "$I\bar{b}a$," but not of "$I\bar{b}\bar{a}$," or there may be cases in which no $\bar{b}$ is an $\bar{a}$ (*i.e.,* "$E\bar{b}\bar{a}$," which is equivalent to "$A\bar{b}a$"; *i.e.,* to "Everything that is not b is a," there being nothing else in the universe but things that are a or b).

Two more diagrams must therefore be introduced, presenting the special subcases of (δ) and (ε):



δ' ε'

in which the circle represents the total universe of discourse. To bring the rest of the diagrams up to this level of abstraction, such a circle should then be added also to the diagrams (α)–(γ):



By introducing negative terms, the rules for the validity of syllogisms are somewhat altered. Since "*Aba*" is equivalent to "*Ebā,*" the valid syllogism "$Em\bar{a} \cdot Eb\bar{m} \supset Eb\bar{a}$" arises. Moreover, "$Em\bar{a} \cdot Eb\bar{m} \supset Aba$." Since these syllogistic forms seem to violate the rules of quality, it should be stipulated that those rules should apply only to a syllogistic without negative terms. An axiom system called *CS*(*n*) has been devised, which differs from the system described above by admitting negative terms, by dropping *Barbara* from the axioms, by using as definitions "$Eba =_{df} Ab\bar{a}$" and "$Iba =_{df} NAb\bar{a}$," and by having a special rule of inference for double negation.

Modifications owing to emptiness

Empty terms and existential import. The fact that no empty terms may replace any of the term-variables is a deficiency of traditional syllogistic that calls for certain precautions. In the mid-20th century, for example, no one knew whether gravitation waves existed; but, from "Every gravitation wave is a physical phenomenon," there is a danger that, without this ban on empty terms, the proposition "Some gravitation wave is a physical phenomenon" might be inferred. This statement, however, implies by conversion that "Some physical phenomenon is a gravitation wave," which in turn would most naturally be interpreted as asserting the existence of gravitation waves.

In order to admit terms regardless of their emptiness or nonemptiness, certain revisions of the earlier laws are required. It is important to specify, for example, which syllogistic expressions are to be false if containing empty terms. Three different specifications have been discussed by logicians: (1) that *A*-, *I*-, and *O*-propositions, but not *E*-propositions, presuppose the existence of both terms for their truth; (2) that *A*- and *I*-propositions, but not *E*- and *O*-propositions, presuppose nonemptiness of their terms; (3) that *I*- and *O*-propositions presuppose nonemptiness, whereas *A*-and *E*-propositions do not.

In cases (1) and (2), all of the laws *CS*(*n*), with the exception of inversion and contraposition ("$Aba \supset Ib\bar{a}$" and "$Aba \supset A\bar{a}\bar{b}$") would remain valid. Since these theorems are not used by *CS*(*n*) for the derivation of syllogisms, however, all syllogistic moods remain. Specification (3) has much more to commend it. *A*- and *E*-propositions may be defined as negations of the corresponding *O*- and *I*-propositions. But "*Iba*" is most naturally interpreted as asserting the existence of objects that are both *b* and *a* and must therefore be false if *a* or *b* is empty. The same applies, correspondingly, to "*Oba*" in relation to *b* and *ā*. Since "*Aba*" and "*Eba*" are negations of "*Oba*" and "*Iba,*" respectively, they must be true if "*Iba*" and "*Oba*" happen to be false through lack of reference of their terms. In this way, the laws of contradictory opposition remain valid. But subalternation is invalid: "*Aba*" may be true merely because "*Oba*" fails for lack of existential reference of *b*; but in this case "*Iba*" is also false. Contrary opposition and subcontrary opposition now vanish from the logical square: "*Aba*" and "*Eba*" may both be true and "*Oba*" and "*Iba*" may both be false in case there is a lack of reference for *b*. The most important law that is sacrificed, however, is "$Aba \supset Iba$," since the derivation of all moods having universal premises and particular conclusions depend on this law. The effect is the same as that of dropping the axiom "*Iaa*" from the system described earlier: *Barbari, Camestrop, Felapton, Celaront, Cesaro, Darapti, Bramantip, Fesapo,* and *Camenop* are then lost and only 15 moods remain. Since specification (3) for the existential import of *A*-, *E*-, *I*-, and *O*-propositions is equivalent to those adopted for general and particular propositions in the modern predicate calculus, the 15 remaining moods are the theorems of the syllogistic when interpreted in the predicate calculus. This is not to say, however, that Aristotle's syllogistic is incorrect, but only that he and his followers never formulated the existence postulates for their syllogistic terms. If these were made explicit and added to the premises of all syllogistic laws, they would all again become laws of predicate logic also.

Boolean algebra and Venn diagrams

When negative and empty terms are admitted, syllogistic becomes a special case of Boolean algebra, which can be defined as any system incorporating the concepts of a universal class, of a null class, and of the complement of a class, and those of class union and class intersection. The null class is denoted by 0, the universal class by 1; and if *a* is a class, $\bar{a}$ is its complement class. Emptiness is expressed by saying that the class is identical with the null class or that "$a = 0$"; universality, expressing the fact that the term *a* applies to everything, is represented by "$a = 1$." Clearly, if a term is empty, its complement class is identical with the universal class. The class that contains the common members of the terms *a* and *b* is denoted by $a \cap b$ and is called the intersection of *a* and *b*. The union of classes *a* and *b*, expressed by $a \cup b$, is the class of all of those things that belong either to *a* or to *b*. If two classes have the same objects as members—that is, if *a* and *b* are extensionally identical—then this fact is expressed by "$a = b$." Nonidentity of classes is expressed by "$a \neq b$."

In this language, the syllogistic propositions may now be expressed in the following way:

$$Aba =_{df} b \cap \bar{a} = 0$$
$$Eba =_{df} b \cap a = 0$$
$$Iba =_{df} b \cap a \neq 0$$
$$Oba =_{df} b \cap \bar{a} \neq 0.$$

Barbara, in this notation, is then:

$$(m \cap \bar{a} = 0 \cdot b \cap \bar{m} = 0) \supset b \cap \bar{a} = 0$$

and *Datisi* becomes:

$$(m \cap \bar{a} = 0 \cdot m \cap b \neq 0) \supset b \cap a \neq 0.$$

Venn proofs of *Barbara* and *Datisi*

To test such Boolean syllogisms for validity, the Venn diagrams, so-called after the British logician John Venn, who devised them in 1880, provide an ingenious method. He represents the extension of the three terms by three interlocking circles as shown in Figure 4. A region that is said to be empty in a premise is shaded out; and a region said to be not empty is marked by a cross inside the region. The Venn diagram for *Barbara* is shown in Figure 5. Its first premise states that the intersection of *m* and $\bar{a}$ is empty; therefore, one shades out all of *m* that is outside of *a*. The second premise states that the intersection of *b*

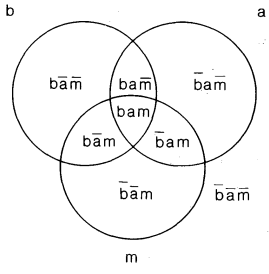


Figure 4: Basis of the Venn diagrams.

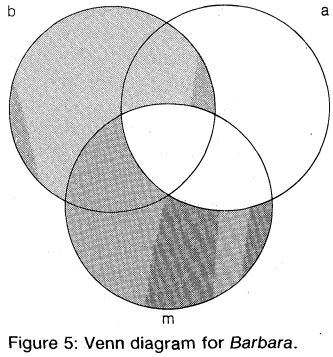


Figure 5: Venn diagram for *Barbara*.

and $\bar{m}$ is empty; therefore, one shades out all of b that is outside of m. The conclusion, finally, states that the intersection of b and $\bar{a}$ is empty. If, accordingly, the logician wants to shade out the region of b that is outside of a, he finds that this has been done already by the stipulations of the premises: the conclusion states nothing that has not been stated by the premises taken together; therefore, *Barbara* is valid.

The proof of *Datisi* proceeds as follows (see Figure 6). The first premise is illustrated by shading out the region of m that is outside of a. The second premise requires that a cross be placed in the region where m and b overlap. Since part of this region has been shaded out already, the cross can be put only into the part still left white; *i.e.*, the part where m and a and b overlap. The conclusion calls for a cross to be put into the overlapping region of a and b. The logician finds, however, that there is a cross already in this part; therefore *Datisi* is valid. (Gü.P./Ed.)

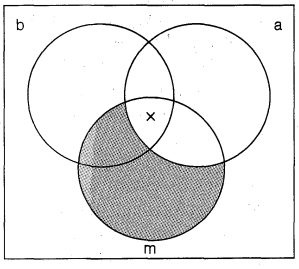


Figure 6: Venn diagram for *Datisi*.

MODAL LOGIC

True propositions can be divided into those—like "2 + 2 = 4"—that are true by logical necessity (*necessary* propositions), and those—like "France is a republic"—that are not (*contingently true* propositions). Similarly, false propositions can be divided into those—like "2 + 2 = 5"—that are false by logical necessity (*impossible* propositions), and those—like "France is a monarchy"—that are not (*contingently false* propositions). Contingently true and contingently false propositions are known collectively as *contingent* propositions. A proposition that is not impossible (*i.e.*, one that is either necessary or contingent) is said to be a *possible* proposition. Intuitively, the notions of necessity and possibility are connected in the following way: to say that a proposition is necessary is to say that it is not possible for it to be false; and to say that a proposition is possible is to say that it is not necessarily false.

If it is logically impossible for a certain proposition, p, to be true without a certain proposition, q, being also true (*i.e.*, if the conjunction of p and not-q is logically impossible), then it is said that *p strictly implies q*. An alternative, equivalent way of explaining the notion of strict implication is by saying that p strictly implies q if and only if it is necessary that p materially implies q. "John's tie is scarlet," for example, strictly implies "John's tie is red," because it is impossible for John's tie to be scarlet without being red (or: it is necessarily true that if John's tie is scarlet it is red); and in general, if p is the conjunction of the premises, and q is the conclusion, of a deductively valid inference, p will strictly imply q.

The notions just referred to—necessity, possibility, impossibility, contingency, strict implication—and certain other closely related ones are known as *modal* notions; and a logic designed to express principles involving them is called a *modal logic*.

The most straightforward way of constructing such a logic is to add to some standard nonmodal system a new primitive operator intended to represent one of the modal notions mentioned above, to define other modal operators in terms of it, and to add certain special axioms or transformation rules or both. A great many systems of modal logic have been constructed; but attention will be restricted here to a few closely related ones in which the underlying nonmodal system is ordinary PC.

Alternative systems of modal logic. All of the systems to be considered here have the same *wff*'s, but differ in their axioms. The *wff*'s can be specified by adding to the symbols of PC a primitive monadic operator L, and to the formation rules of PC the rule that if α is a *wff*, so is $L\alpha$. L is intended to be interpreted as "It is necessary that," so that Lp will be true if and only if p is a necessary proposition. The monadic operator M and the dyadic operator $\dashv$ (to be interpreted as "It is possible that" and "strictly implies," respectively) can then be introduced by the following definitions, which reflect in an obvious way the informal accounts given above of the connections between necessity, possibility, and strict implication: if α is any *wff*, then $M\alpha$ is to be an abbreviation of $\sim L\sim\alpha$; and if α and β are any *wff*'s, then $\alpha \dashv \beta$ is to be an abbreviation of $L(\alpha \supset \beta)$ [or alternatively of $\sim M(\alpha \cdot \sim\beta)$].

The system T and its limitations

The modal system known as T has as axioms some set of axioms adequate for PC (such as those of PM), and in addition

$$(1)\ Lp \supset p$$
$$(2)\ L(p \supset q) \supset (Lp \supset Lq).$$

Axiom (1) expresses the principle that whatever is necessarily true is true and (2) the principle that, if q logically follows from p then, if p is a necessary truth, so is q (*i.e.*, that whatever follows from a necessary truth is itself a necessary truth). These two principles appear to have a high degree of intuitive plausibility, and almost all modal systems contain (1) and (2) as theorems. The transformation rules of T are *uniform substitution, modus ponens,* and a third rule to the effect that if α is a theorem so is $L\alpha$ (the *rule of necessitation*). The intuitive rationale of this last rule is that in a sound axiomatic system it is expected that every instance of a theorem α will be not merely true but necessarily true—and in that case every instance of $L\alpha$ will be true.

Among the simpler theorems of T are:

$$p \supset Mp;\ L(p \cdot q) \equiv (Lp \cdot Lq);\ M(p \vee q) \equiv (Mp \vee Mq);(Lp \vee Lq) \supset L(p \vee q)$$

(but not its converse);

$$M(p \cdot q) \supset (Mp \cdot Mq)$$

(but not its converse);

$$LMp \equiv \sim ML\sim p;\ (p \dashv q) \supset (Mp \supset Mq);\ (\sim p \dashv p) \equiv Lp;\ L(p \vee q) \supset (Lp \vee Mq).$$

There are many modal formulas that are not theorems of T but that have a certain claim to express truths about necessity and possibility. Among these are

$$Lp \supset LLp;\ Mp \supset LMp;\ \text{and}\ p \supset LMp.$$

The first of these means that if a proposition is necessary, its being necessary is itself a necessary truth; the second means that if a proposition is possible, its being possible is a necessary truth; and the third means that if a proposition is true, then not merely is it possible but its being possible is a necessary truth. These are all various elements in the general thesis that a proposition's having the modal characteristics it has (such as necessity, possibility) is not a contingent matter but is determined by logical considerations. Though this thesis may be philosophically controversial, it is at least plausible, and its consequences are worth exploring; and one way of exploring them is to construct modal systems in which the formulas listed above are theorems. None of these formulas, as was said, is a theorem of T; but each could be consistently added to T as an extra axiom to produce a new and more extensive system. The system obtained by adding $Lp \supset LLp$ to T is known as S4; that obtained by adding $Mp \supset LMp$ to T is known as S5; and the addition of $p \supset LMp$ to T gives the Brouwerian system, here called B for short.

The systems S4, S5, and B

The relations between these four systems are as follows: S4 is stronger than T; *i.e.*, it contains all of the theorems of T and others besides. B is also stronger than T. S5 is stronger than S4 and also stronger than B. S4 and B, however, are independent of each other in the sense that each contains some theorems that the other does not have. It is of particular importance that if $Mp \supset LMp$ is added to T then $Lp \supset LLp$ can be derived as a theorem, but

that if one merely adds the latter to T the former cannot then be derived.

Examples of theorems of S4 that are not theorems of T are: $Mp \equiv MMp$; $MLMp \supset Mp$; $(p \dashv q) \supset (Lp \dashv Lq)$; and examples of theorems of S5 that are not theorems of S4 are:

$$Lp \equiv MLp;\ L(p \vee Mq) \equiv (Lp \vee Mq);\ M(p \cdot Lq) \equiv (Mp \cdot Lq);\ (Lp \dashv Lq) \vee (Lq \dashv Lq).$$

One important feature of S5 but not of the other systems mentioned is that any *wff* that contains an unbroken sequence of monadic modal operators (L's and/or M's) is provably equivalent to the same *wff* with all of these operators deleted except the last.

Considerations of space preclude an account of the many other axiomatic systems of modal logic that have been investigated. Some of these are weaker than T: such systems normally contain the axioms of T either as axioms or as theorems but have only a restricted form of the *rule of necessitation.* Another group comprises systems that are stronger than S4 but weaker than S5; some of these have proved fruitful in developing a logic of temporal relations. Yet another group includes systems that are stronger than S4 but weaker than S5; some of these have proved fruitful in developing a logic of temporal relations. Yet another group includes systems that are stronger than S4 but independent of S5 in the sense explained above.

Modal predicate logics can be formed also by making analogous additions to LPC instead of to PC.

Validity in modal logic. The task of defining validity for modal *wff*'s is complicated by the fact that, even if the truth values of all of the variables in a *wff* are given, it is not obvious how one should set about calculating the truth value of the whole *wff.* Nevertheless, a number of definitions of validity applicable to modal *wff*'s have been given, each of which turns out to match some axiomatic modal system in the sense that it brings out as valid those *wff*'s, and no others, that are theorems of that system. Most, if not all, of these accounts of validity can be thought of as variant ways of giving formal precision to the idea that necessity is truth in every "possible world" or "conceivable state of affairs." The simplest such definition

Models for the four systems

is this: Let a *model* be constructed by first assuming a (finite or infinite) set W of "worlds." In each world, independently of all the others, let each propositional variable then be assigned either the value 1 or the value 0. In each world the values of truth functions are calculated in the usual way from the values of their arguments in that world. In each world, however, La is to have the value 1 if a has the value 1, not only in that world but in every other world in W as well, and is otherwise to have the value 0; and in each world Ma is to have the value 1 if a has value 1 either in that world or in some other world in W, and is otherwise to have the value 0. These rules enable one to calculate a value (1 or 0) in any world in W for any given *wff,* once the values of the variables in each world in W are specified. A model is defined as consisting of a set of worlds together with a value assignment of the kind just described. A *wff* is *valid* if and only if it has the value 1 in every world in every model. It can be proved that the *wff*'s that are valid by this criterion are precisely the theorems of S5; for this reason models of the kind here described may be called S5-models, and validity as just defined may be called S5-validity.

A definition of T-validity (*i.e.,* one that can be proved to bring out as valid precisely the theorems of T) can be given as follows: a T-model consists of a set of worlds W and a value assignment to each variable in each world, as before; but it also includes a specification, for each world in W, of some subset of W as the worlds that are "accessible" to that world. Truth functions are evaluated as before; but in each world in the model, La is to have the value 1 if a has value 1 in that world and in every other world in W accessible to it, and is otherwise to have value 0. And in each world, Ma is to have the value 1 if a has value 1 either in that world or in some other world accessible to it and is otherwise to have value 0. (In other words, in computing the value of La or Ma in a given world, no account is taken of the value of a in any other world not accessible to it.) A *wff* is T-valid if and only if it has the value 1 in every world in every T-model.

An S4-model is defined as is a T-model except that it is required that the accessiblity relation be transitive; *i.e.,* that where w_1, w_2, and w_3 are any worlds in W, if w_1 is accessible to w_2 and w_2 is accessible to w_3, then w_1 is accessible to w_3. A *wff* is S4-valid if and only if it has the value 1 in every world in every S4-model. The S4-valid *wff*'s can be shown to be precisely the theorems of S4. Finally, a definition of validity is obtained that will match the system B by requiring that the accessibility relation be symmetrical but not that it be transitive.

Adaptations to other systems

For all four systems, effective decision procedures for validity can be given. Further modifications of the general method described have yielded validity definitions that match many other axiomatic modal systems; and the method can be adapted to give a definition of validity for Intuitionistic PC. For a number of axiomatic modal systems, however, no satisfactory account of validity has been devised. Validity can also be defined for various modal predicate logics by combining the definition of LPC-validity given in *Validity in LPC* with the relevant accounts of validity for modal systems; but a modal logic based on LPC is, like LPC itself, an undecidable system.

SET THEORY

Only a sketchy account of set theory is given here; for further information see the article SET THEORY. Set theory is a logic of classes; *i.e.,* of collections (finite or infinite) or aggregations of objects of any kind, which are known as the *members* of the classes in question. Some logicians use the terms "class" and "set" interchangeably; others distinguish between them, defining a set (for example) as a class that is itself a member of some class and defining a proper class as one that is not a member of any class. It is usual to write "$\in$" for "is a member of" and to abbreviate $\sim(x \in y)$ to $x \notin y$. A particular class may be specified either by listing all of its members or by stating some condition of membership, in which (latter) case the class consists of all of those things and only those that satisfy that condition (it is used, for example, when one speaks of the class of inhabitants of London or the class of prime numbers). Clearly, the former method is available only for finite classes and may be highly inconvenient even then; the latter, however, is of more general applicability. Two classes that have precisely the same members are regarded as the same class or are said to be *identical* with each other, even if they are specified by different conditions; *i.e.,* identity of classes is identity of membership, not identity of specifying conditions. This principle is known as the *principle of extensionality.* A class with no members, such as the class of Chinese popes, is said to be *null.* Since the membership of all such classes is the same, there is only one null class, which is therefore usually called *the* null class (or sometimes the empty class); it is symbolized by Λ or φ. The notation $x = y$ is used for "x is identical with y" and $\sim(x = y)$ is usually abbreviated to $x \neq y$. The expression $x = \Lambda$ therefore means that the class x has no members, and $x \neq \Lambda$ means that x has at least one member.

Interrelations of classes

A member of a class may itself be a class. The class of dogs, for example, is a member of the class of species of animals. An individual dog, however, though a member of the former class is not a member of the latter—because an individual dog is not a species of animal (if the number of dogs increases, the number of species of animals does not thereby increase). Class membership is therefore not a transitive relation. The relation of class inclusion, however (to be carefully distinguished from class-membership), is transitive. A class x is said to be *included* in a class y (written: $x \subseteq y$) if and only if every member of x is also a member of y. (This is not meant to exclude the possibility that x and y may be identical.) If x is included in but is not identical with y—*i.e.,* if every member of x is a member of y but some members of y are not members of x—x is said to be *properly included* in y (written: $x \subset y$).

It is perhaps natural to assume that for every statable condition there is a class (null or otherwise) of objects that satisfy that condition. This assumption is known as the *principle of comprehension.* In the unrestricted form

just mentioned, however, this principle has been found to lead to inconsistencies and hence cannot be accepted as it stands. One statable condition, for example, is *non-self-membership; i.e.,* the property possessed by a class if and only if it is not a member of itself. This in fact appears to be a condition that most classes do fulfill; the class of dogs, for example, is not itself a dog and hence is not a member of the class of dogs.

Russell's paradox: theory of types

Let it now be assumed that the class of all classes that are not members of themselves can be formed and let this class be y. Then any class x will be a member of y if and only if it is not a member of itself; *i.e.,* for any class x, $(x \in y) \equiv (x \notin x)$. The question can then be asked whether y is a member of itself or not, with the following awkward result: If it is a member of itself, then it fails to fulfill the condition of membership of y; and hence it is not a member of y; *i.e.,* not a member of itself. On the other hand, if y is not a member of itself, then it does fulfill the required condition; and therefore it *is* a member of y—*i.e.,* of itself. Hence the equivalence $(y \in y) \equiv (y \notin y)$ results, which is self-contradictory. This perplexing conclusion, which was pointed out by Bertrand Russell, is known as *Russell's paradox.* Russell's own solution to it and to other similar difficulties was to regard classes as forming a hierarchy of types and to posit that a class could only be regarded sensibly as a member, or a nonmember, of a class at the next higher level in the hierarchy. The effect of this theory is to make $x \in x$, and therefore $x \notin x$, ill-formed. Another kind of solution, however, is based upon the distinction made earlier between two kinds of classes, those that are sets and those that are not—a set being defined as a class that is itself a member of some class. The unrestricted *principle of comprehension* is then replaced by the weaker principle that for every condition there is a class the members of which are the individuals or *sets* that fulfill that condition. Other solutions have also been devised; but none has won universal acceptance, with the result that several different versions of set theory are found in the literature of the subject.

Formally, set theory can be derived by the addition of various special axioms to a rather modest form of LPC that contains no predicate variables and only a single primitive dyadic predicate constant ($\in$) to represent membership. Sometimes LPC-with-identity is used and there are then two primitive dyadic predicate constants ($\in$ and $=$). In some versions the variables $x, y, \ldots$ are taken to range only over sets or classes; in other versions they range over individuals as well. The special axioms vary; but the basis normally includes the *principle of extensionality* and some restricted form of the *principle of comprehension,* or some elements from which these can be deduced.

Notation, definitions, theorems

A notation to express theorems about classes can be either defined in various ways (not detailed here) in terms of the primitives mentioned above, or else introduced independently. The main elements of one widely used notation are the following: if α is an expression containing some free occurrence of x, the expression $\{x{:}\alpha\}$ is used to stand for the class of objects fulfilling the condition expressed by α—*e.g.,* $\{x$: "x is a prime number"$\}$ represents the class of prime numbers; $\{x\}$ represents the class the only member of which is x; $\{x, y\}$ the class the only members of which are x and y; and so on. $\langle x, y\rangle$ represents the class the members of which are x and y *in that order* (thus $\{x, y\}$ and $\{y, x\}$ are identical; but $\langle x, y\rangle$ and $\langle y, x\rangle$ are in general not identical). Let x and y be any classes, as (for example) those of the dots on the two arms of a stippled cross. The *intersection* of x and y, symbolized as $x \cap y$, is the class the members of which are the objects common to x and y—in this case the dots within the area where the arms cross—*i.e.,* $\{z : z \in x \cdot z \in y\}$. Similarly, the *union* of x and y, symbolized as $x \cup y$, is the class the members of which are the members of x together with those of y—in this case all of the dots on the cross—*i.e.,* $\{z : z \in x \vee z \in y\}$; the *complement* of x, symbolized as $-x$, is the class the members of which are all those objects that are not members of x—*i.e.,* $\{y : y \notin x\}$; the complement of y in x, symbolized as $x - y$, is the class of all objects that are members of x but not of y—*i.e.,* $\{z : z \in x \cdot z \notin y\}$; the *universal* class, symbolized as Λ, is the class of which everything is a member, definable as the complement of the null class—*i.e.,* as $-\Lambda$. Λ itself is sometimes taken as a primitive individual constant, sometimes defined as $\{x : x \neq x\}$— the class of objects that are not identical with themselves.

Among the simpler theorems of set theory are:

$$(\forall x)(x \cap x = x);\ (\forall x)(\forall y)(x \cap y = y \cap x);$$

and corresponding theorems for $\cup$;

$$(\forall x)(\forall y)(\forall z)[x \cap (y \cap z) = (x \cap y) \cup (x \cap z)];$$
$$(\forall x)(\forall y)[-(x \cap y) = -x \cup -y];$$

and corresponding theorems with $\cap$ and $\cup$ interchanged;

$$(\forall x)\,(--x = x);\ (\forall x)(\forall y)(x - y = x \cap -y);$$
$$(\Lambda x)\,(\Lambda \subset x);\ (\forall x)(x \cap \Lambda = \Lambda);\ (\forall x)(x \cup \Lambda = x).$$

In these theorems the variables range over classes. In several cases, there are obvious analogies to valid *wff*'s of PC. For an account of further developments, more advanced axioms, etc., see SET THEORY.

Apart from its own intrinsic interest, set theory has an importance for the foundations of mathematics in that it is widely held that the natural numbers can be adequately defined in set-theoretic terms. Moreover, given suitable axioms, standard postulates for natural-number arithmetic can be derived as theorems within set theory.

LOGICAL FOUNDATIONS OF MATHEMATICS

It is possible to develop arithmetic in the form of a logical system without recourse to set theory.

Axiomatization of arithmetic. Axiomatic bases for the arithmetic of the natural numbers $(0, 1, 2, \ldots)$ can be constructed as extensions of LPC designed to have a direct numerical interpretation. One such basis, with the intended interpretation indicated, is the following: in addition to the usual operators and quantifiers, the symbols comprise (a) an infinite list of individual variables, the values of which are to be natural numbers; (b) one individual constant, 0, to represent zero; (c) one dyadic predicate constant, $=$, to represent equality; (d) three function constants: a monadic one, $'$, interpreted as "the successor of"; and two dyadic ones, $+$ and $\cdot$, interpreted as "the sum of" and "the product of," respectively. In view of their intended interpretations, it is convenient to write $'$ after its argument, and $=$, $+$, and $\cdot$ between their arguments. The numeral 1 may be defined as $0'$, the numeral 2 as $1'$ (or $0''$), and so on. Terms are defined as explained in *Extensions of LPC*; and since there are no predicate variables, an atomic *wff* will simply be any expression of the form $(t_1 = t_2)$, where t_1 and t_2 are terms. The following special axioms are added to the general axioms for LPC:

Axiomatic basis and interpretations

(1) $(x = y) \supset [(x = z)(y = z)]$;
(2) $(x = y) \supset (x' = y')$;
(3) $0 \neq x'$;
(4) $(x' = y') \supset (x = y)$;
(5) $x + 0 = x$;
(6) $(x + y') = (x + y)'$;
(7) $x \cdot 0 = 0$;
(8) $x \cdot y' = (x \cdot y) + x$.

There is also the following special axiom schema:

(9) Let α be any *wff* and $\mathbf{a}$ any individual variable; let $\alpha(0)$ be α but with 0 replacing every free occurrence of $\mathbf{a}$, and $\alpha(\mathbf{a}')$ be α but with free $\mathbf{a}'$ replacing every free occurrence of $\mathbf{a}$: then

$$\alpha(0) \supset \{(\forall \mathbf{a})\ [\alpha \supset \alpha(\mathbf{a}')] \supset (\forall \sim)\alpha\}$$

is an axiom.

The transformation rules are *modus ponens* and *universal generalization* as before. Axiom schema (9) expresses the principle of *mathematical induction.* It may be paraphrased by saying that if a property (i) belongs to 0 and (ii) belongs to the successor of any natural number to which it also belongs, then it belongs to every natural number.

Scope and limitations of axiomatized arithmetic. In this system (which shall be called S), the axioms for identity given in *LPC-with-identity* are derivable as theorems. Since S contains no variables other than individual variables, every closed *wff*—*i.e.,* every *wff* containing no free

Propositions expressible in S

individual variables—will (given the intended interpretation) express not a proposition form but a determinate proposition of natural-number arithmetic—possibly one containing no variables at all, such as $3 + 5 = 8$, or possibly one containing bound variables, such as

$(\forall x)(\exists y)(4 + x = 3 + y)$ or
$(\forall x)(\forall y)(\forall z)[x \cdot (y + z) = (x \cdot y) + (x \cdot z)]$.

The range of arithmetical propositions expressible by *wff*'s of S is indeed much greater than might appear from the fact that its only primitive predicate is = and its only primitive function symbols are $'$, +, and ·, inasmuch as many relations and properties not so far explicitly mentioned, such as "is less than," "is divisible (without remainder) by," and "is prime," can be defined in terms of the primitives of S; *e.g.*, $x < y$ can be defined as $(\exists z)(x + z' = y)$. Moreover, even if a given numerical function is not directly definable in terms of the primitives of S, it may still be possible to express propositions involving that function as *wff*'s of S, in a way that can be sufficiently explained for present purposes as follows: A numerical function is one that always has a natural number as its value when its arguments are natural numbers. Now let $f(x)$ be such a function with one argument x. It may then be possible to find a *wff* of S, α, containing two free variables, x and y, such that (i) all values of x and y that make $f(x) = y$ true also turn α into a theorem of S, and (ii) whenever α' is formed from α by uniformly replacing x by a numeral, $(\exists^1 y)\ \alpha'$ is a theorem of S. In that case every true proposition of natural-number arithmetic involving only that function will be expressible as a *wff*, appropriately constructed from α, that is a theorem of S; and α is then said to *represent* that function in S. For functions with more than one argument, the above account can be extended in a straightforward way and, provided that a true natural-number proposition involves only functions that are representable in S, the *wff* that expresses it will be a theorem of S.

Effective computability and completeness

A numerical function is said to be *effectively computable* if there is an effective (mechanical and finite) method of calculating the value of the function in all cases when its arguments are given (*e.g.*, the addition function $x + y$ is effectively computable since, given any two numbers, it is possible to calculate their sum mechanically in a finite number of steps). Various attempts have been made to give a precise mathematical analysis of the conditions for effective computability, and it is widely held that a satisfactory analysis of this kind is provided in the theory of recursive functions; *i.e.*, that the (intuitively defined) class of effectively computable functions coincides exactly with the (mathematically defined) class of recursive functions. An account of recursive functions is provided in PHILOSOPHY OF THE BRANCHES OF KNOWLEDGE: *Foundations of mathematics* and is therefore omitted here. In the present context the essential point is that it can be proved that the numerical functions representable in S are precisely those that are recursive. And from this fact, given the assumption that effective computability and recursiveness do in fact coincide, it follows that every true proposition of natural-number arithmetic in which all of the functions are effectively computable can be expressed as a *wff* that is a *theorem* of S. In this sense, S is an adequate basis for natural-number arithmetic.

It would be natural to hope that S were also complete in the sense that every true arithmetical proposition whatever that can be expressed as a *wff* of S would be a theorem. It can be shown, however—at least if S is assumed to be consistent, which there seems no reason to doubt—that this hope is unfounded. That this is so is a direct consequence of the most important finding that has been made to date in this area, a finding known as *Gödel's incompleteness theorem.* An application of Gödel's argument to S shows how to construct a certain closed *wff* of S of which it can be proved both that it expresses a true arithmetical proposition, and also that it is not a theorem of S, if S is a consistent system. Moreover, contrary to what one might expect, the incompleteness of S could not be overcome by adding extra axioms or rules to it; for, although it would be easy to strengthen the basis of S so as to obtain a system S′ in which the *wff* in question is indeed a theorem, Gödel's reasoning would then enable one to construct another closed *wff* which would stand to S′ exactly as the previous one does to S. The position in fact is that any consistent extension of S is bound also to be incomplete, and this fact is often expressed by saying that S is "essentially incomplete." Nor could one obtain a complete system for arithmetic by starting afresh with some other axioms and rules in place of those of S, for the upshot of Gödel's theorem is that any axiomatic system whatever that is sufficient for natural-number addition and multiplication is essentially incomplete; and as a result the otherwise natural ideal of giving a complete axiomatization of the whole, or even of some considerable part, of pure mathematics must be abandoned.

Summarizing distinctions

An important threefold distinction therefore exists among logical systems. There are those, like PC, that are both decidable and completely axiomatizable; there are those, like LPC, that are not decidable but are completely axiomatizable; and there are those, like natural-number arithmetic, that are neither decidable nor completely axiomatizable. The ability to draw this distinction with clarity and to demonstrate to which of the three groups a given system belongs is one of the most noteworthy achievements of contemporary logic. (G.E.H.)

Metalogic

Metalogic is the study of the syntax and the semantics of formal languages and formal systems. It is related to, but does not include, the formal treatment of natural languages.

NATURE, ORIGINS, AND INFLUENCES OF METALOGIC

Specification of a formal language

Syntax and semantics. A formal language usually requires a set of formation rules; *i.e.*, a complete specification of the kinds of expressions that shall count as well-formed formulas (sentences or meaningful expressions), applicable mechanically, in the sense that a machine could check whether a candidate satisfies the requirements. This specification usually contains three parts: (1) a list of primitive symbols (basic units) given mechanically; (2) certain combinations of these symbols, singled out mechanically as forming the simple (atomic) sentences; and (3) a set of inductive clauses—inductive inasmuch as they stipulate that natural combinations of given sentences formed by such logical connectives as the disjunction "Either . . . or," which is symbolized "∨," "Not," symbolized "~," and "For every . . . ," symbolized "(∀ −)," are again sentences. [(∀ −), or (∃ −)—"There is some . . . "—is called a quantifier.] Since these specifications are concerned only with symbols and their combinations and not with meanings, they involve only the syntax of the language.

An interpretation of a formal language is determined by formulating an interpretation of the atomic sentences of the language with regard to a domain of objects; *i.e.*, by stipulating which objects of the domain are denoted by which constants of the language and which relations and functions are denoted by which predicate letters and function symbols. The truth-value (whether "true" or "false") of every sentence is thus determined according to the standard interpretation of logical connectives. For example, $p \cdot q$ is true if and only if p and q are true. (Here, the dot means the conjunction "and," not the multiplication operation "times.") Thus, given any interpretation of a formal language, a formal concept of truth is obtained. Truth, meaning, and denotation are semantic concepts.

If, in addition, a formal system in a formal language is introduced, certain syntactic concepts arise; viz., axioms, rules of inference, and theorems. Certain sentences are singled out as axioms. These are (the basic) theorems. Each rule of inference is an inductive clause, stating that if certain sentences are theorems, then another sentence related to them in a suitable way is also a theorem. If p and "Either not-p or q" ($\sim p \vee q$) are theorems, for example, then q is a theorem. In general, a theorem is either an axiom or the conclusion of a rule of inference whose premises are theorems.

In 1931 Gödel made the fundamental discovery that in

most of the interesting (or significant) formal systems not all true sentences are theorems. It follows from this finding that semantics cannot be reduced to syntax; thus syntax, which is closely related to proof theory, must often be distinguished from semantics, which is closely related to model theory. Roughly speaking, syntax—as conceived in the philosophy of mathematics—is a branch of number theory, and semantics is a branch of set theory, which deals with the nature and relations of aggregates (see NUMBER THEORY and SET THEORY).

Historically, as logic and axiomatic systems became more and more exact, there emerged, in response to a desire for greater lucidity, a tendency to pay greater attention to the syntactic features of the languages employed rather than to concentrate exclusively on intuitive meanings. In this way, logic, the axiomatic method (such as that employed in geometry), and semiotic (the general science of signs) converged toward metalogic.

The axiomatic method. The best known axiomatic system is that of Euclid for geometry. In a manner similar to that of Euclid, every scientific theory involves a body of meaningful concepts and a collection of true or believed assertions. The meaning of a concept can often be explained or defined in terms of other concepts; and, similarly, the truth of an assertion or the reason for believing it can usually be clarified by indicating that it can be deduced from certain other assertions already accepted. The axiomatic method proceeds in a sequence of steps, beginning with a set of primitive concepts and propositions and then defining or deducing all other concepts and propositions in the theory from them.

The realization that arose in the 19th century that there are different possible geometries led to a desire to separate abstract mathematics from spatial intuition; and, in consequence, many hidden axioms were uncovered in Euclid's geometry. These discoveries were organized into a more rigorous axiomatic system by David Hilbert in his *Grundlagen der Geometrie* (1899; *The Foundations of Geometry,* 1902). In this and related systems, however, logical connectives and their properties are taken for granted and remain implicit. If the logic involved is taken to be that of the predicate calculus, the logician can then arrive at such formal systems as that discussed above.

Transformations from semantics to syntax

Once such formal systems are obtained, it is possible to transform certain semantic problems into sharper syntactic problems. It has been asserted, for example, that non-Euclidean geometries must be self-consistent systems because they have models (or interpretations) in Euclidean geometry, which in turn has a model in the theory of real numbers. It may then be asked, however, how it is known that the theory of real numbers is consistent in the sense that no contradiction can be derived within it. Obviously, modelling can only establish a relative consistency and has to come to a stop somewhere. Having arrived at a formal system (say of real numbers), however, the consistency problem then has the sharper focus of a syntactic problem, viz., of considering all of the possible proofs (as syntactic objects) and asking whether any of them ever has (say) $0 = 1$ as the last sentence.

As another example, the question whether a system is categorical—that is, whether it determines essentially a unique interpretation in the sense that any two interpretations are isomorphic—may be explored. This semantic question can to some extent be replaced by a related syntactic question, that of completeness: whether there is in the system any sentence having a definite truth-value in the intended interpretation such that neither that sentence nor its negation is a theorem. Even though it is now known that the semantic and syntactic concepts are different, the vague requirement that a system be "adequate" is clarified by both concepts. The study of such sharp syntactic questions as those of consistency and completeness, which was emphasized by Hilbert, was named "metamathematics" (or "proof theory") by him around 1920.

Logic and metalogic. In one sense, logic is to be identified with the predicate calculus of the first order, the calculus in which the variables are confined to individuals of a fixed domain—though it may include as well the logic of equality symbolized "=," which takes the ordinary properties of equality as part of logic. In this sense, Gottlob Frege achieved a formal calculus of logic as early as 1879. Sometimes logic is construed, however, as including also higher order predicate calculi, which admit variables of higher types, such as those ranging over predicates (or classes and relations) and so on. But then it is a small step to the inclusion of set theory; and, in fact, axiomatic set theory is often regarded as a part of logic. For the purposes of this article, however, it is more appropriate to confine the discussion to logic in the first sense.

It is hard to separate out significant findings in logic from those in metalogic because all theorems of interest to logicians are about logic and therefore belong to metalogic. If p is a mathematical theorem—in particular, one about logic—and P is the conjunction of the mathematical axioms employed for proving p, then every p can be turned into a theorem, "Either not-P or p," in logic. Mathematics is not done, however, by carrying out explicitly all of the steps as formalized in logic; the selection and intuitive grasp of the axioms is important both for mathematics and for metamathematics. Actual derivations in logic, such as those carried out just prior to World War I by Alfred North Whitehead and Bertrand Russell are of little intrinsic interest to logicians. It might therefore appear redundant to introduce the term metalogic. In the present classification, however, metalogic is conceived as dealing not only with findings about logical calculi but also with studies of formal systems and formal languages in general.

Formal systems and logical calculi

An ordinary formal system differs from a logical calculus in that the system usually has an intended interpretation, whereas the logical calculus deliberately leaves the possible interpretations open. Thus, one speaks, for example, of the truth or falsity of sentences in a formal system; but with respect to a logical calculus one speaks of validity (*i.e.,* being true in all interpretations or in all possible worlds) and of satisfiability (or having a model; *i.e.,* being true in some particular interpretation). Hence, the completeness of a logical calculus has quite a different meaning from that of a formal system: a logical calculus permits many sentences, such that neither the sentence nor its negation is a theorem because it is true in some interpretations and false in others and requires only that every valid sentence be a theorem.

Semiotic. Originally, the word "semiotic" meant the medical theory of symptoms; however, an Empiricist, John Locke, used the term in the 17th century for a science of signs and significations. The current usage was recommended especially by Rudolf Carnap—see his *Introduction to Semantics* (1942) and his reference there to Charles William Morris, who suggested a threefold distinction. According to this usage, semiotic is the general science of signs and languages, consisting of three parts: (1) pragmatics (in which reference is made to the user of the language); (2) semantics (in which one abstracts from the user and analyzes only the expressions and their meanings); and (3) syntax (in which one abstracts also from the meanings and studies only the relations between expressions).

Pragmatics is at present a rather loose designation covering a wide range of scientific investigations of language—physiological, psychological, ethnological, and sociological. Since pragmatics is conspicuously incapable of formal treatment, it is usually not considered to be a part of metalogic.

At first Carnap exclusively emphasized syntax. But gradually he came to realize the importance of semantics, and the door was thus reopened to many difficult philosophical problems.

Logical Positivism and the formalization of science

Certain aspects of metalogic have been instrumental in the development of a whole new approach to philosophy commonly associated with the label of Logical Positivism. In his *Tractatus Logico-Philosophicus* (1922; originally published under a different title, 1921), the Austrian-born Cambridge philosopher Ludwig Wittgenstein, one of the most seminal thinkers in the philosophy of language, presented an appealing exposition of logical truths as sentences that are true in all possible worlds. One may say, for example, "It is raining or it is not raining," and in every possible world one of the disjuncts is true. On the

basis of this observation and certain broader developments in logic, Carnap attempted to develop formal treatments of science and philosophy.

It has been thought that the success that metalogic had achieved in the mathematical disciplines could be carried over into physics and even into biology or psychology. In so doing, the logician gives a branch of science a formal language in which there are logically true sentences having universal logical ranges and factually true sentences having universal logical ranges and factually true ones having more restricted ranges. (Roughly speaking, the logical range of a sentence is the set of all possible worlds in which it is true.)

A formal solution of the problem of meaning has also been proposed for these disciplines. Given the formal language of a science, it is possible to define a notion of truth. Such a truth definition determines the truth condition for every sentence; *i.e.,* the necessary and sufficient conditions for its truth. The meaning of a sentence is then identified with its truth condition because, as Carnap wrote:

> To understand a sentence, to know what is asserted by it, is the same as to know under what conditions it would be true .... To know the truth condition of a sentence is (in most cases) much less than to know its truth-value, but it is the necessary starting point for finding out its truth-value.

Influences in other directions. In recent years, metalogic has led to a great deal of work of a mathematical nature in axiomatic set theory, model theory, and recursion theory (in which functions that are computable in a finite number of steps are studied).

Turing machines, philosophy of logic, and ontology

In a different direction, the devising of Turing computing machines, involving abstract designs for the explication of mechanical logical procedures, has led to the investigation of idealized computers, with ramifications in the theory of finite automata and mathematical linguistics.

Among philosophers of language, there is a widespread tendency to stress the philosophy of logic (see PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE: *Philosophy of logic*). The contrast, for example, between intensional concepts and extensional concepts; the role of meaning in natural languages as providing truth conditions; the relation between formal and natural logic (*i.e.,* the logic of natural languages); and the relation of ontology, the study of the kinds of entities that exist, to the use of quantifiers—all of these areas are receiving extensive consideration and discussion. There are also efforts to produce formal systems for empirical sciences such as physics, biology, and even psychology. Many scholars have doubted, however, whether these efforts have been fruitful.

NATURE OF A FORMAL SYSTEM AND OF ITS FORMAL LANGUAGE

Example of a formal system. In order to clarify the abstract concepts of metalogic, a formal system N (with its formal language) may be considered for illustration.

Formation rules. The system may be set up by employing the following formation rules:

1. The following are primitive symbols: $\sim$, $\vee$, $\forall$, and $=$, and the symbols used for grouping, (, and); the function symbols for arithmetical addition and multiplication $+$ and $\cdot$; constants 0, 1; and variables $x,y,z, \ldots$.
2. The following are terms: a constant is a term; a variable is a term; if a and b are terms, $a + b$ and a $\cdot$ b are terms.
3. Atomic sentences are thus specified: if a and b are terms, $a = b$ is a sentence.
4. Other sentences can be defined as follows: if A and B are sentences and v is a variable, then $\sim A$, $A \vee B$, and $(\forall v)\, A$ are sentences.

Axioms and rules of inference. The system may be developed by adopting certain sentences as axioms and following certain rules of inference.

1. The basic axioms and rules are to be those of the first-order predicate calculus, including the propositional calculus (the logic of unanalyzed statements; see above *Formal logic*).
2. The following additional axioms of N are stipulated:

(a) Zero is not a successor (*i.e.,* does not follow any number immediately):

$$\sim(x + 1) = 0.$$

(b) No two numbers have the same successor:

$$\sim(x + 1 = y + 1) \vee x = y.$$

(c) Recursive definition of addition:

$$x + 0 = x$$
$$x + (y + 1) = (x + y) + 1.$$

(d) Recursive definition of multiplication:

$$x \cdot 0 = 0;\ x(y + 1) = (x \cdot y) + x.$$

3. Rule of inference (the principle of mathematical induction): Every number has a property if zero has it and the successor of a number has it if the number has it; more explicitly, if $A(0)$ and $\forall\ x[\sim A(x) \vee A(x + 1)]$ are theorems, then $\forall x\, A(x)$ is a theorem.

The system N as specified by the foregoing rules and axioms is a formal system in the sense that, given any combination of the primitive symbols, it is possible to check mechanically whether it is a sentence of N; and given a finite sequence of sentences, it is possible to check mechanically whether it is a (correct) proof in N; *i.e.,* whether each sentence either is an axiom or follows from preceding sentences in the sequence by a rule of inference. Viewed in this way, a sentence is a theorem if and only if there exists a proof in which it appears as the last sentence. It is not required of a formal system, however, that it be possible to decide mechanically whether a given sentence is a theorem; and, in fact, it has been proved that no such mechanical method exists.

Truth definition of the given language. The formal system N admits of different interpretations, according to findings of Gödel (from 1931) and of the Norwegian mathematician Thoralf Skolem, a pioneer in metalogic (from 1933). The originally intended, or standard, interpretation takes the ordinary non-negative integers $\{0,1,2, \ldots\}$ as the domain, the symbols 0 and 1 as denoting zero and one, and the symbols $+$ and $\cdot$ as standing for ordinary addition and multiplication. Relative to this interpretation, it is possible to give a truth definition of the language of N.

Open and closed sentences

It is necessary first to distinguish between open and closed sentences. An open sentence, such as $x = 1$, is one that may be either true or false depending on the value of x; but a closed sentence, such as $0 = 1$ and $(\forall x)\ (x = 0)$ or "All x's are zero," is one that has a definite truth-value—in this case, false (in the intended interpretation).

1. A closed atomic sentence is true if and only if it is true in the intuitive sense; for example, $0 = 0$ is true, $0 + 1 = 0$ is false.

This specification as it stands is not syntactic; but with some care, it is possible to give an explicit and mechanical specification of those closed atomic sentences that are true in the intuitive sense.

2. A closed sentence $\sim A$ is true if and only if A is not true.
3. A closed sentence $A \vee B$ is true if and only if either A or B is true.
4. A closed sentence $(\forall v)\, A(v)$ is true if and only if $A(v)$ is true for every value of v; *i.e.,* if $A(0)$, $A(1)$, $A(1 + 1)$, ... are all true.

The above definition of truth is not an explicit definition; it is an inductive one. Using concepts from set theory, however, it is possible to obtain an explicit definition that yields a set of sentences that consists of all of the true ones and only them. If Gödel's method of representing symbols and sentences by numbers is employed, it is then possible to obtain in set theory a set of natural numbers that are just the Gödel numbers of the true sentences of N.

There is a definite sense in which it is impossible to define the truth of a language in itself. This is proved by the liar paradox: if the sentence "I am lying," or alternatively

> This sentence is not true (1)

is considered, it is clear—since (1) is "This sentence"—that if (1) is true, then (1) is false; on the other hand, if (1) is false, then (1) is true. In the case of the system N, if the concept of truth were definable in the system itself, then (using a device invented by Gödel) it would be possible to obtain in N a sentence that amounts to (1) and that thereby yields a contradiction.

DISCOVERIES ABOUT FORMAL MATHEMATICAL SYSTEMS

The two central questions of metalogic are those of the completeness and consistency of a formal system based on axioms. In 1931, Gödel made fundamental discoveries in these areas for the most interesting formal systems.

In particular, he discovered that if such a system is ω-consistent—*i.e.,* devoid of contradiction in a sense to be explained below—then it is not complete, and that if a system is consistent, then the statement of its consistency, easily expressible in the system, is not provable in it.

Recursive functions and decidability

Soon afterward, in 1934, Gödel modified a suggestion that had been offered by Jacques Herbrand, a French mathematician, and introduced a general concept of recursive functions; *i.e.,* of functions mechanically computable by a finite series of purely combinatorial steps. In 1936, Alonzo Church, a mathematical logician; Alan Mathison Turing, originator of a theory of computability; and Emil L. Post, a specialist in recursive unsolvability, all argued for this concept (and certain equivalent notions), thereby arriving at stable and exact conceptions of "mechanical," "computable," "recursive," and "formal" that explicate the intuitive concept of what a mechanical computing procedure is. As a result of the development of recursion theory, it is now possible to prove not only that certain classes of problems are mechanically solvable (which could be done without the theory) but also that certain others are mechanically unsolvable (or absolutely unsolvable). The most notable example of such unsolvability is the discovery, made in 1970, that there is no algorithm, or rule of repetitive procedure, for solving all Diophantine equations; *i.e.,* equations of which the coefficients and roots are whole numbers. This solution gives a negative solution to the 10th problem in the famous list presented by Hilbert at the International Mathematical Congress in 1900.

In this way, logicians have finally arrived at a sharp concept of a formal axiomatic system because it is no longer necessary to leave "mechanical" as a vague nonmathematical concept. In this way, too, they have arrived at sharp concepts of decidability. In one sense, decidability is a property of sets (of sentences): that of being subject (or not) to mechanical methods by which to decide, for any closed sentence of a given formal system (*e.g.,* of N), whether it is true; or—as a different question—whether it is a theorem. In another sense, decidability can refer to a single closed sentence: the sentence is called undecidable in a formal system if neither it nor its negation is a theorem. Using this concept, Gödel's incompleteness theorem is sometimes stated thus: "Every interesting (or significant) formal system has some undecidable sentences."

Given these developments, it was easy to extend Gödel's findings, as Church did in 1936, to show that interesting formal systems such as N are undecidable (both with regard to theorems and with regard to true sentences).

The two incompleteness theorems. The first and most central finding in this field is that systems such as N are incomplete and incompletable because Gödel's theorem applies to any reasonable and moderately rich system. The proof of this incompleteness may be viewed as a modification of the liar paradox, which shows that truth cannot be defined in the language itself. Since provability in a formal system can often be expressed in the system itself, one is led to the conclusion of incompleteness.

Let us consider the sentence

This sentence is not provable in the system (2)

In particular, N may be thought of as the system being studied. Representing expressions by numbers and using an ingenious substitution function, Gödel was able to find in the system a sentence p that could be viewed as expressing (2).

Once such a sentence is obtained, some strong conclusions result. If the system is complete, then either the sentence p or its negation is a theorem of the system. If p is a theorem, then intuitively p or (2) is false, and there is in some sense a false theorem in the system. Similarly, if $\sim p$ is a theorem, then it says that $\sim$(2) or that p is provable in the system. Since $\sim p$ is a theorem, it should be true, and there seem then to be two conflicting sentences that are both true, viz., p is provable in the system, and $\sim p$ is provable in it. This can be the case only if the system is inconsistent.

Gödel's exact theorem

A careful examination of this inexact line of reasoning leads to Gödel's exact theorem, which says that if a system is reasonably rich and ω-consistent, then p is undecidable in it. The notion of ω-consistency is stronger than consistency, but it is a very reasonable requirement since it demands merely that one cannot prove in a system both that some number does not have the property A and yet that each number does have the property A; *i.e.,* that $(\exists x) \sim A(x)$ and also all of $A(0), A(1), \ldots$ are theorems. An American mathematician J. Barkley Rosser, who also contributed to number theory and applied mathematics, weakened the hypothesis to mere consistency in 1936, at the expense of complicating somewhat the initial sentence (2).

More exactly, Gödel showed that if the system is consistent, then p is not provable; if it is ω-consistent, then $\sim p$ is not provable. The first half leads to Gödel's theorem on consistency proofs, viz., that if a system is consistent, then the arithmetic sentence expressing the consistency of the system cannot be proved in the system. This is usually stated briefly thus: that no interesting system can prove its own consistency or that there exists no consistency proof of a system that can be formalized in the system itself.

The proof of this theorem consists essentially of a formalization in arithmetic of the arithmetized version of the proof of the statement, "If a system is consistent, then p is not provable"; *i.e.,* it consists of a derivation within number theory of p itself from the arithmetic sentence that says that the system is consistent. Hence, if the arithmetic sentence were provable, p would also be provable—contradicting the previous result. This proof, which was only briefly outlined by Gödel, has been carried out in detail by Paul Bernays in his joint work with Hilbert. Moreover, the undecidable sentence p is always of a relatively simple form, viz., of the form $(\forall x)\, A\,(x)$, "For every x, x is A," in which A is a recursive, in fact a primitive recursive, predicate.

Decidability and undecidability. The first incompleteness theorem yields directly the fact that truth in a system (*e.g.,* in N) to which the theorem applies is undecidable. If it were decidable, then all true sentences would form a recursive set, and they could be taken as the axioms of a formal system that would be complete. This claim depends on the reasonable and widely accepted assumption that all that is required of the axioms of a formal system is that they make it possible to decide effectively whether a given sentence is an axiom.

Undecidability of truth and theoremhood

Alternatively, the above assumption can be avoided by resorting to a familiar lemma, or auxiliary truth: that all recursive or computable functions and relations are representable in the system (*e.g.,* in N). Since truth in the language of a system is itself not representable (definable) in the system, it cannot, by the lemma, be recursive (*i.e.,* decidable).

The same lemma also yields the undecidability of such systems with regard to theorems. Thus, if there were a decision procedure, there would be a computable function f such that $f(i)$ equals 1 or 0 according as the ith sentence is a theorem or not. But then what $f(i) = 0$ says is just that the ith sentence is not provable. Hence, using Gödel's device, a sentence (say the tth) is again obtained saying of itself that it is not provable. If $f(t) = 0$ is true then, because f is representable in the system, it is a theorem of the system. But then, because $f(t) = 0$ is (equivalent to) the tth sentence, $f(t) = 1$ is also true and therefore provable in the system. Hence, the system, if consistent, is undecidable with regard to theorems.

Although the system N is incompletable and undecidable, it has been discovered by the Polish logician M. Presburger and by Skolem (both in 1930) that arithmetic with addition alone or multiplication alone is decidable (with regard to truth) and therefore has complete formal systems. Another well-known positive finding is that of the Polish-American semanticist and logician Alfred Tarski, who developed a decision procedure for elementary geometry and elementary algebra (1951).

A discovery made by Turing in 1936, which at first sight seems very surprising, may be mentioned in passing: that every complete formal system (though not every logical calculus) is decidable. If a system is complete, then for every closed sentence A, either that sentence or its negation is a theorem. Being a formal system, all of its theorems

can be enumerated mechanically. But then, for each A, one sooner or later reaches either A or $\sim A$. In the first case, A is a theorem (and true); in the second case, A is not a theorem (and false). This finding illustrates an impractical aspect of the concept of mechanical procedures, for all that is required is that a conclusion can be reached at some finite time and there need not be any useful estimate of how long it will take.

Consistency proofs. The best known consistency proof is that of the German mathematician Gerhard Gentzen (1936) for the system N of classical (or ordinary, in contrast to intuitionistic) number theory. Taking ω (omega) to represent the next number beyond the natural numbers (called the "first transfinite number"), Gentzen's proof employs an induction in the realm of transfinite numbers ($\omega + 1, \omega + 2, \ldots; 2\omega, 2\omega + 1, \ldots; \omega^2, \omega^2 + 1, \ldots$), which is extended to the first epsilon-number, ε_0 (defined as the limit of $\omega\ \omega^\omega, \omega\ \omega^{\omega^\omega}, \ldots$), which is not formalizable in N. This proof, which has appeared in several variants, has opened up an area of rather extensive work.

Intuitionistic number theory, which denies the classical concept of truth and consequently eschews certain general laws such as "Either A or $\sim A$," and its relation to classical number theory have also been investigated (see PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE: *Foundations of mathematics*). This investigation is considered significant because intuitionism is believed to be more constructive and more evident than classical number theory. In 1932, Gödel found an interpretation of classical number theory in the intuitionistic theory (also found by Gentzen and by Bernays). In 1958, Gödel extended his findings to obtain constructive interpretations of sentences of classical number theory in terms of primitive recursive functionals.

More recently, work has been done to extend Gentzen's findings to ramified theories of types and to fragments of classical analysis, and to extend Gödel's interpretation and to relate classical analysis to intuitionistic analysis. Also, in connection with these consistency proofs, various proposals have been made to present constructive notations for the ordinals of segments of the German mathematician Georg Cantor's second number class, which includes ω and the first epsilon-number and much more. A good deal of discussion has been devoted to the significance of the consistency proofs and the relative interpretations for epistemology (the theory of knowledge).

DISCOVERIES ABOUT LOGICAL CALCULI

The calculi of formal logic. The two main branches of formal logic are the propositional calculus and the predicate calculus (see above *Formal logic*).

The propositional calculus. It is easy to show that the propositional calculus is complete in the sense that every valid sentence in it—*i.e.*, every tautology, or sentence true in all possible worlds (in all interpretations)—is a theorem, as may be seen in the following example. "Either p or not-p" ($p \vee \sim p$) is always true because p is either true or false. In the former case, $p \vee \sim p$ is true because p is true; in the latter case, because $\sim p$ is true. One way to prove the completeness of this calculus is to observe that it is sufficient to reduce every sentence to a conjunctive normal form; *i.e.*, to a conjunction of disjunctions of single letters and their negations. But any such conjunction is valid if and only if every conjunct is valid; and a conjunct is valid if and only if it contains some letter p as well as $\sim p$ as parts of the whole disjunction. Completeness follows because (1) such conjuncts can all be proved in the calculus, and (2) if these conjuncts are theorems, then the whole conjunction is also a theorem.

Consistency and decidability of PC

The consistency of the propositional calculus (its freedom from contradiction) is more or less obvious because it can easily be checked that all of its axioms are valid—*i.e.*, true in all possible worlds—and that the rules of inference carry from valid sentences to valid sentences. But a contradiction is not valid; hence, the calculus is consistent. The conclusion, in fact, asserts more than consistency, for it holds that only valid sentences are provable.

The calculus is also easily decidable. Since all valid sentences, and only these, are theorems, each sentence can be tested mechanically by putting true and false for each letter in the sentence. If there are n letters, there are 2^n possible substitutions. A sentence is then a theorem if and only if it comes out true in every one of the 2^n possibilities.

The independence of the axioms is usually proved by using more than two truth-values. These values are divided into two classes: the desired and the undesired. The axiom to be shown independent can then acquire some undesired value, whereas all of the theorems that are provable without this axiom always get the desired values. This technique is what originally suggested the many-valued logics.

The first-order predicate calculus. The problem of consistency for the predicate calculus is relatively simple. A world may be assumed in which there is only one object a. In this case, both the universally quantified and the existentially quantified sentences $(\forall x)\,A(x)$ and $(\exists x)\,A(x)$ reduce to the simple sentence $A(a)$; and all quantifiers can be eliminated. It may easily be confirmed that, after the reduction, all theorems of the calculus become tautologies (*i.e.*, theorems in the propositional calculus). If F is any predicate, such a sentence as "Every x is F and not every x is F"—*i.e.*, $(\forall x)\,F(x) \cdot \sim(\forall x)\,F(x)$—is then reduced to "$a$ is both A and not-A"—$A(a) \cdot \sim A(a)$—which is not a tautology; therefore, the original sentence is not a theorem; and hence, no contradiction can be a theorem. If F is simple, then F and A are the same. If F is complex and contains $(\forall y)$ or $(\exists z)$, etc., then A is the result obtained by iterating the transformation of eliminating $(\forall y)$, etc. In fact, it can be proved quite directly not only that the calculus is consistent but also that all of its theorems are valid.

The discoveries that the calculus is complete and undecidable are much more profound than the discovery of its consistency. Its completeness was proved by Gödel in 1930; its undecidability was established with quite different methods by Church and Turing in 1936. Given the general developments that occurred up to 1936, its undecidability also follows in another way from Theorem X of Gödel's paper of 1931.

Completeness means that every valid sentence of the calculus is a theorem. It follows that if $\sim A$ is not a theorem, then $\sim A$ is not valid; and, therefore, A is satisfiable; *i.e.*, it has an interpretation or a model. But to say that A is consistent means nothing other than that $\sim A$ is not a theorem. Hence, from the completeness, it follows that if A is consistent, then A is satisfiable. Therefore, the semantic concepts of validity and satisfiability are seen to coincide with the syntactic concepts of derivability and consistency.

Löwenheim–Skolem theorem

A finding closely related to the completeness theorem is the Löwenheim–Skolem theorem (1915, 1920), named after Leopold Löwenheim, a German school teacher, and Skolem, which says that if a sentence (or a formal system) has any model, it has a countable or enumerable model; *i.e.*, a model whose members can be matched with the positive integers. In the most direct method of proving this theorem, the logician is provided with very useful tools in model theory and in studies on relative consistency and independence in set theory.

The Löwenheim–Skolem theorem. In the predicate calculus, there are certain reduction or normal-form theorems. One useful example is the prenex normal form: every sentence can be reduced to an equivalent sentence expressed in the prenex form, viz., in a form such that all of the quantifiers appear at the beginning. This form is specially useful for displaying the central ideas of some of the proofs of the Löwenheim–Skolem theorem.

For illustration one may consider a simple schema in prenex form, "For every x, there is some y such that x bears the (arbitrary) relation M to y"; *i.e.*,

$$(\forall x)\,(\exists y)Mxy \qquad (3)$$

If (3) now has a model with a nonempty domain D, then by a principle from set theory (the axiom of choice), there exists a function f of x, written $f(x)$, that singles out for each x a corresponding y. Hence, "For every x, x bears the relation M to $f(x)$"; *i.e.*,

$$(\forall x)Mxf(x). \qquad (4)$$

If a is now any object in D, then the countable sub-domain $\{a, f(a), f[f(a)], \ldots\}$ already contains enough objects to

satisfy (4) and therefore to satisfy (3). Hence, if (3) has any model, it has a countable model, which is in fact a submodel of the original.

An alternative proof, developed by Skolem in 1922 to avoid appealing to the principles of set theory, has turned out to be useful also for establishing the completeness of the calculus. Instead of using the function f as before, a can be arbitrarily denoted by *1*. Since equation (3) is true, there must be some object y such that the number *1* bears the relation M to y, or symbolically M_1y, and one of these y's may be called 2. When this process is repeated indefinitely, one obtains

Proof independent of set theory

$$M_{12}, M_{12} \cdot M_{23}, M_{12} \cdot M_{23} \cdot M_{34}, \ldots, \quad (5)$$

all of which are true in the given model. The argument is elementary because in each instance one merely argues from "There exists some y such that n is M of y"—*i.e.*, $(\exists y)Mny$—to "Let one such y be $n+1$." Consequently, every member of (5) is true in some model. It is then possible to infer that all members of (5) are simultaneously true in some model; *i.e.*, that there is some way of assigning truth-values to all of its atomic parts so that all members of (5) will be true. Hence, it follows that (3) is true in some countable model.

The completeness theorem. Gödel's original proof of the completeness theorem is closely related to the second proof above. Consideration may again be given to all of the sentences in (5) that contain no more quantifiers. If they are all satisfiable, then, as before, they are simultaneously satisfiable and (3) has a model. On the other hand, if (3) has no model, some of its terms—say $M_{12} \cdot \ldots \cdot M_{89}$—are not satisfiable; *i.e.*, their negations are tautologies (theorems of the propositional calculus). Thus, $\sim M_{12} \vee \ldots \vee \sim M_{89}$ is a tautology, and this remains true if *1,2, . . . ,9* are replaced by variables, such as *r,s, . . . ,z;* hence, $\sim Mrs \vee \ldots \vee \sim Myz$, being a tautology expressed in the predicate calculus as usually formulated, is a theorem in it. It is then easy to use the usual rules of the predicate calculus to derive also the statement, "There exists an x such that, for every y, x is not M of y"; *i.e.*, $(\exists x)\ (\forall y)\ \sim Mxy$. In other words, the negation of (3) is a theorem of the predicate calculus. Hence, if (3) has no model, then its negation is a theorem of the predicate calculus. And, finally, if a sentence is valid (*i.e.*, if its negation has no model), then it is itself a theorem of the predicate calculus.

The undecidability theorem and reduction classes. Given the completeness theorem, it follows that the task of deciding whether any sentence is a theorem of the predicate calculus is equivalent to that of deciding whether any sentence is valid or whether it is satisfiable.

Turing's undecidability proof

Turing's method of proving that this class of problems is undecidable is particularly suggestive. Once the concept of mechanical procedure was crystallized, it was relatively easy to find absolutely unsolvable problems; *e.g.*, the halting problem, which asks for each Turing machine the question of whether it will ever stop, beginning with a blank tape. In other words, each Turing machine operates in a predetermined manner according to what is given initially on the (input) tape; we consider now the special case of a blank tape and ask the special question whether the machine will eventually stop. This infinite class of questions (one for each machine) is known to be unsolvable.

Turing's method shows that each such question about a single Turing machine can be expressed by a single sentence of the predicate calculus so that the machine will stop if and only if that sentence is not satisfiable. Hence, if there were a decision procedure of validity (or satisfiability) for all sentences of the predicate calculus, then the halting problem would be solvable.

In more recent years (1962), Turing's formulation has been improved to the extent that all that is needed are sentences of the relatively simple form $(\forall x)\ (\exists y)\ (\forall z)\ Mxyz$, in which all of the quantifiers are at the beginning; *i.e.*, M contains no more quantifiers. Hence, given the unsolvability of the halting problem, it follows that, even for the simple class of sentences in the predicate calculus having the quantifiers $\forall\ \exists\ \forall$, the decision problem is unsolvable. Moreover, the method of proof also yields a procedure by which every sentence of the predicate calculus can be correlated with one in the simple form given above. Thus, the class of $\forall\ \exists\ \forall$ sentences forms a "reduction class." (There are also various other reduction classes.)

MODEL THEORY

Background and typical problems. In model theory one studies the interpretations (models) of theories formalized in the framework of formal logic, especially in that of the first-order predicate calculus with equality; *i.e.*, in elementary logic. A first-order language is given by a collection S of symbols for relations, functions, and constants, which, in combination with the symbols of elementary logic, single out certain combinations of symbols as sentences. Thus, for example, in the case of the system N (see above *Example of a formal system*), the formation rules yield a language that is determined in accordance with a uniform procedure by the set (indicated by braces) of uninterpreted extralogical symbols:

$$S = \{+,\ \cdot,\ 0,\ 1\}.$$

A first-order theory is determined by a language and a set of selected sentences of the language—those sentences of the theory that are, in an arbitrary, generalized sense, the "true" ones (called the "distinguished elements" of the set). In the particular case of the system N, one theory T_a is built up on the basis of the language and the set of theorems of N, and another theory T_b is determined by the true sentences of N according to the natural interpretation or meaning of its language. In general, the language of N and any set of sentences of the language can be used to make up a theory.

Satisfaction of a theory by a structure: finite and infinite models. A realization of a language (for example, the one based on S) is a structure $\mathfrak{A}$ identified by the five elements so arranged

$$\mathfrak{A} = \langle A, +_{\mathfrak{A}}, \cdot_{\mathfrak{A}}, 0_{\mathfrak{A}}, 1_{\mathfrak{A}} \rangle,$$

in which A is a nonempty set (called the domain of $\mathfrak{A}$), the last two terms are members of A, and the other two terms are functions correlating each member of the Cartesian product $A \times A$ (*i.e.*, from the set of ordered pairs $\langle a,b \rangle$ such that a,b belong to A) with a member of A. The structure $\mathfrak{A}$ satisfies or is a model of the theory T_a (or T_b) if all of the distinguished sentences of T_a (or T_b) are true in $\mathfrak{A}$ (or satisfied by $\mathfrak{A}$). Thus, if $\mathfrak{A}$ is the structure of the ordinary non-negative integers $\langle \omega, +, ., 0, 1 \rangle$, in which ω is the set of all such integers and $+, ., 0$, and 1 the elements for their generation, then it is not only a realization of the language based on S but also a model of both T_a and T_b. Gödel's incompleteness theorem permits nonstandard models of T_a that contain more objects than ω but in which all of the distinguished sentences of T_a (viz., the theorems of the system N) are true. Skolem's constructions (related to ultraproducts, see below) yield nonstandard models for both theory T_a and theory T_b.

The use of the relation of satisfaction, or being-a-model-of, between a structure and a theory (or a sentence) can be traced to the book *Wissenschaftslehre* (1837; "Theory of Knowledge") by Bernard Bolzano, a Czech theologian and mathematician, and, in a more concrete context, to the introduction of models of non-Euclidean geometries around that time. In the mathematical treatment of logic, these concepts can be found in works of the late 19th-century German mathematician Ernst Schröder and in Löwenheim (in particular, in his paper of 1915). The basic tools and results achieved in model theory—such as the Löwenheim–Skolem theorem, the completeness theorem of elementary logic, and Skolem's construction of nonstandard models of arithmetic—were developed during the period from 1915 to 1933. A more general and abstract study of model theory began after 1950, in the work of Tarski and others.

One group of developments may be classified as refinements and extensions of the Löwenheim–Skolem theorem. These developments employ the concept of a "cardinal number," which—for a finite set—is simply the number at which one stops in counting its elements. For infinite sets, however, the elements must be matched from set

to set instead of being counted, and the "sizes" of these sets must thus be designated by transfinite numbers. A rather direct generalization can be drawn that says that, if a theory has any infinite model, then, for any infinite cardinal number, it has a model of that cardinality. It follows that no theory with any infinite model can be categorical or such that any two models of the theory are isomorphic (*i.e.,* matchable in one-to-one correspondence) because models of different cardinalities can obviously not be so matched. A natural question is whether a theory can be categorical in certain infinite cardinalities; *i.e.,* whether there are cardinal numbers such that any two models of the theory of the same cardinality are isomorphic. According to a central discovery made in 1963 by Michael Morley, a U.S. mathematician, if a theory is categorical in any uncountable cardinality (*i.e.,* any cardinality higher than the countable), then it is categorical in every uncountable cardinality. On the other hand, examples are known for all four combinations of countable and uncountable cardinalities: specifically, there are theories that are categorical (1) in every infinite cardinality; (2) in the countable cardinality but in no uncountable cardinality; (3) in every uncountable cardinality but not in the countable; and (4) in no infinite cardinality.

Categorical theories in various cardinalities

In another direction, there are "two-cardinal" problems that arise from the possibilities of changing, from one model to another, not only the cardinality of the domain of the first model but also the cardinality of some chosen property (such as being a prime number). Various answers to these questions have been found, including proofs of independence (based on the ordinary axioms employed in set theory) and proofs of conditional theorems made on the basis of certain familiar hypotheses of set theory.

Elementary logic. An area that is perhaps of more philosophical interest is that of the nature of elementary logic itself. On the one hand, the completeness discoveries seem to show in some sense that elementary logic is what the logician naturally wishes to have. On the other hand, he is still inclined to ask whether there might be some principle of uniqueness according to which elementary logic is the only solution that satisfies certain natural requirements on what a logic should be. The recent development of model theory has led to a more general outlook that enabled Per Lindström, a Swedish logician, to prove, in 1969, a general theorem to the effect that, roughly speaking, within a broad class of possible logics, elementary logic is the only one that satisfies the requirements of axiomatizability and of the Löwenheim–Skolem theorem. Although Lindström's theorem does not settle satisfactorily whether or not elementary logic is the right logic, it does seem to suggest that mathematical findings can help the logician to clarify his concepts of logic and of logical truth.

Compounded models

A particularly useful tool for obtaining new models from the given models of a theory is the construction of a special combination called the "ultraproduct" of a family of structures (see below *Ultrafilters, ultraproducts, and ultrapowers*)—in particular, the ultrapower when the structures are all copies of the same structure (just as the product of $a_1, \ldots, a_n$ is the same as the power a^n, if $a_i = a$ for each i). The intuitive idea in this method is to establish that a sentence is true in the ultraproduct if and only if it is true in "almost all" of the given structures (*i.e.,* "almost everywhere"—an idea that was present in a different form in Skolem's construction of a nonstandard model of arithmetic in 1933). It follows that if the given structures are models of a theory, then their ultraproduct is such a model also, because every sentence in the theory is true everywhere (which is a special case of "almost everywhere" in the technical sense employed). Ultraproducts have been applied, for example, to provide a foundation for what is known as "nonstandard analysis" that yields an unambiguous interpretation of the classical concept of infinitesimals—the division into units as small as one pleases. They have also been applied by two mathematicians, James Ax and S.B. Kochen, to problems in the field of algebra (on p-adic fields).

Nonelementary logic and future developments. There are also studies that develop the model theory of nonelementary logic: such as second-order logic and infinitary logics. Second-order logic contains, in addition to variables that range over individual objects, a second kind of variable ranging over sets of objects so that the model of a second-order sentence or theory also involves, beyond the basic domain, a larger set (called its "power set") that encompasses all of the subsets of the domain. Infinitary logics may include functions or relations with infinitely many arguments, infinitely long conjunctions and disjunctions, or infinite strings of quantifiers. From studies on infinitary logics, William Hanf, an American logician, was able to define certain cardinals, some of which have been studied in connection with the large cardinals in set theory. In yet another direction, logicians are developing model theories for modal logics—those dealing with such modalities as necessity and possibility—and for the intuitionistic logic.

There is a large gap between the general theory of models and the construction of interesting particular models such as those employed in the proofs of the independence (and consistency) of special axioms and hypotheses in set theory. It is natural to look for further developments of model theory that will yield more systematic methods for constructing models of axioms with interesting particular properties, especially in deciding whether certain given sentences are derivable from the axioms. Relative to the present state of knowledge, such goals appear fairly remote. The gap is not unlike that between the abstract theory of computers and the basic properties of actual computers.

Characterizations of the first-order logic. There has been outlined above a proof of the completeness of elementary logic without including sentences asserting equality or identity. The proof can be extended, however, to the full elementary logic in a fairly direct manner. Thus, if F is a sentence containing equality, a sentence G can be adjoined to it that embodies the special properties of equality relevant to the sentence F. The conjunction of F and G can then be treated as a sentence not containing equality (*i.e.,* $=$ can be treated as an arbitrary relation symbol). Hence, the conjunction has a model in the sense of logic-without-equality if and only if F has a model in the sense of logic-with-equality; and the completeness of elementary logic (with equality) can thus be inferred.

Extension to include equality

A concept more general than validity is that of the relation of logical entailment or implication between a possibly infinite set X of sentences and a single sentence p that holds if and only if p is true in every model of X. In particular, p is valid if the empty set, defined as having no members, logically entails p—for this is just another way of saying that p is true in every model. This suggests a stronger requirement on a formal system of logic, viz., that p be derivable from X by the system whenever X logically entails p. The usual systems of logic satisfy this requirement because, besides the completeness theorem, there is also a compactness theorem:

> A theory X has a model if every finite subset of X has a model.

Roughly speaking, this theorem enables the logician to reduce an infinite set X to a finite subset X_1 in each individual case, and the case of entailment when X_1 is finite is taken care of by the completeness of the system.

These findings show that the ordinary systems of elementary logic comprise the correct formulation, provided that the actual choice of the truth functions (say negation and disjunction), of the quantifiers, and of equality as the "logical constants" is assumed to be the correct one. There remains the question, however, of justifying the particular choice of logical constants. One might ask, for example, whether "For most x" or "For finitely many x" should not be counted as logical constants. Lindström has formulated a general concept of logic and shown that logics that apparently extend the first-order logic all end up being the same as that logic, provided that they satisfy the Löwenheim–Skolem theorem and either have the compactness property or are formally axiomatizable. There remains the question, however, of whether or why these requirements (especially that of the Löwenheim–Skolem theorem) are intrinsic to the nature of logic.

Generalizations and extensions of the Löwenheim–Skolem theorem. A generalized theorem can be proved using

basically the same ideas as those employed in the more special case discussed above.

If a theory has any infinite model, then, for any infinite cardinality α, that theory has a model of cardinality α. More explicitly, this theorem contains two parts: (1) If a theory has a model of infinite cardinality β, then, for each infinite cardinal α that is greater than β, the theory has a model of cardinality α. (2) If a theory has a model of infinite cardinality β, then, for each infinite cardinal α less than β, the theory has a model of cardinality α.

It follows immediately that any theory having an infinite model has two nonisomorphic models and is, therefore, not categorical. This applies, in particular, to the aforementioned theories T_a and T_b of arithmetic (based on the language of N), the natural models of which are countable, as well as to theories dealing with real numbers and arbitrary sets, the natural models of which are uncountable; both kinds of theory have both countable and uncountable models. There is much philosophical discussion about this phenomenon.

The possibility is not excluded that a theory may be categorical in some infinite cardinality. The theory T_d, for example, of dense linear ordering (such as that of the rational numbers) is categorical in the countable cardinality. One application of the Löwenheim–Skolem theorem is: If a theory has no finite models and is categorical in some infinite cardinality α, then the theory is complete; *i.e.,* for every closed sentence in the language of the theory, either that sentence or its negation belongs to the theory. An immediate consequence of this application of the theorem is that the theory of dense linear ordering is complete.

Morley's theorem

A theorem that is generally regarded as one of the most difficult to prove in model theory is the theorem by Morley, as follows:

> A theory that is categorical in one uncountable cardinality is categorical in every uncountable cardinality.

The two-cardinal theorems deal with languages that have some distinguished predicate U. A theory is said to admit the pair $<\alpha, \beta>$ of cardinals if the theory has a model (with its domain) of cardinality α in which the value of U is a set of cardinality β. The central two-cardinal theorem says:

> If a theory admits the pair $<\alpha, \beta>$ of infinite cardinals with β less than α, then for each regular cardinal γ, the theory admits $<\gamma^+, \gamma>$, in which γ^+ is the next larger cardinal after γ.

The most interesting case is when γ is the least infinite cardinal, $\aleph_0$. (The general theorem can be established only when the "generalized continuum hypothesis" is assumed, according to which the next highest cardinality for an infinite set is that of its power set.)

Ultrafilters, ultraproducts, and ultrapowers. An ultrafilter on a nonempty set I is defined as a set D of subsets of I such that:

1. the empty set does not belong to D;
2. if A, B are in D, so is their intersection, $A \cap B$, the set of elements common to both;
3. if A is a subset of B, and A is in D, then B is in D;
4. for every subset A of I, either A is in D or I minus A is in D.

Roughly stated, each ultrafilter of a set I conveys a notion of large subsets of I so that any property applying to a member of D applies to I "almost everywhere."

The set $\{\mathfrak{A}_i\}$ where $\mathfrak{A}_i = <A_i, R_i>$ and the i are members of the set I, is taken to be a family of structures indexed by I, and D to be an ultrafilter on I. Consider now the Cartesian product B of $\{A_i\}$ (for example, if I is $\{0,1,2,\ldots\}$, then B is the set of all sequences f such that $f(i)$ belong to A_i). The members of B are divided into equivalence classes with the help of $D: f \equiv g$ if and only if $\{i \mid f(i) = g(i)\} \in D$; in other words, the set of indices i such that $f(i) = g(i)$ being to D [or $f(i)$ and $g(i)$ are equal "almost everywhere"]. Let W be the set of these equivalence classes, *i.e.,* the set of all f^* such that f^* is the set of all members g of B with $g \equiv f$. We introduce similarly a relation S such that Sfg if and only if R_i holds between $f(i)$ and $g(i)$ for "almost all" i, *i.e.,*

$$\{i \mid R_i [f(i), g(i)]\} \in D$$

In this way, we arrive at a new structure $U = <W, S>$ which is called the ultraproduct of the original family $\{\mathfrak{A}_i\}$ over D. In the special case when all the $\mathfrak{A}_i$ are the same, the resulting structure U is called the ultrapower of the original family over D.

The central theorems are the following:

1. If $\mathfrak{A}_i$ $(i \in I)$ are realizations of the same language, then a sentence p is true in the ultraproduct U if and only if the set of i such that p is true in $\mathfrak{A}_i$ belongs to D. In particular, if each $\mathfrak{A}_i$ is a model of a theory, then U is also a model of the theory.
2. Two realizations of the same language are said to be elementarily equivalent if they have the same set of true sentences. A necessary and sufficient condition for two realizations to be elementarily equivalent is that they admit ultrapowers that are isomorphic.

One application of these theorems is in the introduction of nonstandard analysis, which was originally instituted by other considerations. By using a suitable ultrapower of the structure of the field of $\mathfrak{R}$ real numbers, a real closed field that is elementarily equivalent to $\mathfrak{R}$ is obtained that is non-Archimedean; *i.e.,* which permits numbers a and b such that no n can make na greater than b. This development supplies an unexpected exact foundation for the classical differential calculus using infinitesimals, which has considerable historical, pedagogical, and philosophical interest.

Applications to nonstandard analysis and algebra

A widely known application to the area of algebra is that which deals with certain fields of rational numbers Q_p, called the p-adic completion of the rational numbers. The conjecture has been made that every form of degree d (in the same sense as degrees of ordinary polynomials) over Q_p, in which the number of variables exceeds d^2, has a nontrivial zero in Q_p. Using ultraproducts, it has been shown that the conjecture is true for arbitrary d with the possible exception of a finite set of primes p (depending on d). Subsequently, it was found that the original conjecture is not true when extended to full generality.

Other useful tools in model theory include the pigeonhole principles, of which the basic principle is that, if a set of large cardinality is partitioned into a small number of classes, some one class will have large cardinality. Those elements of the set that lie in the same class cannot be distinguished by the property defining that class.

A related concept is that of "indiscernibles," which also has rather extensive applications in set theory. An ordered subset of the domain of a model $\mathfrak{A}$ of a theory is a homogeneous set, or a set of indiscernibles for $\mathfrak{A}$, if $\mathfrak{A}$ cannot distinguish the members of the subset from one another. More exactly, given any $x_1 <\ldots< x_n$, $y_1 <\ldots< y_n$ in the subset, then for any sentence $F(a_1, \ldots, a_n)$ of the language of the theory, that sentence (with argument x) is satisfied by (symbolized $\models$) the structure; *i.e.,*

$$\mathfrak{A} \models F(x_1, \ldots, x_n),$$

if and only if that sentence (with argument y) is also satisfied by it; *i.e.,*

$$\mathfrak{A} \models F(y_1, \ldots, y_n).$$

There is also a first theorem on this notion that says that, given a theory with an infinite model and a linearly ordered set X, there is then a model $\mathfrak{A}$ of the theory such that X is a set of indiscernibles for $\mathfrak{A}$.

(H.Wa.)

Applied logic

As has been shown, pure logic develops a body of strictly abstract machinery—*e.g.,* the syllogism—that displays only the skeletal form of the argument without regard to its content. In applied logic one adapts this machinery for deployment over concrete issues of a certain range of special subject matter—such as beliefs, commands, or duties. Such arguments often are not governed only by strictly formal issues (which alone are relevant in pure logic); they involve also certain subject-matter considerations of a particular area, generally called the material aspects. Applied logic, to be sure, has its formal aspects, as well, and so cannot be equated with material logic as such; nor can it be equated with practical logic (analysis of statements of the form "I choose . . . ."; "I prefer . . . ."; "I do . . . ."), which is only one of its branches.

The applications envisaged in applied logic cover a vast

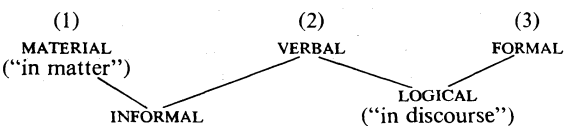
Range of applications

range, relating to reasoning in the sciences and in philosophy, as well as in everyday discourse. They include: (1) the various sorts of reasoning affecting the conduct of ordinary discourse as well as the theory of the logical relations that exist within special realms of discourse—between two commands, for example, or between one question and another; (2) special forms of logic designed for scientific applications, such as temporal logic (of what "was" or "will be" the case), or mereology (the logic of parts and wholes); and (3) special forms for concepts bearing upon philosophical issues, such as logics that deal with statements of the form "I know that . . . ."; "I believe that . . . ."; "It is permitted to . . . ."; "It is obligatory to . . . ."; or "It is prohibited to . . . ."

THE CRITIQUE OF FORMS OF REASONING

Correct and defective argument forms. An argument is a body of discourse that presents evidence—or purported evidence—in support of some thesis. Correspondingly, an argument has two components: a conclusion, the thesis argued for; and certain premises, the considerations adduced on behalf of the conclusion. An argument is deductively valid when its premises provide conclusive evidence for the conclusion—*i.e.,* if the conclusion must be true whenever the premises are true. An argument that fails to be conclusively deduced is invalid; it is also said to be fallacious (unless reclassified as a nondeductive argument and reinstated by appeal to material or substantive considerations beyond its merely formal aspects).

An argument may be fallacious in three ways: in its material content, through a misstatement of the facts; in its wording, through an incorrect use of terms; or in its structure (or form), through the use of an improper process of inference. As shown in the diagram below,



fallacies are correspondingly classified as (1) material, (2) verbal, and (3) formal. Groups (2) and (3) are called logical fallacies, or fallacies "in discourse," in contrast to the substantive, or material, fallacies of group (1), called fallacies "in matter"; and groups (1) and (2) are called informal fallacies, in contrast to group (3).

Kinds of fallacies. *Material fallacies.* The material fallacies are also known as fallacies of presumption, because the premises "presume" too much—they either covertly assume the conclusion or avoid the issue in view.

Accident, converse accident, irrelevant conclusion

The classification that is still widely used is that of Aristotle's *Sophistical Refutations:* (1) The fallacy of accident is committed by an argument that applies a general rule to a particular case in which some special circumstance ("accident") makes the rule inapplicable. The truth that "men are capable of seeing" is no basis for the conclusion that "blind men are capable of seeing." This is a special case of the fallacy of *secundum quid* (more fully: *a dicto simpliciter ad dictum secundum quid,* which means "from a saying [taken too] simply to a saying according to what [it really is]"; *i.e.,* according to its truth as holding only under special provisos). This fallacy is committed when a general proposition is used as the premise for an argument without attention to the (tacit) restrictions and qualifications that govern it and invalidate its application in the manner at issue. (2) The converse fallacy of accident argues improperly from a special case to a general rule. Thus, the fact that a certain drug is beneficial to some sick persons does not imply that it is beneficial to all men. (3) The fallacy of irrelevant conclusion is committed when the conclusion changes the point that is at issue in the premises. Special cases of irrelevant conclusion are presented by the so-called fallacies of relevance. These include: (*a*) the argument *ad hominem* (speaking "against the man" rather than to the issue), in which the premises may only make a personal attack on a person who holds some thesis, instead of offering grounds showing why what he says is false; (*b*) the argument *ad populum* (an appeal "to the people"), which, instead of offering logical reasons, appeals to such popular attitudes as the dislike of injustice; (*c*) the argument *ad misericordiam* (an appeal "to pity"), as when a trial lawyer, rather than arguing for his client's innocence, tries to move the jury to sympathy for him; (*d*) the argument *ad verecundiam* (an appeal "to awe"), which seeks to secure acceptance of the conclusion on the grounds of its endorsement by persons whose views are held in general respect; (*e*) the argument *ad ignorantiam* (an appeal "to ignorance"), which argues that something (*e.g.,* extrasensory perception) is so since no one has shown that it is not so; and (*f*) the argument *ad baculum* (an appeal "to force"), which rests on a threatened or implied use of force to induce acceptance of its conclusion. (4) The fallacy of circular argument, known as *petitio principii* ("begging the question"), occurs when the premises presume, openly or covertly, the very conclusion that is to be demonstrated (example: "Gregory always votes wisely." "But how do you know?" "Because he always votes Libertarian."). A special form of this fallacy, called a vicious circle, or *circulus in probando* ("arguing in a circle"), occurs in a course of reasoning typified by the complex argument in which a premise p_1 is used to prove p_2; p_2 is used to prove p_3; and so on, until p_{n-1} is used to prove p_n; then p_n is subsequently used in a proof of p_1, and the whole series $p_1, p_2, \ldots, p_n$ is taken as established. (Example: "McKinley College's baseball team is the best in the Association [$p_n = p_3$]; they are the best because of their strong batting potential [p_2]; they have this potential because of the ability of Jones, Crawford, and Randolph at the bat [p_1]." "But how do you know that Jones, Crawford, and Randolph are such good batters?" "Well, after all, these men are the backbone of the best team in the Association [p_3 again].") Strictly speaking, *petitio principii* is not a fallacy of reasoning but an ineptitude in argumentation: thus the argument from p as a premise to p as conclusion is not deductively invalid but lacks any power of conviction, since no one who questioned the conclusion could concede the premise. (5) The fallacy of false cause (*non causa pro causa*) mislocates the cause of one phenomenon in another that is only seemingly related. The most common version of this fallacy, called *post hoc ergo propter hoc* ("after which hence by which"), mistakes temporal sequence for causal connection—as when a misfortune is attributed to a "malign event," like the dropping of a mirror. Another, more sophisticated, and purely logical version of this fallacy occurs, however, in a complex form of the *reductio ad absurdum*—an argument that, in its simple form, declares that an assumption is false if a contradiction can be drawn from it. In the present case, a contradiction drawn jointly from several premises—say, p_1, p_2, p_3, and p_4—is taken to establish the falsity of one of them—say, p_2—when another—say, p_3—is really the source of the difficulty. Thus, the premises p_1 through p_4 may be those, for example, of the following two syllogisms:

Circular argument

Other material fallacies

p_1: "All statesmen are political figures who act in the public interest."
p_2: "Jones is a statesman."
C_1: Therefore, "Jones is a political figure who is acting in the public interest."

and

p_3: "Anyone who shamelessly campaigns for reelection is not acting in the public interest."
p_4: "Jones is shamelessly campaigning for reelection."
C_2: Therefore, "Jones is not acting in the public interest."

But conclusions C_1 and C_2 are contradictory. A critic might therefore argue that p_2 is false and that Jones is not a statesman. But he could be wrong, for the trouble could lie in p_3: far from acting at variance with the public interest, a man might campaign in the belief that he cannot accomplish his legislative program, which he believes to be in the public interest, unless he wins another term in office. His campaigning may, thus, be in the public interest; and he may be a statesman after all. (6) The fallacy of many questions (*plurimum interrogationum*) consists in demanding or giving a single answer to a question when this answer could either be divided (example: "Do you like the twins?" "Neither yes nor no; but Ann yes and Mary no.") or refused altogether, because a mistaken presupposition is involved

(example: "Have you stopped beating your wife?"). (7) The fallacy of *non sequitur* ("it does not follow"), still more drastic than the preceding, occurs when there is not even a deceptively plausible appearance of valid reasoning, because there is a virtually complete lack of connection between the given premises and the conclusion drawn from them. Some authors, however, identify *non sequitur* with the fallacy of the consequent (see below *Formal fallacies*).

Verbal fallacies. These fallacies, called fallacies of ambiguity, arise when the conclusion is achieved through an improper use of words.

Equivocation, amphiboly, and other verbal fallacies

The principal instances are as follows: (1) Equivocation occurs when a word or phrase is used in one sense in one premise and another sense in some other needed premise or in the conclusion (example: "The loss made Jones mad [= angry]; mad [= insane] people should be institutionalized; so Jones should be institutionalized."). The figure-of-speech fallacy is the special case arising from confusion between the ordinary sense of a word and its metaphorical, figurative, or technical employment (example: "For the past week Joan has been living on the heights of ecstasy." "And what is her address there?"). (2) Amphiboly occurs when the grammar of a statement is such that several distinct meanings can obtain (example: "The king hopes you the enemy will slay. So he wishes you well."). (3) Accent is a counterpart of amphiboly arising when a statement can bear distinct meanings depending upon which word is stressed (example: "Men are *considered* equal." "Men are considered *equal*."). (4) Composition occurs when the premise that the parts of a whole are of a certain nature is improperly used to infer that the whole itself must also be of this nature (example: a story made up of good paragraphs is thus said to be a good story). (5) Division—the reverse of composition—occurs when the premise that a collective whole has a certain nature is improperly used to infer that a part of this whole must also be of this nature (example: in a speech that is long winded it is presumed that every sentence is long). But this fallacy and its predecessor can be viewed as versions of *equivocation,* in which the distributive use of a term—*i.e.,* its application to the elements of an aggregate (example: "the crowd," viewed as individuals)—is equivocally confused with its collective use ("the crowd," as a unitary whole)—compare "The crowd were filing through the turnstile" with "The crowd was compressed into the space of a city block."

Formal fallacies. Strictly logical, or formal, fallacies arise not from the specific matter of the argument but from a structural pattern of reasoning that is generically incorrect.

A classic case is Aristotle's fallacy of the consequent, relating to reasoning from premises of the form: "If p_1, then p_2." The fallacy has two forms: (1) denial of the antecedent, in which one mistakenly argues from the premises "If p_1, then p_2" and "not-p_1" (symbolized $\sim p_1$), to the conclusion "not-p_2" (example: "If George is a man of good faith, he can be entrusted with this office; but George is not a man of good faith; therefore, George cannot be entrusted with this office"); and (2) affirmation of the consequent, in which one mistakenly argues from the premises "If p_1, then p_2" and "p_2," to the conclusion "p_1" (example: "If Amos was a prophet, then he had a social conscience; he had a social conscience; hence, Amos was a prophet"). Most of the traditionally considered formal fallacies, however, relate to the syllogism (see above *Syllogistic*). One example may be cited, that of the fallacy of illicit major (or minor) premise, which violates the rules for "distribution." (A term is said to be distributed when reference is made to all members of the class. Example: in "Some crows are not friendly," reference is made to all friendly things, but not to all crows.) The fallacy arises when a major (or minor) term that is undistributed in the premise is distributed in the conclusion (example: "All tubers are high-starch foods [undistributed]; no squashes are tubers; therefore, no squashes are high-starch foods [distributed]").

EPISTEMIC LOGIC

Epistemic logic deals with the logical issues arising within the gamut of such epistemological concepts as knowledge, belief, assertion, doubt, question-and-answer, or the like. Instead of dealing with the essentially factual issues of alethic logic (Greek *alētheia,* "truth")—*i.e.,* with what is actually or must necessarily or can possibly be the case—it relates to what people know or believe or maintain or doubt to be the case.

The logic of belief. From the logical standpoint, a belief is generally analyzed as a relationship obtaining between the person who accepts some thesis on the one hand, and the thesis that he accepts, on the other. Correspondingly, given a person x, it is convenient to consider the set $\mathbf{B}_x$ of x's beliefs and represent the statement "x believes that p" as $p \in \mathbf{B}_x$. (The symbol $\in$ represents membership in a set, $\notin$ its denial.)

Alternative ways of construing belief. Before the formulation of any logical theory of belief can get under way, however, it is essential to specify the precise sense at issue when a person is said to believe something. It is necessary to choose between three alternative constructions of "x believes that p":

1. The disposition-to-explicit-assent construction, according to which "x believes that p" obtains if he overtly professes to do so (and is candid) or would overtly so profess (if candid) once the matter is duly explained to him. That is, x is simply asked if he believes p, and the matter is settled by his response, if there is no reason to suspect disingenuousness.
2. The commitment-to-assent construction, according to which "x believes that p" if he either (a) explicitly assents to p in the preceding manner or (b) is logically committed to giving assent, because p follows logically from propositions to which he does give assent.
3. The behaviouristic construction, according to which "x believes that p" if x acts as though p were the case.

Shortcomings of the three constructions

The behaviouristic construction has serious shortcomings insofar as it strays from the ordinary, presystematic concept at issue in ordinary discourse about beliefs. Fiji Islanders may well act as though Julius Caesar had never existed but cannot be said to have beliefs in the matter. The other two approaches also have shortcomings. The commitment-to-assent construction gives rise, for example, to the entailment principle (EP): "If p is one of the beliefs held by x, and if q follows deductively from p, then q is one of the beliefs held by x"; or, symbolically:

$$\text{(EP)} \qquad \text{If } p \in \mathbf{B}_x \text{ and } p \vdash q, \text{ then } q \in \mathbf{B}_x,$$

in which the symbol $\vdash$ signifies that the proposition on the right side follows as a deductively valid consequence from those on the left. Example: "That there are more people in Kōbe than the number of hairs that any human head can hold (p)" is a member of ($\in$) the set of Joe's beliefs ($\mathbf{B}_{\text{Joe}}$). It can be deduced from p, however, that "there are at least two people in Kōbe having the same number of hairs in their head (q)." The conclusion can be seen by supposing that one person is found (so far as possible) for each number of hairs, from a man with one hair to one with the maximum possible number. If these people are then seated in an amphitheatre in order by their number of hairs, there will still be people left unseated—each a contender for a seat already filled. Hence, Joe believes q, that "there are at least two people in Kōbe (viz., those pairs contending for seats) having the same number of hairs in their head," whether he realizes it or not. But while a person is certainly committed to accepting the consequences of his beliefs, it is implausible to say that he believes them; very possibly Joe may disavow that there are two such people in Kōbe. It is only in some strange and artificial sense of the term that one's belief in certain axioms can be held to entail belief in some extremely remote consequence thereof.

On the other hand, with the preparedness-for-explicit-assent construction, it seems impossible to build up any sort of logic whatsoever. On this construction, it may well be that:

1. "p is among the beliefs of x and not-p is among the beliefs of x"; *i.e.,* $p \in \mathbf{B}_x$ and $\sim p \in \mathbf{B}_x$ (example: "x believes that he can jump the chasm"; "x believes that he cannot jump the chasm");

or that

2. A man may assent to (believe in) each of two statements, p and q, taken separately—*i.e.,* both $p \in \mathbf{B}_x$ and $q \in \mathbf{B}_x$—but

fail to assent to their conjunction—$(p \cdot q) \notin \mathbf{B}_x$ in which the centred dot means "and." For example, someone might assent to each of the statements "God is benevolent" and "There is evil in the world" but fail to assent to (or, perhaps, even deny) their conjunction; *i.e.,* "God is benevolent but there is evil in the world."

Clearly, one may fail to develop a logic of belief when the beliefs at issue turn out to be highly illogical—as well they may.

Theory of belief. To articulate a viable logic of belief, it is, at the very least, essential to postulate certain minimal conditions of rationality regarding the parties whose beliefs are at issue:

Conditions of rationality

1. Consistency: "If x believes that p, then x does not believe that not-p"; *i.e.,*

$$\text{If } p \in \mathbf{B}_x, \text{ then } \sim p \notin \mathbf{B}_x.$$

"If not-p, then x does not believe that p"; *i.e.,*

$$\text{If } \vdash \sim p, \text{ then } p \notin \mathbf{B}_x.$$

Example: If "Jesus was a Zealot" (p) is among ($\in$) the beliefs of Ralph ($\mathbf{B}_{\text{Ralph}}$), then "Jesus was not a Zealot" ($\sim p$) is not among ($\notin$) Ralph's beliefs. It is an accepted thesis ($\vdash$) that "Jesus was not a Zealot." Hence, "Jesus was a Zealot" is not among Ralph's beliefs. (Note that the symbol $\vdash$, in cases where it appears as an isolated prefix, represents "thesis-hood"; *i.e.,* a deductive consequence—as it were—from no premises.)

2. Conjunctive composition and division: "If x believes that p_1, and x believes that p_2, etc., to x believes that p_n, then x believes that p_1 and p_2, etc., and p_n"; *i.e.,*

$$\text{If } (p_1 \in \mathbf{B}_x, p_2 \in \mathbf{B}_x, \ldots, p_n \in \mathbf{B}_x), \text{ then } (p_1 \cdot p_2 \cdot \ldots \cdot p_n) \in \mathbf{B}_x,$$

and conversely. Example: If "cats are affectionate" (p_1), "cats are clean" (p_2), etc., to "cats are furry" (p_n) are among ($\in$) Bob's beliefs ($\mathbf{B}_{\text{Bob}}$), then "cats are affectionate and clean, etc., and furry" ($p_1 \cdot p_2 \cdot \ldots \cdot p_n$) is also a belief of Bob's.

3. Minimal inferential capacity: "If x believes that p, and q is an obvious consequence of p, then x believes that q"; *i.e.,*

$$\text{If } p \in \mathbf{B}_x \text{ and } p \models q, \text{ then } q \in \mathbf{B}_x.$$

Example: "If x believes that his cat is on the mat, and his cat's being on the mat has an obvious consequence that *something* is on the mat, then x believes that something is on the mat."

Here item three is a form of the entailment principle, but with $\models$ representing entailment of the simplest sort, designating obvious consequence—say, deducibility by fewer than two (or n) inferential steps, employing only those primitive rules of inference that have been classified as obvious. (In arguments about beliefs, however, all repetitions of the application of this version of the entailment principle must be avoided.) These principles endow the theory with such rules as:

1. "If x believes that not-p, then he does not believe that p"; *i.e.,*

$$\text{If } \sim p \in \mathbf{B}_x, \text{ then } p \notin \mathbf{B}_x.$$

2. "If x believes that p, and x believes that q, then x believes that both p and q taken together"; *i.e.,*

$$\text{If } p \in \mathbf{B}_x \text{ and } q \in \mathbf{B}_x, \text{ then } p \cdot q \in \mathbf{B}_x.$$

3. "If x believes that p, then x believes that either p or q"; *i.e.,*

$$\text{If } p \in \mathbf{B}_x, \text{ then } p \vee q \in \mathbf{B}_x,$$

given "$p \vdash p \vee q$" as an "obvious" rule of inference (where $\vee$ means "either—or").

Iterative beliefs

One key question of the logical theory of belief relates to the area of iterative beliefs (example: "Andrews believes that I believe that he believes me to be untrustworthy"). Clearly, one would not want to have such theses as:

1. "If y believes that x believes that p, then x believes that p"; *i.e.,*

$$\text{If } (p \in \mathbf{B}_x) \in \mathbf{B}_y, \text{ then } p \in \mathbf{B}_x \qquad (y \neq x)$$

nor

2. "If y believes that x believes that p, then y believes that p"; *i.e.,*

$$\text{If } (p \in \mathbf{B}_x) \in \mathbf{B}_y, \text{ then } p \in \mathbf{B}_y \qquad (y \neq x)$$

But when the iteration is subject-uniform rather than subject-diverse, it might be advantageous to postulate certain special theses, such as:

$$\text{If } p \in \mathbf{B}_x, \text{ then } (p \in \mathbf{B}_x) \in \mathbf{B}_x,$$

which in effect limits the beliefs at issue to conscious beliefs. The plausibility of this thesis also implicates its converse; viz., whether there are circumstances under which someone's believing that he believes something would necessarily vouch for his believing of it—*i.e.,* whether it is legitimate to argue that "If x believes that he believes that p, then he believes that p"; *i.e.,*

$$\text{If } (p \in \mathbf{B}_x) \in \mathbf{B}_x, \text{ then } p \in \mathbf{B}_x.$$

According to this thesis, the belief set $\mathbf{B}_x$ is to have the feature of second-order—as opposed to direct—applicability. From $q \in \mathbf{B}_x$, it is not, in general, permissible to infer q, but one is entitled to do so when q takes the special form $p \in \mathbf{B}_x$; *i.e.,* when the belief at issue is one about the subject's own beliefs.

A theory of belief along these lines is a halfway house between the commitment-to-assent conception of belief, which is too inclusive and the explicit-assent approach, which is too restrictive. The theory is predicated on the view that belief is subject to logical compulsion but that the range of this compulsion is limited since men are not logically omniscient. Belief here is like sight: man has a limited range of logical vision; he can see clearly in the immediate logical neighbourhood of his beliefs but only dimly afar.

The logic of knowing. The propositional sense of knowing (*i.e.,* knowing that something or other is the case), rather than the operational sense of knowing (*i.e.,* knowing how something or other is done), is generally taken as the starting point for a logical theory of knowing. Accordingly, the logician may begin with a person x and consider a set of propositions $\mathbf{K}_x$ to represent his "body of knowledge." The aim of the theory then is to clarify and to characterize the relationship "x knows that p" or "p is among the items known to x," which is here represented as: $p \in \mathbf{K}_x$.

"Knowing as" and "being" true

There can be false knowledge only in the sense that "he *thought* he knew that p, but he was mistaken." When the falsity of purported knowledge becomes manifest, the claim to knowledge must be withdrawn. "I know that p, but it may be false that p" is a contradiction in terms. When something is asserted or admitted as known, it follows that this must be claimed to be true. But what sort of inferential step is at issue in the thesis that "x knows p" leads to "p is true"? Is the link deductive, inductive, presuppositional, or somehow "pragmatic"? Each view has its supporters: on the deductive approach, $p \in \mathbf{K}_x$ logically implies (deductively entails) p; on the inductive approach, $p \in \mathbf{K}_x$ renders p extremely probable, though not necessarily certain; on the presuppositional approach, $p \in \mathbf{K}_x$ is improper (nonsensical) whenever p is not true; and on the pragmatic approach, the assertion of $p \in \mathbf{K}_x$ carries with it a rational commitment to the assertion of p (in a manner, however, that does not amount to deductive entailment). From the standpoint of a logic of knowing, the most usual practice is to assume the deductive approach and to lay it down as a rule that if $p \in \mathbf{K}_x$, then p is true. This approach construes knowledge in a very strong sense.

According to a common formula, knowledge is "true, justified belief." This formulation, however, seems defective. Let the expression $\mathbf{J}xp$ be defined as meaning "x has justification for accepting p," then:

$$p \in \mathbf{K}_x = p \cdot \mathbf{J}xp \cdot p \in \mathbf{B}_x.$$

Example: The proposition "Jane knows that ($\mathbf{K}_{\text{Jane}}$) the gown is priceless (p)" means (=) "The gown is priceless, and Jane has justification for accepting that it is priceless ($\mathbf{J}_{\text{Jane}}\ p$) and Jane believes that it is priceless ($p \in \mathbf{B}_{\text{Jane}}$)." One cannot but assume that the conceptual nature of $\mathbf{J}$ is such as to underwrite the rule: "If x is justified in accepting p, then he is justified in accepting 'Either p or q' "; *i.e.,*

$$\text{(J)} \qquad \text{If } \mathbf{J}xp, \text{ then } \mathbf{J}x\,(p \vee q),$$

in which q can be any other proposition whatsoever. The components p, q, and x may be such that all of the following obtain:

1. not-p
2. q
3. x believes that p; *i.e.,* $p \in \mathbf{B}_x$
4. x does not believe that q; *i.e.,* $q \notin \mathbf{B}_x$ and, indeed, x believes that not-q; *i.e.,* $\sim q \in \mathbf{B}_x$

5. x is justified in accepting q; *i.e.*, $\mathbf{J}xq$
6. x believes that either p or q; *i.e.*, $p \vee q \in \mathbf{B}_x$.

Clearly, on any reasonable interpretation of **B** and **J**, this combination of six premises is possible. But the following consequences would then obtain:

7. $p \vee q$	by item 2 above
8. $\mathbf{J}x(p \vee q)$	by item 5 above and by **J**
9. $(p \vee q) \in \mathbf{K}_x$	by items 6, 7, and 8.

However the conclusion nine is wrong: x cannot properly be said to know that either p or q when $p \vee q$ is true solely because of the truth of q (which x rejects), but $p \vee q$ is believed by x solely because he accepts p (which is false). This example shows that the proposed definition of knowledge as "true, justified belief" cannot be made to work; and the best plan, therefore, seems to be to treat the logic of knowing directly, rather than through the mediation of acceptance (belief) and justification.

Overt versus tacit knowledge

Since Aristotle's day, stress has been placed on the distinction between actual, overt knowledge that requires an explicit, consciously occurring awareness of what is known and potential, tacit knowledge that requires only implicit dispositional awareness. Unless $p \in \mathbf{K}_x$ is construed in the tacit sense, the following principles will not hold:

If $p \in \mathbf{K}_x$ and $p \vdash q$, then $q \in \mathbf{K}_x$.

If $p \in \mathbf{K}_x$ and $q \in \mathbf{K}_x$, then $(p \cdot q) \in \mathbf{K}_x$.

These two rules, if accepted, however, suffice to guarantee the principle:

If $p_1, p_2, \ldots, p_n \vdash q$, then $p_1 \in \mathbf{K}_x, p_2 \in \mathbf{K}_x, \ldots, p_n \in \mathbf{K}_x \vdash q \in \mathbf{K}_x$.

Similar considerations regarding the potential construction of knowledge govern the answer to the question of whether, when something is known, this fact itself is known: if $p \in \mathbf{K}_x$, then $(p \in \mathbf{K}_x) \in \mathbf{K}_x$. This principle is eminently plausible, provided that the membership of $\mathbf{K}_x$ is construed in the implicit (tacit) rather than in the explicit (overt) sense.

The logic of questions. Whether a given grouping of words is functioning as a question may hinge upon intonation, accentuation, or even upon context, rather than upon overt form: at bottom, questions represent a functional rather than a purely grammatical category. The very concept of a question is correlative with that of an answer; and every question correspondingly delimits a range of possible answers. One way of classifying questions is in terms of the surface characteristics of this range. On this basis, the logician can distinguish (among others):

1. Yes/no questions (example: "Is today Tuesday?");
2. Item-specification questions (example: "What is an instance of a prime number?");
3. Instruction-seeking questions (example: "How does one bake an apple pie?"); etc.

Propositional approach

From the logical standpoint, however, a more comprehensive policy and one leading to greater precision is to treat every answer as given in a complete proposition ("Today is not Tuesday"; "Three is an example of a prime number"; etc.). From this standpoint, questions can be classed in terms of the nature of the answers. There would then be factual questions (example: "What day is today?") and normative questions (example: "What ought to be done in these circumstances?").

The advantage of the propositional approach to answers is that it captures the intrinsically close relationship between question and answer. The possible answers to (1) "What is the population of A-ville?" and (2) "What is the population of B-burgh?" are seemingly the same; namely, numbers of the series 0, 1, 2, ... But once complete propositions are taken to be at issue, then an answer to (1), such as "The population of A-ville is 5,238," no longer counts as an answer to (2), since the latter must mention B-burgh. This approach has the disadvantage, on the other hand, of obscuring similarities in similar questions. One can now no longer say of two brothers that the questions "Who is Tom's father?" and "Who is John's father?" have the same answer.

With every question Q can be correlated the set of propositions $\mathbf{A}(Q)$ of possible answers to Q. Thus, "What day of the week is today?" has seven conceivable answers, of the form "The day of the week today is Monday," and the like. A possible answer to a question must be a possibly true statement. Accordingly, the question "What is an example of a prime number?" does not have "The Washington Monument is an example of a prime number" among its possible answers.

Truth-values in question theory

A question can be said to be true if it has a true answer; *i.e.*, if $(\exists p)[p \cdot p \in \mathbf{A}(Q)]$—which (taking the existential quantifier $\exists$ to mean "there exists ...") can be read "There exists a proposition p such that p is true and p is among the answers of Q"; otherwise it is false; *i.e.*, when all of its answers are false. If he never came at all, the question "On what day of the week did he come?" is a false question in the sense that it lacks any true answer.

A true question can be called contingent if it admits of possible answers that are false, as in "Where did Jones put his pen?" In logic and mathematics there are, presumably, no contingent questions.

Questions can have presuppositions, as in "Why does Smith dislike Jones?" Any possible answer here must take the form "Smith dislikes Jones because ..." and so commits one to the claim that "Smith dislikes Jones." Every such question with a false presupposition must be a false question: all of its possible answers (if any) are false.

Besides falsity, questions can exhibit an even more drastic sort of "impropriety." They can be illegitimate in that they have no possible answers whatsoever. (Example: "What is an example of an even prime number different from two?") The logic of questions is correspondingly three valued: a question can be true (*i.e.*, have a true answer); illegitimate (*i.e.*, have no possible answer at all); or false (*i.e.*, have possible answers but no true ones).

One question, Q_1, will entail another, Q_2, if every possible answer to the first deductively yields a possible answer to the second, and every true answer to the first deductively yields a true answer to the second. In this sense the question "What are the dimensions of that box?" entails the question "What is the height of that box?"

PRACTICAL LOGIC

The theory of reasoning with concepts of practice—of analyzing the logical relations obtaining among statements about actions and their accompaniments in choosing, planning, commanding, permitting, etc.—comprises the domain of practical logic.

The logic of preference. The logic of preference—also called the logic of choice, or prohairetic logic (Greek *proairesis,* "a choosing")—seeks to systematize the formal rules that govern the conception "x is preferred to y." A diversity of things can be at issue here: (1) Is x preferred to y by some individual (or group), or is x preferable to y in terms of some impersonal criterion? (2) Is on-balance preferability at issue or preferability in point of some particular factor (such as economy or safety or durability)? The resolution of these questions, though vital for interpretation, does not affect the formal structure of the preference relationships.

Symbolization and approach taken in prohairetic logic. The fundamental tools of the logic of preference are as follows: (1) (strong) preference: x is preferable to y, symbolically $x \gg y$; (2) indifference: x and y are indifferent, $x \cong y$, defined as "neither $x \gg y$ nor $y \gg x$"; (3) weak preference: x is no less preferred than y, $x \gtrsim y$, defined as "either $x \gg y$ or $x \cong y$." Since preference comprises a relationship, its three types can be classed in terms of certain distinctions commonly drawn in the logic of relations: that of reflexivity (whether holding of itself: "John supports himself"), of symmetry (whether holding when its terms are interchanged: "Peter is the cousin of Paul"; "Paul is the cousin of Peter"), and of transitivity (whether transferable: "a > b"; "b > c"; therefore, "a > c"). Once it is established that the (strong) preference relation ($\gg$) is an ordering (*i.e.*, is irreflexive, asymmetric, and transitive), it then follows that weak preference ($\gtrsim$) is reflexive, nonsymmetric, and transitive and that indifference ($\cong$) is an equivalence relation (*i.e.*, reflexive, symmetric, and transitive).

The "measure of merit"

One common approach to establishing a preference relation is to begin with a "measure of merit" to evaluate the relative desirability of the items $x, y, z, \ldots$, that are at issue. Thus for any item x, a real-number quantity is ob-

tained, symbolized # (x). (Such a measure is called a utility measure, the units are called utiles, and the comparisons or computations involved comprise a preference calculus.) In terms of such a measure, a preference ordering is readily introduced by the definitions that: (1) $x \gg y$ is to be construed as # (x) > # (y); (2) $x \geqslant y$ as # (x) ≥ # (y); and (3) $x \cong y$ as # (x) = # (y), in which ≥ means "is greater than or equal to." Given these definitions, the relationships enumerated above must all obtain. Thus, the step from a utility measure to a preference ordering is simple.

Construction of a logic of preference. In constructing a logic of preference, it is assumed that the items at issue are propositions $p, q, r, \ldots$ and that the logician is to introduce a preferential ordering among them, with $p \gg q$ to mean "p's being the case is preferred to q's being the case." The problem is to systematize the logical relationships among such statements in order to permit a determination of whether, for example, it is acceptable to argue that: If either p is preferable to q or p is preferable to r, then p is preferable to either q or r,

$$(p \gg q \vee p \gg r) \supset [p \gg (q \vee r)]$$

(in which ⊃ means "implies" or "if . . . then"), or to argue similarly that

$$(p \gg q \cdot r \gg q) \supset [(p \cdot r) \gg q].$$

Example: "If eating pears (p) is preferable to eating quinces (q) and eating rhubarb (r) is preferable to eating quinces, then eating both pears and rhubarb is preferable to eating quinces." The task is one of erecting a foundation for the systematization of the formal rules governing such a propositional preference relation—a foundation that can be either axiomatic or linguistic (*i.e.*, in terms of a semantical criterion of acceptability).

Approach in terms of possible worlds

One procedure—adapted from the ideas of the well-known contemporary Finnish philosopher, Georg Henrik von Wright (born 1916), a prolific contributor to applied logic—is as follows: beginning with a basic set of possible worlds (or states of affairs) $w_1, w_2, \ldots, w_n$, all of the propositions to be dealt with are first defined with respect to these by the usual logical connectives (∨, ·, ⊃, etc.). Given two elementary propositions p and q, there are just the following possibilities: both are true, p is true and q is false, p is false and q is true, or both are false. Corresponding to each of these possibilities is a possible world; thus,

$$w_1 = p \cdot q$$
$$w_2 = p \cdot \sim q$$
$$w_3 = \sim p \cdot q$$
$$w_4 = \sim p \cdot \sim q.$$

The truth of p then amounts to the statement that one of the worlds w_1, w_2 obtains, so that p is equivalent to $w_1 \vee w_2$. Moreover, a given basic preference/indifference ordering among the w_i is assumed. On this basis the following general characterization of propositional preference is stipulated: If delta (δ) is taken to represent any (and thus every) proposition independent of p and q, then p is preferable to q, $p \gg q$, if for every such δ it is the case that every possible world in which p and not-q and δ are the case, $p \cdot \sim q \cdot \delta$ is w-preferable to every possible world in which not-p and q and δ is the case, $\sim p \cdot q \cdot \delta$; *i.e.*, when $p \cdot \sim q$ is always preferable to $\sim p \cdot q$ provided that everything else is equal. It is readily shown that through this approach such general rules as the following are obtained:

1. If p is preferable to q, then q is not preferable to p; *i.e.*,

$$p \gg q \vdash \sim (q \gg p).$$

2. If p is preferable to q, and q is preferable to r, then p is preferable to r; *i.e.*,

$$(p \gg q \cdot q \gg r) \vdash (p \gg r).$$

3. If p is preferable to q, then not-q is preferable to not-p; *i.e.*,

$$p \gg q \vdash \sim q \gg \sim p.$$

4. If p is preferable to q, then having p and not-q is preferable to having not-p and q; *i.e.*,

$$p \gg q \vdash (p \cdot \sim q) \gg (\sim p \cdot q).$$

The preceding construction of preference requires only a preference ordering of the possible worlds. If, however, a measure for both probability and desirability (utility) of possible worlds is given, then one can define the corresponding #-value (see below) of an arbitrary proposition p as the probabilistically weighed utility value of all of the possible worlds in which the proposition obtains. As an example, p may be the statement, "The Franklin Club caters chiefly to business people" and q the statement, "The Franklin Club is sports-oriented." It may then be supposed as given that the following values hold:

World	Probability	Desirability
$w_1 = p \cdot q$	1/6	−2
$w_2 = p \cdot \sim q$	2/6	+1
$w_3 = \sim p \cdot q$	3/6	−1
$w_4 = \sim p \cdot \sim q$	1/6	+3.

The #-value of a proposition is determined by first multiplying the probability times the desirability of each world in which the proposition is true, and then taking the sum of these. For example, the #-value of p is determined as follows: p is true in each of w_1 and w_2 (and only these); the probability times the desirability of w_1 is $1/6 \times (-2)$ and of w_2 is $2/6 \times (+1)$; thus #-(p) is $1/6 \times (-2) + 2/6 \times (+1) = 0$. (The #-value corresponds to the decision theorists' notion of expected value.) By this procedure it can easily be determined that

$$\#(p) = 0 \qquad \#(\sim p) = 1/6$$
$$\#(q) = -(4/6) \qquad \#(\sim q) = 5/6.$$

Since both #(p) > #(q) and #($\sim q$) > #($\sim p$), one correspondingly obtains both $p \gg q$ and $\sim q \gg \sim p$ in the example at issue; *i.e.*, "That the Franklin Club should cater chiefly to business people is preferable to its being sports oriented" and "Its not being sports oriented is preferable to its not catering chiefly to business people." (The result is, of course, relative to the given desirability schedule specified for the various possible-world combinations in the above tabulation.)

A more complex mode of preference results, however, if—when some basic utility measure, #(x), is given—instead of having $p \gg q$ correspond to the condition that #(p) > #(q), it is taken to correspond to #(p) − #($\sim p$) > #(q) − #($\sim q$). This mode will be governed by characteristic rules, specifically including all of those listed above.

The logic of commands. Some scholars have maintained that there cannot be a logic of commands (instructions, orders), inasmuch as there can be no logic in which validity of inference cannot be defined. Validity, however, requires that the concept of truth be applicable (an argument being valid when its conclusion must be true if its premises are true). But since commands—and, for that matter, also instructions, requests, etc.—are neither true nor false, it is argued that the concept of validity cannot be applied; so there can be no valid inference in this sphere. This line of thought, however, runs counter to clear intuitions that arise in specific cases, in which one unhesitatingly reasons from commands and sets of commands. If an examination has the instructions "Answer no fewer than three questions! Answer no more than four questions!" one would not hesitate to say that this implies the instruction, "Answer three or four questions!"

Truth involvements

This seeming impasse can be broken, in effect, by importing truth into the sphere of commands through the back door: with any command one can associate its termination statement, which, with future-tense reference, asserts it as a fact that what the command orders will be done. Thus, the command "Shut all of the windows in the building!" has the termination statement "All of the windows in the building will be shut." In case of a pure command argument—*i.e.*, one that infers a command conclusion from premises that are all commands—validity can be assessed in the light of the validity of the purely assertoric syllogism composed of the corresponding termination statements. Thus the validity of the command argument given above derives from the validity of the inference, from the premises "No fewer than three questions will be answered; no more than four questions will be answered" to the conclusion "Three or four questions will be answered."

The logical issues of pure command inference can be handled in this manner. But what of the mixed cases

in which some statement—premise or conclusion—is not a command?

Special cases

Special case 1. One mixed case is that in which the premises nontrivially include noncommands, but the inferred conclusion is a command. Some writers have endorsed the rule that there is no validity unless the command conclusion is forthcoming from the command premises alone. This, however, invalidates such seemingly acceptable arguments as "Remove all cats from the area; the shed is in the area; so, remove all cats from the shed." It is more plausible, however, to stipulate the weaker condition that an inference to a command conclusion cannot count as valid unless there is at least one command premise that is essential to the argument. Subject to this restriction, a straightforward application of the above-stated characterization of validity can again be made. This approach validates the above-mentioned command inference via the validity of the assertion inference: "All cats will be removed from the area; the shed is in the area; so, all cats will be removed from the shed." (The rule under consideration suffices to block the unacceptable argument from the factual premise "All the doors will be shut" to the command conclusion "Shut all the doors.")

Special case 2. Another mixed case is that in which the premises nontrivially include commands, but the inferred conclusion is an ordinary statement of fact. Some authorities stipulate that no indicative conclusion can be validly drawn from a set of premises which cannot validly be drawn from the indicative among them alone. This rule would seem to be acceptable, though subject to certain significant provisos: (1) It must be restricted to categorical rather than conditional commands. "If you want to see one of the world's tallest buildings, look at the Empire State Building" conveys (*inter alia*) the information that "The Empire State Building is one of the world's tallest buildings." (2) Exception must be made for those commands that include in their formulation—explicitly or by way of tacit presupposition—reference to a factual datum. "John, give the book to Tom's brother Jim" yields the fact that Jim is Tom's brother; and "John, drive your car home" (= "John, you own a car: drive it home") yields "John owns a car." With suitable provisos, however, the rule can be maintained to resolve the issues of the special case in view.

Deontic logic. The propositional modalities relating to normative (or valuational) classifications of actions and states of affairs, such as the permitted, the obligatory, the forbidden, or the meritorious, are characterized as deontic modalities (Greek *deontos,* "of that which is binding") and systematized in deontic logic. Though this subject was first treated as a technical discipline in 1926, its current active development dates from a paper published in 1951. As a highly abstracted branch of logical theory, it leaves to substantive disciplines—such as ethics and law—the concrete questions of what specific acts or states of affairs are to be forbidden, permitted, or the like (just as deductive logic does not meddle with what contingent issues are true but tells only what follows when certain facts or assumptions about the truth are given). It seeks to systematize the abstract, purely conceptual relations between propositions in this sphere, such as the following: if an act is obligatory, then its performance must be permitted and its omission forbidden. In given circumstances, any act is either permitted itself or its omission is permitted.

Reduction to permissibility terms

The systematization and relation to alethic modal logic. In the systematization of deontic logic, the symbols $p, q, r, \ldots$ may be taken to range over propositions dealing both with impersonal states of affairs and with the human acts involved in their realization. Certain special deontic operations can then be introduced: $\mathbf{P}(p)$ for "It is permitted that p be the case"; $\mathbf{F}(p)$ for "It is forbidden that p be the case"; and $\mathbf{O}(p)$ for "It is obligatory that p be the case." In a systematization of deontic logic, it is only necessary to take one of these three operations as primitive (*i.e.,* as an irreducible given), because the others can then be introduced in terms of it. For example, when **P** alone is taken as primitive (as is done here), the following can be introduced by definition: "It is obligatory that p" means "It is not permitted that not-p"; and "It is forbidden that p" means "It is not permitted that p"; *i.e.,*

$$\mathbf{O}(p) = \sim\mathbf{P}(\sim p) \text{ and } \mathbf{F}(p) = \sim\mathbf{P}(p).$$

The logical grammar of **P** is presumably to be such that one wants to insist upon the rule:

$$\text{Whenever } \vdash p \supset q, \text{ then } \vdash \mathbf{P}(p) \supset \mathbf{P}(q).$$

Example: "Since one's helping Jones who has been robbed entails that one help someone who has been robbed, being permitted to help Jones (who has been robbed) entails that one be permitted to help someone who has been robbed." This yields such principles as: "If both p and q are permitted, then p is permitted and q is permitted"; and "If p is permitted, then either p or q is permitted"; *i.e.,*

$$\vdash \mathbf{P}(p \cdot q) \supset [\mathbf{P}(p) \cdot \mathbf{P}(q)] \text{ and } \vdash \mathbf{P}(p) \supset \mathbf{P}(p \vee q).$$

And, once it is postulated that "A p exists that is permitted"—*i.e.,* $\vdash (\exists p)\,\mathbf{P}(p)$—then the statement that "It is not permitted that both p and not-p"—*i.e.,* $\sim\mathbf{P}(p \cdot \sim p)$—is also yielded. Moreover, on any adequate theory of **P** it is necessary to have such principles as "Either p or not-p is permitted"; *i.e.,* $\vdash \mathbf{P}(p \vee \sim p)$.

On the other hand, certain principles must be rejected, such as that "If p is permitted and q is permitted, then both p and q taken together is permitted"; *i.e.,* $\vdash [\mathbf{P}(p) \cdot \mathbf{P}(q)] \supset \mathbf{P}(p \cdot q)$, in which $\vdash$ symbolizes the rejection of a thesis, and that "If either p or q is permitted, then p is permitted"; *i.e.,* $\vdash \mathbf{P}(p \vee q) \supset \mathbf{P}(p)$.

The first of these, accepted unqualifiedly, would lead to the untenable result that there can be no permission-indifferent acts; *i.e.,* no acts such that both they and their omission are permitted—since this would then lead to $\mathbf{P}(p \cdot \sim p)$. The second thesis would have the unacceptable result of asserting that, when at least one member of a pair of acts is permitted, then both members are permitted.

Analogy with possibility and necessity

In all respects so far considered, deontic logic is wholly analogous to the already well developed field of alethic modal logic, which deals with statements of the form "It is possible that . . ." (symbolized $\Diamond$), "It is necessary that . . ." (symbolized $\Box$), etc. (see above *Modal logic*), with **P** in the role of possibility ($\Diamond$) and **O** in that of necessity ($\Box$). This parallel, however, does not extend throughout. In alethic logic, the principle that "necessity implies actuality" obviously holds (*i.e.,* $\vdash \Box p \supset p$). But its deontic analogue, that "obligation implies actuality" (*i.e.,* $\vdash \mathbf{O}p \supset p$), must be rejected, or rather an analogous thesis holds only in the weakened form that "obligation implies permissibility" (*i.e.,* $\vdash \mathbf{O}p \supset \mathbf{P}p$). Controversy exists about the relation of deontic to alethic modal logic, principally in the context of Kant's thesis that "ought implies can" (*i.e.,* $\vdash \mathbf{O}p \supset p$), but also about the theses that "necessity implies permissibility" (*i.e.,* $\vdash \Box p \supset \mathbf{P}p$) and *Ad impossibile nemo obligatur*—"no one is obliged to [do] the impossible" (*i.e.,* $\vdash \sim\Diamond p \supset \sim\mathbf{O}p$). Although this thesis is generally accepted, some scholars want to strengthen the thesis to "necessity implies obligation" (*i.e.,* $\vdash \Box p \supset \mathbf{O}p$), or, equivalently, to "permissibility implies possibility" (*i.e.,* $\vdash \sim\Diamond p \supset \sim\mathbf{P}p$), with the result that only what is possible can count as permitted, so that the impossible is forbidden. Some would deny that it is wrong (*i.e.,* impermissible) to act to realize the impossible, rather than merely unwise.

It has been proposed that deontic logic may perhaps be reduced to alethic modal logic. This approach is based on the idea of a normative code delimiting the range of the permissible. In this context, what signalizes an action as impermissible is that it involves a violation of the code: the statement that the action has occurred entails that the code has been violated and so leads to a "sanction." This line of thought leads to the definition of a modal operator $\mathbf{F}p = \Box(p \supset \sigma)$, "$p$ necessarily implies a sanction," in which sigma (σ) is the sanction produced by code violation. Correspondingly, one then obtains: "For p to be permitted means that p does not imply by necessity a sanction"—*i.e.,* $\mathbf{P}p = \sim\Box(p \supset \sigma)$—and "For p to be obligatory means that not doing p implies by necessity a sanction"—*i.e.,* $\mathbf{O}p = \Box(\sim p \supset \sigma)$. Assuming a systematization of the alethic modal operator $\Box$, these definitions immediately produce a corresponding system of deontic

logic that—if $\Box$ is a normal modality—has many of the features that are desirable in a modal operator. It also yields, however—through the "paradoxes of strict implication" (see above *Alternative systems of modal logic*)—the disputed principle that "The assumption that p is not possible implies that p is not permissible"; *i.e.*, $\vdash \sim\Diamond p \supset \sim \mathbf{P}p$. This and other similar consequences of the foregoing effort to reduce deontic logic to modal logic have been transcended by other scholars, who have resorted to a mode of implication (symbolized as $\rightarrow$) that is stronger than strict implication (as necessary material implication is called) and then defining $\mathbf{F}p$ as $p \rightarrow \sigma$ instead of as above.

Alternative deontic systems. Each of the three principal deontic systems that have been studied to date is analogous to one of the alethic modal systems that were developed in the mid-20th century.

Analogies with **M**, **S4**, and **S5**

These foundational alethic systems differ by virtue of the different axioms and rules adopted for such modalities as necessity, possibility, and contingency. In the system designated **M**, for example, developed by the aforementioned Finnish logician G.H. von Wright, the adverb "possibly," symbolized M, is taken as the fundamental undefined modality in terms of which the other modalities are constructed. "Necessarily p," symbolized Lp, for example, is defined in the system **M** as "not possibly not-p"; *i.e.*, $Lp = \sim M \sim p$. Alternatively, in an equivalent system, **T**, "necessarily p" is taken as primitive, and "possibly p" is defined as "not necessarily not-p"; *i.e.*, $Mp = \sim L \sim p$. Several nonequivalent systems have been developed by the conceptual Pragmatist C.I. Lewis (1883–1964), primary author of *Symbolic Logic* (1932), the foundational work in this field. Of these systems, that known as **S4** includes all of the system **M** but adds also the axiom that " 'Necessarily p' implies 'It is necessary that necessarily p' "—*i.e.*, $Lp \supset LLp$—whereas, that known as **S5** adds still another axiom, that " 'Possibly p' implies 'It is necessary that possibly p' "—*i.e.*, $Mp \supset LMp$ (see above *Alternative systems of modal logic*). The analogous deontic systems are then as follows:

1. **DM** (the deontic analogue of the system **M** of von Wright or of the system **T**). To a standard system of propositional logic the following rule is added: "Any proposition, if true, ought to be true"; *i.e.*, If $\vdash p$ then $\vdash \mathbf{O}p$. Example: Given that "To forgive is divine," (p) then "To forgive ought to be divine" ($\mathbf{O}p$). Axioms:
(A1): "If p is obligatory, then not-p is not obligatory"; *i.e.*, $\mathbf{O}p \supset \sim\mathbf{O}\sim p$.
(A2): "If p ought to imply q, then if p is obligatory q is obligatory"; *i.e.*, $\mathbf{O}(p \supset q) \supset (\mathbf{O}p \supset \mathbf{O}q)$.
2. **DS4** (the deontic analogue of Lewis's system **S4**). To **M** one adds the axiom:
(A3): "If p is obligatory, then p ought to be obligatory"; *i.e.*, $\mathbf{O}p \supset \mathbf{OO}p$. Example: "If John ought to pay his debts," ($\mathbf{O}p$) then it is obligatory that John ought to pay his debts" ($\mathbf{OO}p$).
3. **DS5** (the deontic analogue of Lewis's system **S5**). To **M** one adds the axiom:
(A4): "If p is not obligatory, then p ought to be nonobligatory"; *i.e.*, $\sim\mathbf{O}p \supset \mathbf{O}\sim\mathbf{O}p$.

Semantic systematization

A straightforward semantical systematization of systems of deontic logic can be provided as follows: given a domain of complex propositions built up from atomic propositions (p, q, r, ...) with the use of propositional connectives ($\sim$, $\cdot$, $\vee$, $\supset$) and **O**, a deontic model set Δ for this domain can be characterized as any set chosen from these propositions that meets the following conditions (in which "iff" means "if and only if"):

1. Not-p is in the set if and only if p is not in the set; *i.e.*, $\sim p \in \Delta$ iff $p \notin \Delta$.
2. "Both p and q together" is in the set if and only if p is in the set and q is in the set; *i.e.*, $(p \cdot q) \in \Delta$ iff $p \in \Delta$ and $q \in \Delta$.
3. "Either p or q" is in the set if and only if either p is in the set or q is in the set; *i.e.*, $(p \vee q) \in \Delta$ iff $p \in \Delta$ or $q \in \Delta$.
4. "That p implies q" is in the set if and only if either p is not in the set or q is in the set; *i.e.*, $(p \supset q) \in \Delta$ iff $p \notin \Delta$ or $q \in \Delta$.
5. "That p is obligatory" is in the set whenever p is posited; *i.e.*, $\mathbf{O}p \in \Delta$ whenever $\vdash p$.
6. "That not-p is not obligatory" is in the set whenever "p is obligatory" is in the set; *i.e.*, $\sim\mathbf{O} \sim p \in \Delta$ whenever $\mathbf{O}p \in \Delta$.
7. "That q is obligatory" is in the set whenever both "p is obligatory" is in the set and "That p implies q is obligatory" is in the set; *i.e.*, $\mathbf{O}q \in \Delta$ whenever both $\mathbf{O}p \in \Delta$ and $\mathbf{O}(p \supset q) \in \Delta$.

A proposition can be characterized as a deontic thesis (D-thesis) if it can be shown that, in virtue of these rules, it must belong to every deontic model set. It can be demonstrated that the D-thesis in this sense will coincide exactly with the theorems of **DM**—No. 1 of the above three systems. Furthermore, if one adds one of the additional rules:

8′. "That p ought to be obligatory" is in the set whenever "p is obligatory" is in the set; *i.e.*, $\mathbf{OO}p \in \Delta$ whenever $\mathbf{O}p \in \Delta$.
8″. "That p ought to be non-obligatory" is in the set whenever "p is not obligatory" is in the set; *i.e.*, $\mathbf{O}\sim\mathbf{O}p \in \Delta$ whenever $\sim\mathbf{O}p \in \Delta$.

then the corresponding D′ or D″ theses will coincide exactly with the theorems of the deontic systems **DS4** and **DS5**, respectively—Nos. 2 and 3 above.

LOGICS OF PHYSICAL APPLICATION

Certain systems of logic are built up specifically with particular physical applications in view. Within this range lie temporal logic; spatial, or topological, logic; mereology, or the logic of parts and wholes generally; as well as the logic of circuit analysis.

Since the field of topological logic is new and relatively undeveloped, the reader is referred to the bibliography for a recent source that provides some materials and references to the literature.

Of the physical applications of logic, the logic of circuit analysis is perhaps the most important and certainly the most extensively developed. Since the course of the analysis in this case, however, soon veers off from logical into mathematical considerations, this subject is not treated here.

Temporal logic. The object of temporal logic—variously called chronological logic or tense logic—is to systematize reasoning with time-related propositions. Such propositions generally do not involve the timeless "is" (or "are") of the mathematicians' "three is a prime," but rather envisage an explicitly temporal condition (examples: "Bob *is* sitting," "Robert *was* present," "Mary *will have been* informed"). In this area, statements are employed in which some essential reference to the before-after relationship or the past-present-future relationship is at issue; and the ideas of succession, change, and constancy enter in.

Classic historical treatments. Chronological logic originated with the Megarians of the 4th century BC, whose school (not far from Athens) reflected the influence of Socrates and of Eleaticism.

In the Megarian conception of modality, the *actual* is that which is realized *now*, the *possible* is that which is realized *at some time or other*, and the *necessary* is that which is realized *at all times*. These Megarian ideas can be found also in Aristotle, together with another temporalized sense of necessity according to which certain possibilities are *possible* prior to the event, *actual* then, and *necessary* thereafter, so that their modal status is not omnitemporal (as in the Megarian concept), but changes in time. The Stoic conception of temporal modality is yet another cognate development, according to which the *possible* is that which is realized at *some* time in the present or future and the *necessary* that which is realized at *all* such times. The Diodorean concept of implication (named after the 4th-century BC Megarian logician Diodorus Cronus) holds, for example, that the conditional "If the sun has risen, it is daytime" is to be given the temporal construction "All times after the sun has risen are times when it is daytime." The Persian logician Avicenna (980–1037), the foremost philosopher of medieval Islām, treated this chronological conception of implication in the framework of a general theory of categorical propositions (such as "All A is B," etc.) of a temporalized type and considerably advanced and developed the Megarian-Stoic theory of temporal modalities.

Fundamental concepts and relations of temporal logic. The statements "It sometimes rains in London," "It always rains in London," and "It is raining in London on January 1, AD 3000, is a fact" are all termed chronologically definite in that their truth or falsity is independent

Definite versus indefinite reference

of their time assertion. By contrast, the statements "It is now raining in London," "It rained in London yesterday," and "It will rain in London sometime next week" are all chronologically indefinite, in that their truth or falsity is not independent of their time of assertion. The notation $| t \vdash p$ is here introduced to mean that the proposition p, often in itself chronologically indefinite, is represented as being asserted at the time t. For example, if p_1 is the statement "It is raining in London today," and t_1 is January 1, 1900, then "$| t_1 \vdash p_1$" represents the assertion made on January 1, 1900, that it is raining today—an assertion that is true if and only if the statement "It is raining in London on January 1, 1900, is a fact" is true. If the statement p is chronologically definite, then (by definition) the assertions "$| t \vdash p$" and "$| t' \vdash p$" are materially equivalent (*i.e.,* have the same truth value) for all values of t and t'. Otherwise p is chronologically indefinite. The time may be measured, for example, in units of days, so that the time variable is discretized. Then $(t + 1)$ will represent "the day after t-day," $(t - 1)$ will represent "the day before t-day," and the like. And, further, the statements p_1, q_1, and r_1, can then be as follows:

p_1: "It rains in London today."
q_1: "It will rain in London tomorrow."
r_1: "It rained in London yesterday."

The following assertions can now be made:

(P): $| t \vdash p_1$
(Q): $| t - 1 \vdash q_1$
(R): $| t + 1 \vdash r_1$.

Clearly, for any value of t whatsoever, the assertions (P), (Q), and (R) must (logically) be materially equivalent (*i.e.,* have the same truth value). This illustration establishes the basic point—that the theory of chronological propositions must be prepared to exhibit the existence of logical relationships among such propositions of such a kind that the truth of the assertion of one statement at one time may be bound up essentially with the truth (or falsity) of the assertion of some very different statement at another time.

Dated versus pseudo-dated statements

A (genuine) date is a time specification that is chronologically stable (such as "January 1, 3000," or "the day of Lincoln's assassination"); a pseudodate is a time specification that is chronologically unstable (such as "today" or "six weeks ago"). These lead to very different results depending upon the nature of the fundamental reference point—the "origin" in mathematical terms. If the origin is a pseudodate—say, "today"—the style of dating will be such that its chronological specifiers are pseudodates; *e.g.,* tomorrow, day before yesterday, four days ago, etc. If, on the other hand, the origin is a genuine date, say that of the founding of Rome, or the accession of Alexander, the style of dating will be such that all of its dates are of the type: two hundred and fifty years *ab urbe condita* ("since the founding of the city"). Clearly, a chronology of genuine dates will then be chronologically definite, and one of pseudodates will be chronologically indefinite.

Let p be some chronologically indefinite statement. Then, in general, another statement can be formed, asserting that p holds (obtains) at the time t. Correspondingly, let the statement-forming operation $\mathbf{R}_t$ be introduced. The statement $\mathbf{R}_t(p)$, which is to be read "p is realized at the time t," will then represent the statement stating explicitly that p holds (obtains) specifically at the time t. Thus, if t_1 is 3 PM Greenwich time on January 1, 2000, and p_1 is the (chronologically indefinite) statement "all men are (*i.e.,* are now) playing Chess," then "$\mathbf{R}_{t1}(p_1)$" is the statement "It is the case at 3 PM Greenwich time on January 1, 2000, that all men are playing Chess."

If t is a proper date (not a pseudodate), then "$\mathbf{R}_t(p)$" is always temporally definite. For example, if p_1 is the temporally indefinite statement "It is raining in London today," and t_1 is as specified in the last paragraph, then "$\mathbf{R}_{t1}(p_1)$" is the temporally definite statement "It is raining in London at 3 PM Greenwich time on January 1, 2000." On the other hand, if t is a pseudodate, "$\mathbf{R}_t(p)$" is temporally indefinite if p is indefinite. Thus, if t_2 is the pseudodate tomorrow, then "$\mathbf{R}_{t2}(p_1)$" means "It is the case tomorrow that it is raining in London today." How is the **R**-operation to be construed in a chronology based on the indexical pseudodate "now" as origin? When two t's are involved, shall the second t be measured from the original now-point or from the now point newly established by the first t? That is, should "$\mathbf{R}_t'[\mathbf{R}_t(p)]$" be interpreted as:

1. It will be the case t' time units from now that it will be the case t time units from (the original) now that p—or, equivalently, simply as "$\mathbf{R}_{t'}(p)$"; or as
2. It will be the case t' time units from now that it will be the case t time units thence (*i.e.,* from the new now) that p—or, equivalently, as "$\mathbf{R}_{(t'+t)}(p)$"?

Since each of these represents a feasible alternative policy for the construction of **R**, it is necessary in fact to deal with the consequences of the adoption of either.

Systematization of temporal reasoning. On the basis of these ideas, the logical theory of chronological propositions can be developed in a systematic, formal way. It may be postulated that the operator **R** is to be governed by the following rules:

(T1) The negation of a statement p is realized at a given time if and only if it is not the case that the statement is realized at that time; *i.e.,* $\mathbf{R}_t(\sim p) \equiv \sim \mathbf{R}_t(p)$, in which $\equiv$ signifies equivalence and means "holds if and only if."

(T2) A conjunction of two statements is realized at a given time if and only if each of these two statements is realized at that time: $\mathbf{R}_t(p \cdot q) \equiv [\mathbf{R}_t(p) \cdot \mathbf{R}_t(q)]$. Example: "John and Jane will be at the railroad station at 10 AM—$\mathbf{R}_t(p \cdot q)$—if and only if John is at the station at 10 AM—$\mathbf{R}_t(p)$—and Jane is at the station at 10 AM—$\mathbf{R}_t(q)$."

If a statement is realized universally—*i.e.,* at any and every time whatsoever—it can then be expressed more simply as being true without any temporal qualifications; hence, the rule:

(T3) If for every time t the statement p is realized (**R**), then p obtains unqualifiedly; *i.e.,* $(\forall t)\mathbf{R}_t(p) \supset p$, in which $\forall$ is the universal quantifier (meaning "for all . . .").

If two times are involved, however (as in 1 and 2 above), then the left-hand term in rule T3 can be expressed within the second time frame as "It will be the case t' from now that, for every time t $(\forall t)$, it will be the case t from the first now that p"; *i.e.,* $\mathbf{R}_{t'}[(\forall t)\mathbf{R}_t(p)]$. It is an algebraic rule, however, that an $\mathbf{R}_t$ operator can be moved to the right past an irrelevant quantifier; hence,

$$\mathbf{R}_{t'}[(\forall_t)\mathbf{R}_t(p)] \equiv (\forall_t)\{\mathbf{R}_{t'}[\mathbf{R}_t(p)]\};$$

and, correspondingly, with the existential quantifier $\exists$: "It will be the case τ from now that there exists a time t $(\exists t)$ such that p will be realized at t" is equivalent to saying "There exists a time t such that it will be the case τ from now that p will be realized t from the first now" (in which τ is a second time); *i.e.,*

(T4) $\quad \mathbf{R}_\tau[(\exists_t)\mathbf{R}_t(p)] \equiv (\exists_t)\{\mathbf{R}_\tau[\mathbf{R}_t(p)]\}.$

It is notable that the left-hand side of this equivalence is itself equivalent with $(\exists t)\mathbf{R}_t(p)$ since what follows the initial $\mathbf{R}_\tau$ is a chronologically definite statement.

Finally, there are—as shown above—two distinct ways of construing iterations of the $\mathbf{R}_t$ operator, which provides a choice between the rules:

(T5-I) $\quad \mathbf{R}_\tau[\mathbf{R}_t(p)] \equiv \mathbf{R}_t(p)$
(T5-II) $\quad \mathbf{R}_\tau[\mathbf{R}_t(p)] \equiv \mathbf{R}_{\tau+t}(p)$.

Taking these rules as a starting point, two alternative axiomatic theories are generated for the logic of the operation of chronological realization.

Rules and modalities

Apart from strictly technical results establishing the formal relationships of the various systems of chronological logic to one another, the most interesting findings regarding the systems of tense logic relate to the theory of temporal modalities. The most striking finding here concerns the logical structure of the system of modalities, be it Megarian or Stoic:

Megarian	possibly p:	$(\exists t)\mathbf{R}_t(p)$
	necessarily p:	$(\forall t)\mathbf{R}_t(p)$
Stoic	possibly p:	$(\exists t)[\mathbf{F}(t) \cdot \mathbf{R}_t(p)]$
	necessarily p:	$(\forall t)[\mathbf{F}(t) \supset \mathbf{R}_t(p)]$

in which $\mathbf{F}(t)$ signifies "t is future." It has been shown that the forms, or structures, of both of these systems of tem-

poral modalities are given by the aforementioned system **S5** of C.I. Lewis.

Leśniewski's encounter with the paradoxes

Mereology. The founder of mereology was the Polish logician Stanisław Leśniewski. As noted earlier, Leśniewski was much exercised about Russell's paradox of the class of all classes not elements of themselves—if this class is a member of itself, then it is not; and if it is not, then it is. (Example: "This barber shaves everyone in town who does not shave himself." Does the barber then shave himself? If he does, he does not; if he does not, he does.)

Basic concepts and definitions. The paradox results, Leśniewski argued, from a failure to distinguish the distributive and the collective interpretations of class expressions. The statement "x is an element of the class of X's" is correspondingly equivocal. When its key terms (element of, class of) are used distributively, it means simply that x is an X. But if these terms are used collectively, it means that x is a part (proper or improper) of the whole consisting of the X's; *i.e.,* that x is a part of the object that meets the following two conditions: (1) that every x is a part of it; and (2) that every part of it has a common part with some x. On either construction of class membership, one of the inferences essential to the derivation of Russell's paradox is blocked.

Leśniewski presented his theory of the collective interpretation of class expressions in a paper published in 1916. Eschewing symbolization, he formulated his theorems and their proofs in ordinary language. Later he sought to formalize the theory by embedding it within a broader body of logical theory. This theory comprised two parts, viz., protothetic, a logic of propositions (not analyzed into their parts); and ontology, which contains counterparts to the predicational logic (of subjects and predicates), including the calculus of relations and the theory of identity. On his own approach, mereology was developed as an extension of ontology and protothetic; but the practice of most later writers has been to develop as a counterpart to mereology a theory of parts and wholes that is simply an extension of the more familiar machinery of quantificational logic employing $\exists$ and $\forall$. This is the course adopted here.

An undefined relation **Pt** serves as the basis for an axiomatic theory of the part relation. This relation is operative with respect to the items of some domain **D,** over which the variables $\alpha, \beta, \gamma, \ldots$ (alpha, beta, gamma, etc.) are assumed to range. Thus, $\alpha\mathbf{Pt}\beta$ is to be read "alpha is a part of beta"—with "part" taken in the wider sense in which the whole counts as part of itself. Two definitions are basic:

Disjointness and summation

1. "α is disjoint from β"; *i.e.,* $\alpha \mid \beta$ is defined as obtaining when "there exists no item γ such that γ is a part of α and γ is a part of β"; *i.e.,* $\sim(\exists\gamma)(\gamma\mathbf{Pt}\alpha \cdot \gamma\mathbf{Pt}\beta)$. Example: "The transmission (α) is disjoint from the motor (β) if there exists no machine part (γ) such that it is a part of the transmission and also a part of the motor."

2. "S has the sum of (or sums to) α"; *i.e.,* $S\,\Sigma\alpha$ is defined as obtaining when "for every γ, this γ is disjoint from α if and only if, for every β, to be a member of S is to be disjoint from γ"; *i.e.,*

$$(\forall\gamma)[\gamma \mid \alpha \equiv (\forall\beta)(\beta \in S \supset \beta \mid \gamma)].$$

$S\,\Sigma\alpha$ thus obtains whenever everything disjoint from α is disjoint from every S-element (β) as well, and conversely. Example: "A given group of buildings (S) comprises (Σ) Oxford University (α) when, for every room in the world (γ)—office, classroom, etc.—this room is disjoint from the university if and only if, in the case of each building (β), for it to be a member ($\in$) of the group that comprises the university (S) it must not have this room as a part ($\beta \mid \gamma$)."

Axiomatization of mereology. A comprehensive theory of parts and wholes can now be built up from three axioms:

The first axiom expresses the fact that "For every α and every β, if α is a part of β and β is a part of α, then α and β must be one and the same item"; *i.e.,*

$$(\forall\alpha)(\forall\beta)(\alpha\mathbf{Pt}\beta \cdot \beta\mathbf{Pt}\alpha \supset \alpha = \beta);$$

hence, the axiom:

(A1) Items that are parts of one another are identical. (This is analogous to the principle in number theory that given any pair of natural numbers n and m, if n is less than or equal to m and m is less than or equal to n, then n is equal to m.)

The second axiom expresses the fact that "For every α and every β, α is a part of β if and only if, for every γ, if this γ is disjoint from β it is then disjoint from α as well"; *i.e.,*

$$(\forall\alpha)(\forall\beta)[\alpha\mathbf{Pt}\beta \equiv (\forall\gamma)(\gamma \mid \beta \supset \gamma \mid \alpha)];$$

hence, the axiom:

(A2) One item is part of another only if every item disjoint from the second is also disjoint from the first. (Example: "This grain of sand (α) is a part of this pile of sand (β) just in case whatever is disjoint from the pile is disjoint from the grain.")

The third axiom expresses the fact that "If there exists an α that is a member of a nonempty set of items S, then there also exists a β that is the sum of this set"; *i.e.,*

$$(\exists\alpha)(\alpha \in S) \supset (\exists\beta)S\Sigma\beta;$$

hence, the axiom:

(A3) Every nonempty set has a sum. (If one considers the set of grains of sand in a given pile of sand, since such a set is nonempty, it will have a sum—namely, the pile of sand itself.)

Deduced theorems

Several theorems follow from these axioms:

The first states that "For every α, α is a part of α"; *i.e.,*

$$(\forall\alpha)\ \alpha\mathbf{Pt}\alpha;$$

hence, the theorem:

(T1) Every item is part of itself. (Example: "Not only are the pages parts of a book [α], but the book is a part of the book.")

The second theorem states that "For every α, for every β, and for every γ, if α is a part of β, and β is a part of γ, then α is a part of γ"; *i.e.,*

$$(\forall\alpha)(\forall\beta)(\forall\gamma)[(\alpha\mathbf{Pt}\beta \cdot \beta\mathbf{Pt}\gamma) \supset \alpha\mathbf{Pt}\gamma];$$

hence, the theorem:

(T2) The **Pt**-relation is transitive.

The third theorem states that "For every α, for every β, and for every γ, if γ is a part of α only when it is also a part of β then α is identical with β";*i.e.,*

$$(\forall\alpha)(\forall\beta)(\forall\gamma)[(\gamma\mathbf{Pt}\alpha \equiv \gamma\mathbf{Pt}\beta) \supset \alpha = \beta];$$

hence, the theorem:

(T3) Any item is completely determined by its parts; items are identical when they have the same parts in common.

The fourth theorem states that "For every α and every β, there exists a γ that is the sum of α and β";*i.e.,*

$$(\forall\alpha)(\forall\beta)(\exists\gamma)(\{\alpha,\beta\}\Sigma\gamma);$$

hence, the theorem:

(T4) Any two items whatsoever may be summed up.

In this form as a formal theory of the part relation, the history of mereology can be dated from some drafts and essays of Liebniz prepared in the late 1690s.

HYPOTHETICAL REASONING AND COUNTERFACTUAL CONDITIONALS

A simply conditional, or "if," statement asserts a strictly formal relationship between antecedent ("if" clause) and consequent ("then" clause) ("If p, then q"), without any reference to the status of the antecedent. The knowledge status of this antecedent, however, may be problematic (unknown), or known-to-be-true, or known-to-be-false. In these three cases, one obtains, respectively, the problematic conditional ("Should it be the case that p—which it may or may not be—then q"); the factual conditional ("Since p, then q"); and the counterfactual conditional ("If it were the case that p—which it is not—then q"). Counterfactual conditionals have a special importance in the area of thought experiments in history as well as elsewhere.

Material implication $p \supset q$, construed simply as the truth-functional "either not-p or q," is clearly not suited to represent counterfactual conditionals because any material implication with a false antecedent is true: when p is false, then $p \supset q$ and $p \supset \sim q$ are both true, regardless of what one may choose to put in place of q. But even when a stronger mode of implication is invoked, such as strict implication or its cognates, the problem of auxiliary hypotheses (soon to be explained) would still remain.

Hiż's analysis

It seems most natural to view a counterfactual conditional in the light of an inference to be drawn from the contrary-to-fact thesis represented by its antecedent. Thus, "If this rubber band were made of copper, then it would conduct electricity" would be construed as an incomplete presentation of the argument resulting from its expansion into:

Assumption: "This rubber band is made of copper";
Known fact: "Everything made of copper conducts electricity";
Conclusion: "This rubber band conducts electricity."

On the analysis, the conclusion (= the consequent of the counterfactual) appears as a deductive consequence of the assumption (= the antecedent of the counterfactual). This truncated-argument analysis of counterfactuals is a contribution, in essence, of a Polish linguistic theorist, Henry Hiż (born 1917). On Hiż's analysis, counterfactual conditionals are properly to be understood as metalinguistic; *i.e.,* as making statements about statements. Specifically, "If A were so, then B would be so" is to be construed in the context of a given system of statements S, saying that when A is adjoined as a supplemental premise to S, then B follows. This approach has been endorsed by Roderick Chisholm of Brown University, an important writer in applied logic, and has been put forward by many recent writers, most of whom incline to take S, as above, to include all or part of the corpus of scientific laws.

Evaluation through hypotheses, facts, and laws

The approach warrants a closer scrutiny. On fuller analysis, the following situation, with a considerably enlarged group of auxiliary hypotheses, comes into focus:

Known facts:
1. "This band is made of rubber";
2. "This band is not made of copper";
3. "This band does not conduct electricity";
4. "Things made of rubber do not conduct electricity";
5. "Things made of copper do conduct electricity";

Assumption: Not-(2); *i.e.*, "This band is made of copper."

When this assumption is introduced within the framework of known facts, a contradiction obviously ensues. How can this situation be repaired? Clearly, the logician must begin by dropping items 1 and 2 and replacing them with their negations—the assumption itself so instructs him. But a contradiction still remains. The following alternatives are open:

Alternative 1:	Retain: 3,4.
	Reject: 1,2,5.
Alternative 2:	Retain: 4,5.
	Reject: 1,2,3.

That is, the analyst actually has a choice between rejecting 3 in favour of 5 or 5 in favour of 3, resulting in the following conditionals:

1. "If this rubber band were made of copper, then it would conduct electricity" (since copper conducts electricity), or
2. "If this rubber band were made of copper, then copper would not (always) conduct electricity" (since this band does not conduct electricity).

If the first conditional seems more natural than the second, this is owing to the fact that, in the face of the counterfactual hypothesis at issue, the first invites the sacrifice of a particular fact (viz., that the band does not conduct electricity) in favour of a general law (viz., that copper conducts electricity), whereas the second counterfactual would have sacrificed a law to a purely hypothetical fact. On this view, there is a fundamental epistemological difference between actual and hypothetical situations: in actual cases one makes laws give way to facts, but in hypothetical cases one makes the facts yield to laws.

In more complex cases, however, the fact/law distinction may not help matters. For example, assume a group of three laws L_1, L_2, L_3, where $\sim L_1$ is inconsistent with the conjunction of L_2 and L_3. If asked to hypothesize the denial of L_1—so that the "fact" that one is opposing is itself a law—then what remains is a choice between laws; the distinction between facts and laws does not resolve the issue, and some more sophisticated mechanism for a preferential choice among laws is necessary. (N.R.)

BIBLIOGRAPHY

History of logic: CARL VON PRANTL, *Geschichte der Logik im Abendlande,* 4 vol. (1855–70, reprinted 1927 and 1955), notorious for misleading and outdated evaluations, but still useful as a vast collection of quotations; HEINRICH SCHOLZ, *Geschichte der Logik* (1931; Eng. trans., *The Concise History of Logic,* 1961); I.M. BOCHENSKI, *Formale Logik* (1956; rev. Eng. trans., *A History of Formal Logic,* 1961); WILLIAM and MARTHA KNEALE, *The Development of Logic* (1962). See also articles by various authors in *The Encyclopedia of Philosophy,* ed. by PAUL EDWARDS, 8 vol. (1967). Later monographs include D. GABBAY and F. GUENTHNER (eds.), *Elements of Classical Logic* (1983); RICHARD A. WRIGHT and KEN TOHINAKA, *Logical Thinking: An Integrated Introduction* (1984); F.H. GEORGE, *The Science of Philosophy* (1981); VINCENT E. BARRY, *Invitation to Critical Thinking* (1984).

Studies on logic in specific historical periods and cultures are I.M. BOCHENSKI, *Ancient Formal Logic* (1951); JAN LUKASIEWICZ, *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic,* 2nd ed. enlarged (1957); BENSON MATES, *Stoic Logic,* (1957); BENSON MATES, *Stoic Logic,* 2nd ed. (1961); PHILOTHEUS BOEHNER, *Medieval Logic* (1952); ERNEST A. MOODY, *Truth and Consequence in Medieval Logic* (1953); NICHOLAS RESCHER, *The Development of Arabic Logic* (1964); CLARENCE I. LEWIS, *A Survey of Symbolic Logic* (1918; corrected reprint with the omission of ch. 5 and 6, 1960); GEOFFREY T. KNEEBONE, *Mathematical Logic and the Foundations of Mathematics* (1963); and PETER H. NIDDITCH, *The Development of Mathematical Logic* (1962). See also JONATHAN LEAR, *Aristotle and Logical Theory* (1980); STANLEY ROSEN, *Plato's Sophist: The Drama of Original and Image* (1983); ALFONSO MAIERÙ (ed.), *English Logic in Italy in the 14th and 15th Centuries: Acts of the 5th European Symposium on Medieval Logic and Semantics, Rome, 1980* (1982); MAURICE A. FINOCCHIARO, *Galileo and the Art of Reasoning: Rhetorical Foundations of Logic and Scientific Method* (1980); WARREN E. STEINKRAUS and KENNETH I. SCHMITZ (eds.), *Art and Logic in Hegel's Philosophy* (1980); ROBERT S. COHEN and MARX W. WARTOFSKY (eds.), *Hegel and the Sciences* (1984); ERROL E. HARRIS, *An Interpretation of the Logic of Hegel* (1983).

Several useful bibliographical sources are available: *The Journal of Symbolic Logic* (issued quarterly since 1936), gives complete bibliographies of logic and of its history from before Leibniz; Bochenski's *History of Formal Logic* (*op. cit.*), gives a full bibliography including bibliography to the Indian and Chinese logics. See also GOPINATH KAVIRAJ, *The History and Bibliography of Nyāya-Vaiśesika* (1982), a bibliography of Indian logical writings.

(Cz.L.)

Logic systems: Formal logic is covered in ALFRED TARSKI, *Elementy logiki matematycznej,* 2nd ed. (1958; Eng. trans., *Introduction to Logic and to the Methodology of Deductive Sciences,* 3rd ed. rev., 1965); W.V.O. QUINE, *Elementary Logic,* rev. ed. (1965); ALICE AMBROSE and MORRIS LAZEROWITZ, *Fundamentals of Symbolic Logic* (1948); IRVING M. COPI, *Symbolic Logic,* 3rd ed. (1967); JOHN A. FARIS, *Truth-Functional Logic* (1962), propositional calculus only, and *Quantification Theory* (1964), lower predicate calculus only; E.J. LEMMON, *Beginning Logic* (1965); GEORGE E. HUGHES and D.G. LONDEY, *The Elements of Formal Logic* (1965). All of these, except where noted, contain introductions to both the propositional and lower predicate calculi. The natural deduction rules given in this article are derived from the book by Lemmon. Also introductory, but more advanced in character, are: W.V.O. QUINE, *Methods of Logic* (1950); ARTHUR N. PRIOR, *Formal Logic,* 2nd ed. (1962); RUDOLF CARNAP, *Introduction to Symbolic Logic and Its Applications* (1958); and PATRICK C. SUPPES, *Introduction to Logic* (1957). The most influential work in the field is ALFRED NORTH WHITEHEAD and BERTRAND RUSSELL, *Principia Mathematica,* 3 vol. (1910–13). In *Principia Mathematica to* *56 (1962), those parts of the original that are of most concern to logicians as distinct from mathematicians are reprinted. Another important and influential work is DAVID HILBERT and W. ACKERMANN, *Grundzüge der theoretischen Logik,* 2nd ed. (1938; Eng. trans., *Principles of Mathematical Logic,* 1950). Probably the most definitive account of the propositional and lower predicate calculi is presented in ALONZO CHURCH, *Introduction to Mathematical Logic,* vol. 1 (1956). For information on syllogistic, see JOHN N. KEYNES, *Studies and Exercises in Formal Logic* (1906), for a comprehensive traditional account; and JAN LUKASIEWICZ, *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic,* 2nd enl. ed. (1957), and OTTO A. BIRD, *Syllogistic and Its Extensions* (1964), for modern treatments. For modal logic, ROBERT FEYS, *Modal Logics* (1965), contains a comprehensive survey of axiomatic systems; GEORGE E. HUGHES and M.J. CRESSWELL, *An Introduction to Modal Logic* (1968), is a textbook and general survey. For brief introductions to set theory, see P.R. HALMOS, *Naive Set Theory* (1960), and E.J. LEMMON, *Introduction to Axiomatic Set Theory* (1969); fuller accounts are given in PATRICK C. SUPPES,

Axiomatic Set Theory (1960); and W.V.O. QUINE, *Set Theory and Its Logic* (1963). For an advanced treatment, see A.A. FRAENKEL, *Abstract Set Theory,* 3rd ed. (1965). For the logical foundations of mathematics (number theory and recursive functions), see ELLIOTT MENDELSON, *Introduction to Mathematical Logic* (1964); and J.W. ROBBIN, *Mathematical Logic: A First Course* (1969), though neither is easy reading. A standard advanced work in this field is S.C. KLEENE, *Introduction to Metamathematics* (1952). Later comprehensive surveys include MICHAEL T. CARLSEN-JONES, *Introduction to Logic* (1983); VIRGINIA KLENK, *Understanding Symbolic Logic* (1983); BARRY SMITH (ed.), *Parts and Moments: Studies in Logic and Formal Ontology* (1982); H.-D. EBBINGHAUS, J. FLUM, and W. THOMAS, *Mathematical Logic* (1984; trans. from the German); RUDY RUCKER, *Infinity and the Mind* (1982); FRED SOMMERS, *The Logic of Natural Language* (1982).

The various elements of metalogic are treated in RUDOLF CARNAP, *Introduction to Semantics* (1942); ALFRED TARSKI, *Undecidable Theories* (1953); KURT SCHUETTE, *Beweistheorie* (1960) and *Vollständige Systeme modaler und intuitionistischer Logik* (1968); HAO WANG, *A Survey of Mathematical Logic* (1962); MICHAEL MORLEY, "Partitions and Models," *Proc. Summer School in Logic* (1968); PAUL J. COHEN, *Set Theory and the Continuum Hypothesis* (1966); JOSEPH B. SHOENFIELD, *Mathematical Logic* (1967); and J.L. BELL and A.B. SLOMSON, *Models and Ultraproducts: An Introduction* (1969). See also HOWARD POSPESEL, *Introduction to Logic: Propositional Logic* (1984); A.C. GRAYLING, *An Introduction to Philosophical Logic* (1982); HUGUES LEBLANC, *Existence, Truth, and Provability* (1982); DUMITRU CHISE and ANGELA BOTEZ (eds.), *Dialectics, System, Science* (1981); ROBERT S. COHEN and MARX W. WARTOFSKY (eds.), *Language, Logic, and Method* (1983); DAVID PEARCE and VEIKKO RANTALA, *On a New Approach to Metascience* (1981); ALFRED TARSKI, *Logic, Semantics, Metamathematics,* 2nd ed. by JOHN CORCORAN (1983).

References on applied logic include ARISTOTLE, *De Sophisticis Elenchis* (Eng. trans., *Aristotle on Fallacies,* 1866); JEREMY BENTHAM, *The Book of Fallacies,* vol. 2 of his *Works,* ed. by J. BOWRING (1843); J.S. MILL, bk. 5 of *A System of Logic* (1843); R.H. THOULESS, *Straight and Crooked Thinking* (1930); C.L. HAMBLIN, *Fallacies* (1970), the only monographic study of the subject. K.J.J. HINTIKKA, *Knowledge and Belief* (1962), a technically sophisticated approach; N. RESCHER, *Topics in Philosophical Logic* (1968); R.M. CHISHOLM, "The Logic of Knowing," *Journal of Philosophy,* 60:773–795 (1963), *The Theory of Knowledge* (1966), basic but authoritative; N.D. BELNAP, JR., *An Analysis of Questions* (1963); D. HARRAH, *Communication: A Logical Model* (1963); L. AQVIST, *A New Approach to the Logical Theory of Interrogatives,* pt. 1 (1965); G.H. VON WRIGHT, *The Logic of Preference* (1963), the best introduction to the subject; S. HALLDEN, *On the Logic of "Better"* (1957); N. RESCHER, *Introduction to Value Theory* (1969); R.M. HARE, *The Language of Morals* (1952), a modern classic; N. RESCHER, The *Logic of Commands* (1966), with an extensive bibliography; G.H. VON WRIGHT, *Norm and Action* (1963); B.F. CHELLAS, *The Logical Form of Imperatives* (1969); E. MALLY, *Grundgesetze des Sollens: Elemente der Logik des Willens* (1926); G.H. VON WRIGHT, "Deontic Logic," *Mind,* 60:1–15 (1951); A.R. ANDERSON, "The Formal Analysis of Normative Systems," in N. RESCHER (ed.), *The Logic of Decision and Action* (1966), comprehensive and monographic in scope; F.B. FITCH, "Natural Deduction Rules for Obligation," *American Philosophical Quarterly,* 3:27–38 (1966); A. TARSKI, "Appendix E to J.H. Woodger," *Axiomatic Method in Biology* (1937); E.C. LUSCHEI, *The Logical Systems of Leśniewski* (1962); E.S. LEONARD and N. GOODMAN, "The Calculus of Individuals and its Uses," *Journal of Symbolic Logic,* 5:45–55 (1940); N. GOODMAN, "The Problem of Counterfactual Conditionals," *Journal of Philosophy,* 44:113–128 (1947); G.H. VON WRIGHT, *Logical Studies* (1957); and N. RESCHER, *Hypothetical Reasoning* (1964), a comprehensive study with further references to the literature. See also A.A. ZINOVIEV, *Logical Physics* (1983; trans. from the Russian); THOMAS M. OLSHEWSKY, *Good Reasons and Persuasive Force: A Pragmatic Approach* (1983); ULRICH KLUG, *Juristische Logic,* 4th ed. (1982); J.F.A.K. VAN BENTHEM, *The Logic of Time: A Model-Theoretic Investigation into the Varieties of Temporal Onthology and Temporal Discource* (1983); E.M. BARTH and E.C.W. KRABBE, *From Axiom to Dialogue: A Philosophical Study of Logics and Argumentation* (1982); JOHN WOODS, *Argument, the Logic of the Fallacies* (1982); DOUGLAS N. WALTON, *Logical Dialogue: Games and Fallacies* (1984); JOHN ERIC NOLT, *Informal Logic: Possible Worlds and Imagination* (1984).

(G.E.H./H.Wa./N.R./Ed.)

London

The capital city of the United Kingdom and once the industrial, commercial, and political centre of a vast empire, London lies astride the River Thames in southeastern England, 40 miles (65 kilometres) from its estuary on the North Sea. Urban tensions are moderated by a traditional sense of civic responsibility and by the relatively human proportions of a cityscape rich in greenery and open spaces. Although, like most other very large cities of the world, London has failed to come to terms with the late 20th century, it has at least managed to conserve many of the virtues of earlier centuries.

On a map, the outline of Greater London is a ragged oval, Central London being an irregular area in the middle, about a sixth of the whole, and containing the historic core region (now associated with financial institutions) known as the City. Aerial views reveal an increasing number of skyscrapers each year, although there are still relatively few for a city of such size. Houses preponderate, most built originally for a single family but nowadays often divided into flats (apartments) or maisonettes (flats that are not all on the same floor). Despite the great density of population, there is considerable open space. The River Thames itself is one of the largest continuous open spaces in the area.

Many changes have been made to alter the appearance of London and to affect city living. The 25 years following World War II constituted the period of dissolution of empire and withdrawal from world power; the point of view and the sense of national mission that had been shaped over three centuries suddenly no longer pertained. Once the war damage had been repaired (a massive enterprise that included the erection of low-rent housing groups), there ensued a terrible sense of battle fatigue. The lights had gone on again, as had been promised in wartime song and speech, but the hopes of a bright new world were stupefyingly frustrated. The famous British "muddling through" was supposedly replaced by scientific planning, and "talented amateurism" was ostensibly supplanted by technocratic professionalism; nevertheless, error and fumbling persisted. The mid-20th century was a period of bewilderment, imitation, and vigorous half measures, but by the 1970s a new period had begun, one that was marked by a reaffirmation of London's ancient virtues of self-confidence, imagination, and courage.

The article is divided into the following sections:

Physical and human geography 291
- The landscape 291
  - Site
  - Climate
  - Plant and animal life
- Traditional neighbourhoods 292
  - The Thames and its environs
  - The City of London
  - The City of Westminster
  - Shopping districts and markets
  - The East End
  - South London
  - London's squares and parks
- The people 303
- Administrative conditions 303

History 303
- The early period 303
  - Foundation and early settlement
  - Medieval London
  - Tudor London
  - 17th-century London
- Evolution of the modern city 304
  - Organization, innovation, and reform
  - Reconstruction after World War II

Bibliography 305

Physical and human geography

THE LANDSCAPE

Site. The oval chalk basin through which the Thames meanders from west to east is filled with great thicknesses of younger sediments, including solid rock from the Eocene epoch (57.8 million to 36.6 million years ago) of the Tertiary period, sands, clays, terraces of pebble gravels, and Thames alluvium. Central London has up to 20 feet (6 metres) of "made ground" accumulated in historic times. The earliest settlements were built on the well-watered gravel hills, and the clay produced bricks for building the city for many centuries.

Climate. The climate of the basin is typical of the eastern edge of the British Isles. St. James's Park, in the middle of Central London, receives 23½ inches (597 millimetres) of rain a year, but, over the rest of the basin, the rainfall is irregular from year to year, with extremes of 70 to 130 percent of the average. Temperatures are mild (37.4°–72.5° F [3°–22.5° C] January–July mean) and vary little, although it is warmer in the city than in the suburbs. Since the passage of the Clean Air Act of 1956, there is as much sunshine in the city as in the nearby countryside: average visibility in winter has increased from one mile to four miles, and the amount of smoke in the air has been reduced to between 33 and 25 percent of that measured in 1964. Since the act went into effect, there has been no smog in London, and the celebrated pea-soup fogs have become much rarer.

The Clean Air Act

Plant and animal life. London is of special interest to naturalists for the plants and animals that thrive there in close association with humans.

In the centre, wild plants are found in the parks, squares, and private gardens, on building sites, railway banks, neglected gutters, and broken walls. Many of the most common ones have wind-borne fruits, so that they quickly colonize any soil available. Oxford ragwort, rosebay willow herb, bracken, the shrub *Buddleia davidii* (butterfly bush), and Canadian fleabane were the five most abundant plants growing on bombed sites during and after World War II, when 269 species were found in the "square mile" of the City.

In the suburbs the number of species is much greater. Wimbledon, Mitcham, and Barnes commons, Hampstead Heath, and Epping Forest have a good representation of wild plants. Within 20 miles of St. Paul's Cathedral, 1,835 species of flowering plants and ferns, 221 mosses, and 68 liverworts have been recorded since 1900.

Four kinds of birds have learned to live in great numbers in close association with humans. Feral domestic pigeons have been a feature of London birdlife for at least 600 years and are familiar to visitors to Trafalgar Square, St. Paul's Cathedral, Victoria Station, and other places. House sparrows abound in Central London, and whiteheaded gulls have been regular winter visitors since 1895 and are seen in large numbers by the Thames and elsewhere. Immense numbers of starlings roost on buildings in Central London; the numbers are increased by immigrants from Europe in the winter. Ownership of the mute swans on the Thames is divided between the crown and the Vintners and Dyers livery companies, which have a royalty, or grant of ownership from the crown, of long standing. An ancient ceremony known as Swan-Upping takes place each July when the Vintners and Dyers mark the birds. The ceremony formerly started opposite Fishmonger's Hall, at London Bridge, but since the early 1970s, because of the decline in the swan population, it begins at Walton-on-Thames, in Surrey, and moves upstream, away from central London. The ravens at the Tower of London are semidomesticated birds introduced from Wales.

In inner London 160 species of birds are recorded, and

37 of these species have bred. St. James's Park and Kew Gardens have good collections of waterfowl. Within 20 miles of St. Paul's 245 species have been recorded since 1900 in a wide range of habitats. Of these species, 100 are believed to nest annually.

The brown rat is a serious pest, and the black rat is found in the upper parts of buildings over a large area, including the West End. The house mouse is common, hedgehogs are frequent in the suburbs, and several species of bats occur. A notable increase in the established fox population was reported in the early 1980s.

Butterflies are often seen, and some species, such as the migrant red admiral, congregate around the flowers of buddleia and Michaelmas daisies. Many moths occur in Central London, where the lime hawk, eyed hawk, and poplar hawk (*Laothoë populi*) are among the larger species. The trend toward industrial melanism (evolutionary darkening of species in response to a sooty environment) is less general in moths in London than in other industrial cities but is shown in a few species.

TRADITIONAL NEIGHBOURHOODS

The Thames and its environs. The central historical fact of London is the Thames. In the classic pattern, at the spot where a road crossed the stream, foreign invaders (the Claudian conquest of AD 43) displaced the natives and built a bridge and a city (Londinium). In Roman times the river was broader and shallower. Two small streams, the Walbrook and the Fleet, both of which still flow into the Thames but now through underground conduits, enabled the Romans to bring boats inside their fortified city. After the departure of the Romans, the foreign commerce they had established continued, even through the historically dim Anglo-Saxon period, and made the city rich and therefore powerful.

Two miles upstream at the Westminster river crossing, where there had been a small Roman settlement, a religious community in Saxon times, and a royal palace in the 11th century, the City of Westminster grew. Between the commercial city and the royal city the town houses of nobles and bishops were built, each with its own water gate, and eventually a road (the Strand) ran from Charing Cross to Temple Bar, one of the City of London's 10 gates. Well into the 18th century it was the river rather than the road that served as the main highway of transport and commerce.

All the bridges spanning the Thames from London to Southwark were wooden until, between 1176 and 1209, a single stone bridge was finally built. The tidal waters roared tumultuously through its 19 arches, and until 1831 "shooting the bridge" in a small boat was one of the thrills of London. A chapel was built on the bridge to St. Thomas à Becket, martyred in 1170. Soon after, shops lined both sides of the roadway between the fortified gates at either end, and then houses were built above the shops, 138 separate premises being recorded in 1358. In Queen Elizabeth I's time water mills were installed that added to the uproar. The houses were removed from the structure in 1760, the same year that the City gates were dismantled, but the bridge continued to serve for another 71 years, a total lifetime of 622 years. Its replacement was in its turn replaced piecemeal between 1969 and 1972. The picturesque Tower Bridge, one-half mile downstream, was built between 1886 and 1894 by Sir Horace Jones, who had previously designed Billingsgate, Smithfield, and Leadenhall markets. It has twin Gothic towers and a central draw-bridge and is a good deal more quaint and "Olde" English than London Bridge or the two other spans owned and maintained by the City: Blackfriars (originally constructed 1756–69) and Southwark (first built 1813–19).

The role of London Bridge

After the Norman Conquest the port's foreign trade increased, and embankment of the river on the seaward side of London Bridge began in the 12th century and was completed in the 14th, reclaiming from the water 42½ square miles (110 square kilometres) of marshland at Rotherhithe and Deptford on the south bank and the Isle of Dogs on the north. Still boggy centuries later, this land was easily excavated for dock building.

Further embankment of the river was not achieved until the 19th century, when at last London stopped dumping its wastes into the Thames and constructed main sewers. From 1864 to 1870 the mains were laid along the north bank behind stone retaining walls and roadways laid over them from Blackfriars Bridge to Westminster. About 37½ acres (15 hectares) of mudbank were reclaimed and converted into parkland (Victoria Embankment Gardens between Waterloo Bridge and the Hungerford Railway Bridge). In 1874 Chelsea Embankment, also laid over new sewer mains, was completed. The embankment system, interrupted briefly at the Houses of Parliament, is 4½ miles (seven kilometres) long. On the riverside opposite Parliament, Albert Embankment was built in 1869 as a flood-control installation, extended—for pedestrian traffic only—in front of County Hall (the headquarters of the

J. Alex Langley—Aspect Picture Library



Tower Bridge over the River Thames.

former Greater London Council, or GLC), downstream of Westminster Bridge, in 1910. The last of the embankments was made in 1951, when the South Bank cultural complex was started on a bombed site between County Hall and Waterloo Bridge.

Straightening the shoreline and jacketing the river with stone walls moved the tidal limit from London Bridge 19 miles upstream to Teddington Weir. The average high tide at London Bridge is rising at the rate of about three feet each century, as London sinks about one foot in the same period. High tide at London Bridge is four feet higher than it is downstream at Southend on the broad estuary 43 miles away. London's embankments were raised an additional 1½ feet after 1971 to guard against floods that might come when a wavelike surge mounts the river from the North Sea. A 1,706-foot (520-metre) movable flood barrier was also built at Silvertown, in the Woolwich Reach section of the Thames, eight miles downstream from London Bridge. It became operational in 1982.

Flood protection

Development of the docks. During the reign of Queen Elizabeth I, "legal quays" were established on the north bank of the Thames between London Bridge and the Tower as the only place where ships could land dutiable goods legally. The traffic soon grew too great for the quays, and in 1663 Parliament allowed the establishment of alternative "sufferance wharves" on both banks. Port activity doubled between 1700 and 1770, and, by the end of that time, the Upper Pool (that part of the river that stretches a little less than a mile below London Bridge) held as many as 1,775 ships in a space allocated to 600. Unloaded into the 3,500 lighters that carried goods from ship to wharf, cargoes sometimes remained caught in the maritime traffic jam for weeks at a time, subject to pilfering and to raids by river pirates. To enable ships to discharge directly into guarded quays, where goods could be stored in secure warehouses, the West India Dock was opened in 1802 at the northern end of the Isle of Dogs, a marshy river peninsula opposite Greenwich Hospital. In 1805 the London Dock opened in Wapping, and downstream from the Isle of Dogs the East India Dock was inaugurated in 1806. In 1807 the existing Greenland Dock (where whale blubber was rendered) and Howland Wet Dock in Rotherhithe on the south bank were combined, later to be enlarged as the Surrey Commercial Docks, covering 410 acres, a tract bigger than Hyde Park. The St. Katharine Docks were built under the lee of the Tower in 1828. Except for the West India, all of these and the dock at the entrance to Regent's Canal are closed. The Port of London operates from three large dock systems: West India, Millwall (1868), and Poplar, on the Isle of Dogs; Royal Victoria (1855), Royal Albert (1880), and King George V (1921), six miles farther downstream; and Tilbury Docks (1886), 26 miles from London Bridge. In 1909 the Port of London Authority (PLA) was created to take over ownership of all existing docks and control of the river and the port.

Waterfront development. Having closed down the upstream docks in the 1960s, riverfront properties covering 850 acres, the PLA decided to sell, providing a historic opportunity to remake a rich portion of London. In 1969 the GLC bought the first parcel, the St. Katharine Docks (25 acres, 10 of them water), for £60,000 an acre; three years later, it was worth £250,000 an acre. Some of the original majestic warehouses were retained, and a theatre, hotel, cinema, yachting marina, restaurants, and pubs were built as part of a village for 2,000 residents occupying houses, flats, and studios with extraordinary river views at a wide variety of rents. At the same time, the GLC began construction of Thamesmead, a development on the site of the 300-year-old Royal Arsenal of Woolwich, that was designed to be a model town of social and architectural balance. Simultaneously, in the disused Royal Victorian Victualling Yards, Deptford, the GLC brilliantly converted the interiors of the 1795 rum warehouses into 65 flats, a sailing school, and yachting centre.

Originality in urban renewal plans

The originality and habitability of these projects generated a new enthusiasm for long-ignored areas. Private developers bought and restored "forgotten" squares and terraces in waterfront neighbourhoods. The 12 miles of waterfront from Waterloo to Woolwich, formerly declining docks, factories, and decayed housing, became the subject for frantic bidding and fevered planning that ranged from the utopian to the crassly exploitative. The borough councils, the GLC, and those ministries that might be concerned pondered all the proposals that would eventually return a vast tract of the city to the living community.

The river on which all this excitement was concentrated was until the 1960s foul-smelling and almost black with pollutants. Since the late 1960s it has become cleaner each year, thanks to strenuous antipollution policing; fish are returning in increasing numbers, and birds, not seen for many years, are nesting along the less clamorous portions of the river.

The City of London. Colloquially called "The Square Mile," the City of London is actually slightly larger (677 acres), being about 1⅝ miles at its longest and ⅞ of a mile at its widest. The City is a marvelous mélange of ancient elegance and modern efficiency, of outdated practices and newfangled irritations, a place where for 1,000 years tradition has been honoured and innovation has been essential. Thus, while the messengers ("waiters") of the Bank of England are dressed in top hats and pink tailcoats and, as recently as 1973, a detachment of the Brigade of Guards reported nightly to repel mobs (a precautionary measure dating from the Gordon Riots of 1780), the bank employs computers and other contemporary aids to expedite its labours.

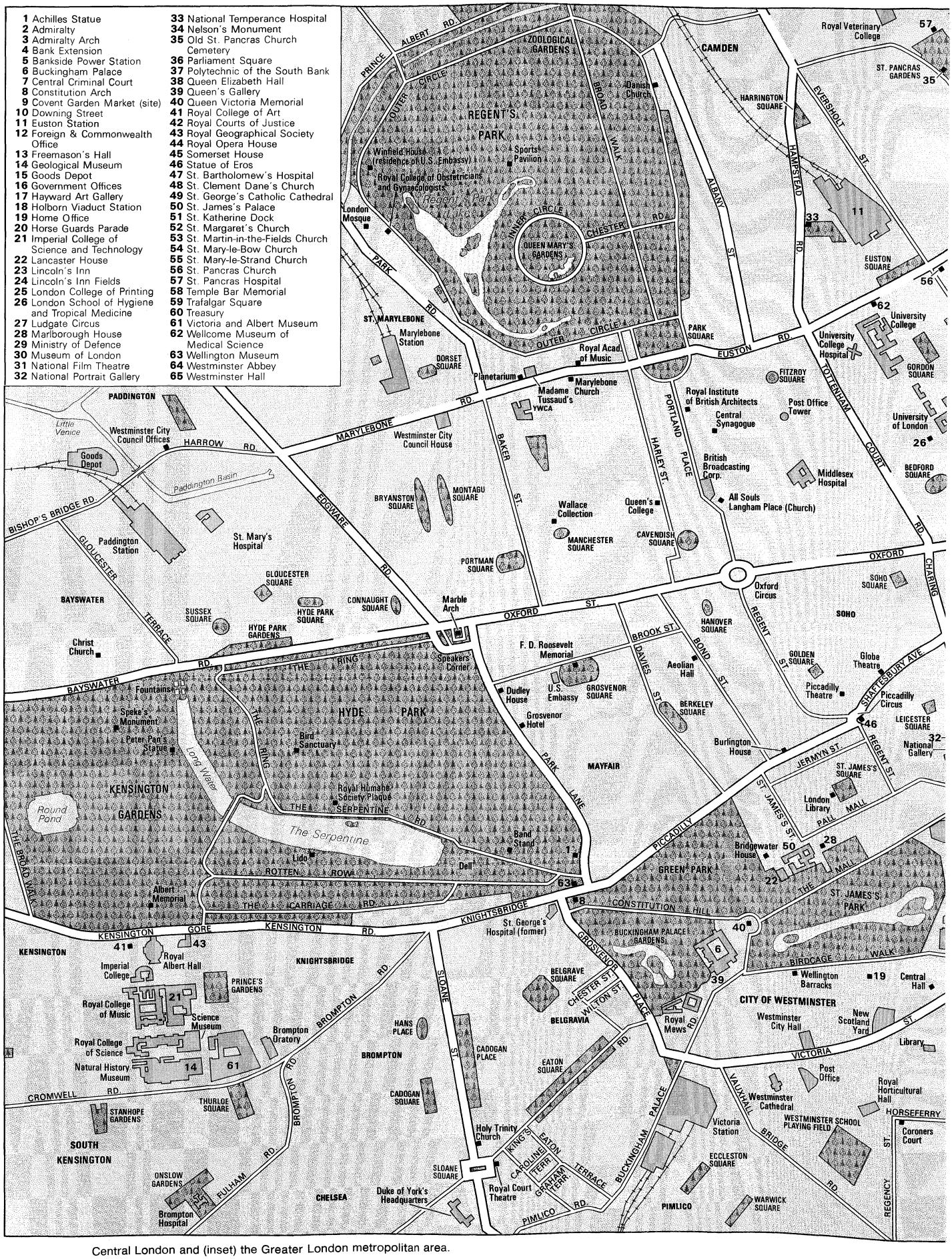
The City is one of the world's great financial centres despite the deliberate dismantling of the sterling bloc (countries whose currency and reserves were tied to the pound sterling). It also remains a pivotal point for world trade in commodities in spite of the fact that raw goods no longer originate in British possessions and that few of the commodities dealt with by the exchanges are actually physically handled at the Port of London.

The City's institutions are as varied as they are ancient. Five "wise men" set the world price of bullion in the opulent Gold Room of N.M. Rothschild & Sons, St. Swithin's Lane, at 10:30 each morning, but, before these gentlemen are out of bed, predawn buyers on the other side of the City eye hook-hung carcasses at Smithfield, the world's largest dressed-meat market. For 378 years, until 1981, the gentlemen of the Fishmongers Guild, their boots silvered with fish scales, exercised their immemorial functions down by the river at Billingsgate, London's fish market. In early 1982, however, the Billingsgate market was moved to new quarters at the West India Dock in adjoining Tower Hamlets Borough. Nearby, nurses begin to prepare patients for surgery at St. Bartholomew's ("Bart's"), London's first hospital (founded in 1123) and the place where, in the 17th century, William Harvey first demonstrated the circulation of the blood. Closer to St. Paul's Cathedral, the vans begin to deliver prisoners whose cases will be heard that day at the Old Bailey, as the Central Criminal Court is known, where most of Britain's sensational murder trials have been held.

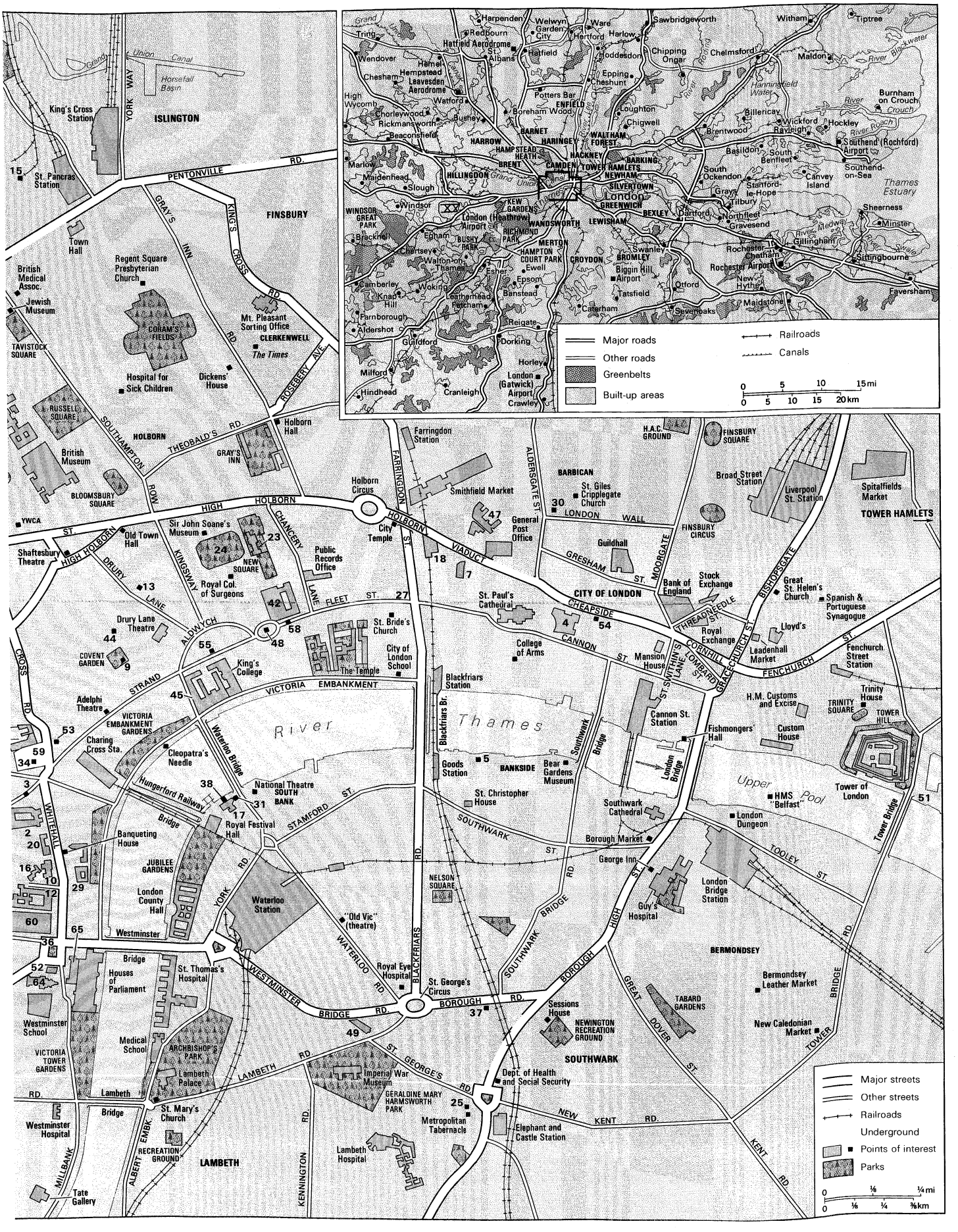
As the daylight broadens, the first commuters begin their daily invasion, descending from eight main railway stations and from buses and emerging from the underground railway. They stream across London Bridge and Waterloo Bridge, just west of the City, in particular, in what seems to be a solid mass of pedestrians. The bankers, the insurance underwriters from Lloyd's, the Stock Exchange members, and the commodity exchange brokers arrive a little later, some by underground and others in their cars.

The daily flood of commuters

The Barbican scheme. Since 1973 there have been some City men who arrive at their desks after a short stroll from their homes in Barbican, a City redevelopment on World War II bomb craters around the 16th-century church of St. Giles Cripplegate. The modern buildings accord with the hues of the old church stones and bricks, and the project is planned to bring inhabitants back to the City. There are private patios and terraces, an artificial lake, greenery, and subsurface service roads and parking facilities. In addition to offices, shops, and restaurants, there is the new building for the old City of London School for Girls (boys resolutely some distance away at Blackfriars on the site of Henry VII's Baynard's Castle, which stood



Central London and (inset) the Greater London metropolitan area.


Major roads
Other roads
Greenbelts
Built-up areas
Railroads
Canals
0 5 10 15mi
0 5 10 15 20km
Major streets
Other streets
Railroads
Underground
Points of interest
Parks
0 ⅛ ¼mi
0 ⅛ ¼ ⅜km
ISLINGTON
FINSBURY
HOLBORN
CLERKENWELL
BLOOMSBURY
LAMBETH
SOUTHWARK
BERMONDSEY
BARBICAN
CITY OF LONDON
TOWER HAMLETS
BANKSIDE
SOUTH BANK
TRINITY SQUARE
TOWER HILL
Thames
River
Upper Pool
King's Cross Station
St. Pancras Station
Town Hall
British Medical Assoc.
Jewish Museum
TAVISTOCK SQUARE
Regent Square Presbyterian Church
CORAM'S FIELDS
Hospital for Sick Children
Dickens' House
Mt. Pleasant Sorting Office
The Times
Holborn Hall
GRAY'S INN
RUSSELL SQUARE
British Museum
BLOOMSBURY SQUARE
YWCA
Old Town Hall
Shaftesbury Theatre
Sir John Soane's Museum
NEW SQUARE
Royal Col. of Surgeons
Public Records Office
Drury Lane Theatre
COVENT GARDEN
King's College
Adelphi Theatre
VICTORIA EMBANKMENT GARDENS
Charing Cross Sta.
Cleopatra's Needle
Waterloo Bridge
Hungerford Railway Bridge
Banqueting House
JUBILEE GARDENS
London County Hall
Waterloo Station
Westminster
Bridge Houses of Parliament
St. Thomas's Hospital
Westminster School
VICTORIA TOWER GARDENS
Medical School
ARCHBISHOP'S PARK
Lambeth Palace
Lambeth
Bridge
Westminster Hospital
St. Mary's Church
RECREATION GROUND
MILLBANK
Tate Gallery
ALBERT EMBK.
KENNINGTON RD.
Lambeth Hospital
Farringdon Station
Holborn Circus
City Temple
Smithfield Market
General Post Office
St. Giles Cripplegate Church
LONDON WALL
H.A.C. GROUND
FINSBURY SQUARE
FINSBURY CIRCUS
Broad Street Station
Liverpool St. Station
Spitalfields Market
Guildhall
Bank of England
Stock Exchange
Great St. Helen's Church
Spanish & Portuguese Synagogue
St. Paul's Cathedral
CHEAPSIDE
Royal Exchange
Lloyd's
College of Arms
Mansion House
Leadenhall Market
Fenchurch Street Station
Blackfriars Station
Blackfriars Br.
Cannon St. Station
H.M. Customs and Excise
Trinity House
Fishmongers' Hall
Custom House
Goods Station
Bear Gardens Museum
St. Christopher House
Southwark Bridge
London Bridge
Tower of London
HMS "Belfast"
Tower Bridge
Southwark Cathedral
London Dungeon
Borough Market
George Inn
London Bridge Station
Guy's Hospital
NELSON SQUARE
"Old Vic" (theatre)
Royal Eye Hospital
St. George's Circus
Sessions House
NEWINGTON RECREATION GROUND
TABARD GARDENS
Bermondsey Leather Market
New Caledonian Market
Dept. of Health and Social Security
Imperial War Museum
GERALDINE MARY HARMSWORTH PARK
Metropolitan Tabernacle
Elephant and Castle Station
Harpenden
Welwyn Garden City
Ware
Sawbridgeworth
Witham
Tiptree
Redbourn
Hatfield Aerodrome
St. Albans
Hatfield
Harlow
Hertford
Chipping Ongar
Chelmsford
Maldon
Wendover
Hemel Hempstead
Chesham
Leavesden Aerodrome
Watford
Epping
Cheshunt
Potters Bar
High Wycombe
Chorleywood
Rickmansworth
Bushey
Boreham Wood
ENFIELD
Loughton
Chigwell
Billericay
Burnham on Crouch
Wickford
Hockley
Rayleigh
Brentwood
Southend (Rochford) Airport
Basildon
South Benfleet
Southend-on-Sea
Canvey Island
Beaconsfield
HARROW
BARNET
HARINGEY
WALTHAM FOREST
HAMPSTEAD HEATH
HACKNEY
BARKING
TOWER HAMLETS
NEWHAM
Marlow
Maidenhead
Slough
BRENT
HILLINGDON
CAMDEN
SILVERTOWN
South Ockendon
Stanford-le-Hope
Grays
Tilbury
Thames Estuary
Windsor
WINDSOR GREAT PARK
KEW GARDENS
GREENWICH
BEXLEY
Dartford
Northfleet
Gravesend
Sheerness
Minster
London (Heathrow) Airport
WANDSWORTH
LEWISHAM
RICHMOND PARK
Bracknell
Egham
BUSHY PARK
Chertsey
Walton-on-Thames
HAMPTON COURT PARK
MERTON
CROYDON
BROMLEY
Swanley
Rochester
Chatham
Gillingham
Sittingbourne
Esher
Ewell
Epsom
Biggin Hill Airport
Rochester Airport
Camberley
Woking
Knaq Hill
Leatherhead
Banstead
Tatsfield
New Hythe
Otford
Faversham
Farnborough
Aldershot
Caterham
Sevenoaks
Maidstone
Reigate
Guildford
Dorking
Horley
Milford
Hindhead
Cranleigh
London (Gatwick) Airport
Crawley

at the river edge from 1487 to the Great Fire). Also at the Barbican is the Barbican Centre: an art gallery, three cinemas, a library, two exposition halls, quarters for the famed City-run Guildhall School of Music and Drama, two theatres for the Royal Shakespeare Company, and a 2,000-seat concert hall, the residence of the London Symphony Orchestra.

The hub of the City. The central spot in the City is an open space from which eight streets radiate. On the south side is the Mansion House, which was designed in the mid-18th century by George Dance as the lord mayor's residence, office, and court (he is first magistrate) and which is still so employed by each lord mayor during his one-year term. Clockwise from the Mansion House is Lombard Street, the traditional banking street; on each of its establishments hang the fanciful signboards contrived in earlier centuries. Between Cornhill and Threadneedle Street sits the third restoration of Sir Thomas Gresham's 1567 Royal Exchange. Across the street, the original 18th-century ground-floor screen hides the Bank of England headquarters, which was built in the period 1921 to 1937 to replace Sir John Soane's 1782–1802 masterpiece, some believe the worst loss to the nation ever caused by the bank. Across a narrow lane east of the bank rises the Stock Exchange (1968), and a short distance to the northeast stands the twice-restored Guildhall, the seat of the Corporation of London, that was originally built in 1425–45.

The Corporation of London, the guilds and livery companies

The Corporation of London operates without a charter of incorporation, the only one in the nation to do so. The voters elect common councilmen and aldermen (one of whom is chosen as lord mayor by the aldermen from two candidates nominated annually by the liverymen of the guilds). Voters are enfranchised by residence in the City or by tenancy or ownership of property of annual gross value of £10. Members of livery companies must be freemen of the City; freedom of the City, acquired through completion of guild apprenticeship, through inheritance or through purchase, was until the mid-19th century a prerequisite for trading or working at a craft in the City. There are now more than 90 livery companies, successors of religious or social fraternities founded in some cases as early as the 12th century.

Some guilds—the Apothecaries, Goldsmiths, Fishmongers, Spectacle Makers, for example—still exercise a controlling influence in their trades, especially in the maintenance of high standards of quality. A few livery companies conduct training schools; others provide scholarships and grants for research and education in schools and universities.

The Inns of Court. The Temple, with an entrance off Fleet Street, represents another segment of the City's power machinery. Originally the English headquarters of the Knights Templar, after dissolution of the order it eventually became the first of the Inns of Court, where barristers had their chambers and where they trained (and still do train) the coming generation of trial lawyers. Middle Temple, Inner Temple, Lincoln's Inn, and Gray's Inn, the latter two just north of the borders of the City, are oases of calm, civility, and continuity. The Inns of Court were rebuilt almost entirely in the 18th century, but the Temple retains its much-rebuilt, round Norman church with a 13th-century addition.

Just across Fleet Street, Chancery Lane begins, with the Royal Courts of Justice on the west and the Public Record Office to the east. The Law Courts, as the Royal Courts of Justice are usually called, were designed by G.E. Street in the 19th century, a chill, gloomy Victorian Gothic mass. It is there that the Lord Mayor's Show comes to an end on the second Saturday of November, when the new lord mayor alights from the state coach (a brewery lends the six horses) to pledge fealty to the monarch, represented by the lord chief justice. Until 1838 the show was waterborne, and the lord mayor went up river to Westminster in the state barge.

The City of Westminster. Once the court stopped following the sovereign whenever he moved and settled itself in Westminster, executive departments such as the Chancery and Exchequer were able to develop. By the 16th century Parliament became institutionalized. Peers of the realm and high dignitaries of the church found it expedient to maintain palaces close to the source of power.

The Strand

The Strand, a link between the City of London and the City of Westminster, became more thickly built upon as time went on, but in Anglo-Saxon times (probably by the 9th century) there had already been a parish church outside the Roman wall, St. Clement Dane's, referring to a period of intermittent Danish occupation during the 9th to 11th century. Its "twin" further west in the Strand, St. Mary-le-Strand, was first mentioned in 1147. Today both stand on tiny islands of land in the middle of the roadway, St. Clement's rebuilt around its 15th-century tower by Wren in 1680 and St. Mary's rebuilt (1714–17) by James Gibbs, in a handsome blend of Italian and English Baroque. St. Clement's was rebuilt again in 1958 after World War II bombing. Between the two churches, but back from the Strand, long stood the town houses of Elizabeth I's favourite, the ill-fated earl of Essex, and the earls of Arundel, whose descendant, the Duke of Norfolk, who still owns the land.

On the south side of the Strand is Somerset House, begun in 1776, possibly the first building in modern times specifically built for government offices. It was designed by Sir William Chambers and is most widely known for its archives of wills and, until the mid-1970s, birth and death records. Wings were added in the 19th century, and King's College (now part of London University) was installed in one in 1835. In 1970 two late 17th-century houses were removed from the Strand to permit construction of the college's laboratories.

North and west of this part of the Strand, a parallelogram about three-quarters of a mile long and not quite half a mile wide would enclose most of London's theatres and first-run cinemas. Until almost the middle of the 19th century, the Strand was one of London's smartest theatre, hotel, and shopping streets.

No amount of gracious building in this area can make good the loss of Robert and James Adam's magnificent 18th-century urban set piece, Adelphi. Red brick and Palladian, massive in total bulk but delicate in detail and proportion, it rose on arches above the riverbank in 1768 (completed 10 years later). In 1936 it was razed to clear space for more profitable development. In three narrow streets named for the brothers, nine of their houses survive, wistful reminders of a regrettably departed glory.

The City of Westminster is the wealthiest of any borough in the kingdom. It stretches along the Thames from the Temple to Chelsea Bridge and north through the titillations of Soho and the consultations of Harley Street to the Grand Union Canal. It includes some of the most desirable residential properties in the kingdom and even the shabbier parts are, by virtue of their position, enormously valuable.

The area contains Westminster Abbey (Anglican) and Westminster Cathedral (Roman Catholic), Buckingham, Westminster, and St. James's palaces, the principal government offices, the most important shopping districts of the country, New Scotland Yard, most of London's luxury hotels, Madame Tussaud's waxworks museum, and four museums of art, two at least of world rank. The National Gallery in Trafalgar Square has a superb collection of masterpieces of Western painting. The Tate Gallery, built in 1893–97 on the river near Vauxhall Bridge, has a large collection of British paintings, modern foreign works, and sculpture. It was named after its donor, Sir Henry Tate.

Shopping districts and markets. The most important retail shopping concentration in the United Kingdom is Oxford Street, where its first palatial department store building was constructed in 1909 by Gordon Selfridge, a "retired" Chicago retail merchant. The heavy shopping activity flows southward along Regent Street to Piccadilly Circus, turns right along Piccadilly, with northward branches along Sackville Street and Savile Row, where eminent tailors still make some of the world's finest men's clothing. Just alongside the Royal Academy is one of retailing's most charming and luxurious shopping areas, the Burlington Arcade. With two score of tiny shops under a glass-roofed promenade, it was built in 1815 at the same time that John Nash built his Royal Opera Arcade. Paral-

Oxford Street, Regent Street, and the Burlington Arcade

lel and a little farther west, Bond Street is still a magnet for lavish spenders from around the world. South of Piccadilly, and parallel to it, Jermyn Street is studded with long-established makers of costly shirts, hats, and other male accoutrements.

Several streets east of Piccadilly Circus, an area along and around Charing Cross Road, is one of the world's great gleaning grounds for secondhand books, a treasure hunt for the collector and a university for the curious.

Kensington High Street and Knightsbridge are two other heavily patronized shopping districts that attract clients from all over. Harrod's department store in Knightsbridge is the largest in western Europe and certainly the most famous. Kensington High Street is just west of Kensington Gardens, with several modish clothing and accessory shops, an antique hypermarket, and a department store.

About a mile to the east, the Knightsbridge shops cluster where Knightsbridge, Belgravia, and the grander parts of Chelsea convene haughtily south of Hyde Park. West along Brompton Road and south down Sloane Street the shops progress toward King's Road in Chelsea. Traced from St. James's Palace to Hampton Court Palace across farmland and through several hamlets, the King's Road was, until Victoria's accession, reserved to court traffic. Aside from a few antique shops and a large department store in Sloane Square, it was a typical London "village high street" until the 1960s, when it was invaded by fashion shops, private dancing and drinking clubs, and coffee bars. From the mid-19th century Chelsea had been a relatively unglamorous district, though with a few elegant old streets, the homes of some celebrated writers, the studios of noted painters, and a fringe of artsy-craftsy aspirants, but after World War II it became increasingly fashionable and expensive. The effect spread westward by the end of the 1960s, raising rents and the decibel count in neighbouring Fulham.

The character of Chelsea

There are still several outdoor markets in Central London, most of them selling only fruit and vegetables but some offering a very general assortment of merchandise. Two of the best known general markets are on Petticoat Lane (really Middlesex Street), in the East End, and on Portobello Road, off Notting Hill Gate.

The East End. Beyond the easternmost City gate down to the River Lea is London's East End. The neighbourhood names persist: Aldgate, Spitalfields, Whitechapel, Mile End, Bethnal Green, Wapping, Shadwell, Stepney, Limehouse, Poplar, and Isle of Dogs, but officially they comprise the borough of Tower Hamlets.

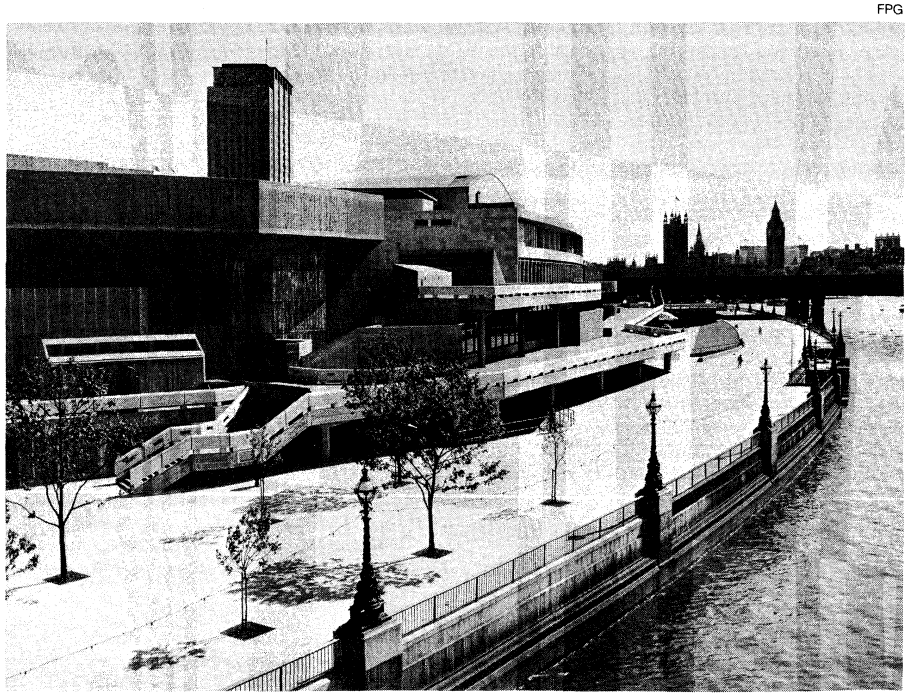
In the Middle Ages, the East End was part of vast Stepney Parish based on the Saxon church of St. Dunstan's, and, by the early 17th century, it was already a place to which the poor gravitated. With the development of the docks in the 19th century, offering casual employment, and the growth of the clothing and furniture industries, which battened on sweated labour, the increasing number of poor people competed for an intermittent pittance.

Among the overcrowded houses there was no drainage, and, despite the early foundation of hospitals and charities, the average age of death in Bethnal Green in 1840 was 16 years, 50 percent of all deaths among the labouring classes being children under five. Through the latter half of the 19th century, there were continuing waves of immigrant groups, to whose poverty the onus of racial, religious, and antiforeign prejudice was added. This was the London of Jack the Ripper, where life and gin were equally cheap.

The East End was the most savagely bombed and burned part of London in World War II. The rebuilding reflected far lower population densities, and many families accepted new homes and jobs in the New Towns started at this period; thus the population in 1960 was less than half that of 1901. Industry as well as population continued initially to drain away from Tower Hamlets, accompanied by a decline in the borough tax revenues. Plans for developing the docklands were intended to reverse this decline in the 1980s. The housing construction program continues, Spitalfields and Stepney–Poplar being two of the areas of concentration. Spitalfields, which has the borough's highest density of population, has been a receiving station for immigrants since the Huguenot silk weavers fled France at the end of the 17th century. The newest immigrants have been Indian and Pakistani, mingling with an earlier wave of West Indians and little settlements left over from the turn-of-the-century Irish and east European Jewish arrivals, spreading down into Whitechapel. The entire district, after consultation with the residents, is being remade. The Stepney–Poplar district, old Cockney territory, is to emerge as a "New Town." The new redevelopment of the dock areas has led to the private purchase and restoration of Georgian buildings by prosperous Londoners of a kind that last inhabited the district 200 years ago. A "green corridor" of open space currently runs from Victoria Park through Mile End Park, and there are plans for this to extend down to the river at Limehouse and Shadwell. In 1987 a light rapid transit railway was built that linked the docklands and the City.

South London. The inner London boroughs of Green-

FPG



Royal Festival Hall on the South Bank of the Thames.

wich, Lewisham, Southwark, Lambeth, and Wandsworth occupy the south bank of the Thames, but the term South Bank refers above all to the music and arts complex begun in 1951 for the Festival of Britain. It begins just downstream from London's County Hall. The first permanent building on the South Bank site was the Royal Festival Hall, which presents the world's great orchestras and ballet troupes in some 450 performances a year. The Queen Elizabeth Hall (1967) and its chamber-music auditorium, the Purcell Room, offer about 400 and 300 performances, respectively, each year. The Hayward Gallery, which opened in 1968, stages major art exhibitions and has three courts for displaying sculpture. The National Film Theatre, tucked into a corner, shows classic films from all nations. The National Theatre, which was lodged for years at the Old Vic Theatre (built 1816), behind Waterloo Station, was moved into its new home just east of Waterloo Bridge on the South Bank in the mid-1970s. Next to the National Theatre, on space originally designated for residential use, are the studios and offices of an independent television station. There is a riverside promenade for pedestrians that starts opposite the Houses of Parliament and continues along the South Bank almost to Blackfriars Bridge. From Blackfriars Bridge waterfront redevelopment was planned to stretch more than 10 miles down to Woolwich.

The South Bank Complex

Bankside, as the riverfront between Blackfriars and London bridges is called, is where Shakespeare's Globe Theatre and its rivals stood and where the palace of the bishop of Winchester neighboured the 13th-century Southwark Cathedral (largely rebuilt in the 19th century). Around this end of London Bridge, coaching inns, mentioned by Chaucer in the 14th century and by Shakespeare in the 17th, once clustered. One such inn, the George, rebuilt after the Southwark fire of 1667 and truncated by the railroad in 1889, continues to operate in an atmosphere of bustling bonhomie. The railway also displaced St. Thomas's Hospital, which had stood by the bridge since 1213. The hospital was removed to its present riverside site between Westminster and Lambeth bridges in 1871, and its medical school is just across the street from the much-restored 13th-century Lambeth Palace of the archbishop of Canterbury. Southwark still has a famed teaching hospital, Guy's, founded in 1721.

Southwark and Lambeth were both heavily bombed during World War II, and many popular proletarian resorts, Elephant & Castle and Lambeth Walk, for example, were later rebuilt out of all recognition. Lambeth is one of the most densely populated of the boroughs. The Brixton district, where much of South London's shopping is still done, houses one of the city's principal concentrations of blacks. The borough programs called for demolition and replacement of some of its housing, but a number of its old squares have been renovated rather than razed and have been brought back to life.

Upstream, the adjacent borough of Wandsworth shares its waterfront in part with the somewhat seedy Battersea Park, which contains a children's zoo and used (until 1974) to contain a fun fair that was erected for the Festival of Britain. The riverfront west of the Albert Bridge, an iron suspension bridge of fanciful design, is now taken up with factories, but plans have been made to build a new neighbourhood of apartment housing in a park setting behind a public riverside walk.

London's squares and parks. Although almost nothing remains in London of the works of the architect Inigo Jones (1573–1652), what does remain impresses the eye. Jones came back from an Italian tour with a new architectural idol, Andrea Palladio. The two great examples of Jones's buildings, the Queen's House at Greenwich (designed 1616) and the Banqueting House of Whitehall Palace (1619–23), show that he did not slavishly follow Palladio's classicism. He was no mere imitator but the personal interpreter of a concept and a style. His great contribution to London was not Palladian Venetian but Medicean Tuscan: the city square. Derived from Cosimo I de' Medici's Piazza d'Arme built in Livorno in 1571, which had already inspired Henry IV's Place des Vosges (1603) in Paris, the city square was to set the pattern of London city building for the next 200 years, sometimes fashioned into ellipses and crescents but always enhancing the aesthetic pleasure of the urban composition.

Inigo Jones's contribution to the city layout

Jones's Covent Garden Piazza (1630) was surrounded on three sides by tall houses with an arcaded street floor, filled on the west by the low, solemn-porticoed St. Paul's Church. When it had stood there 50 years, Charles II gave permission for a vegetable market (Covent Garden Market) to operate in the square "forever," which turned out in fact to mean until the 1970s, when it was replaced by the Nine Elms market across the river.

Not long after the Jones square was completed, a cock-fighting hangar was opened in adjacent Drury Lane and was converted into a theatre in 1662. It was destroyed by fire several times, and the present building dates from 1812. External additions were constructed later. For 115 years it was the only licensed theatre in London, until the opening in 1732 of the Covent Garden theatre, where many of Handel's operas and oratorios were introduced. The present Royal Opera House, built by E.M. Barry in 1858, is the home of the Royal Ballet and the Royal Opera.

Some experts contend that Lindsey House on Lincoln's Inn Fields, the largest square in Central London, was built following a Jones design. With its rusticated lower floor and giant pilasters supporting the entablature, it was the prototype for a whole race of London town houses. Charles I gave a direct license in 1641 for the development of Lincoln's Inn Fields as a city square, but work was halted by an appeal to the House of Commons from the society of Lincoln's Inn (this Inn of Court, founded in the 14th century, still contains some 15th-century buildings), and so the square was not completed until 1656. Within the Inn's enclave is New Square, an elegant oblong of brick houses unchanged—except for the effects of German bombing in 1941—since it was built in 1685.

New Square, like Covent Garden and other West End squares, was built as a private speculation. Some of the loveliest portions of London—entire cherished neighbourhoods—were conceived for the benefit of (usually) noble and sometimes royal landlords. The Earl of Southampton in 1660 built Bloomsbury Square, the first to bear the word square in its name. The district was passed by marriage to the dukes of Bedford, who built a splendid series of garden squares, of which Bedford Square (1775) is one of London's handsomest and Russell Square (1800) one of the largest. The whole atmosphere of Bloomsbury, distinguished and intimate, was for a long time enhanced by the presence of teachers and scholars, painters, actors, and writers. Bloomsbury gave its name to a group of artists and intellectuals who met at private homes in the vicinity during the period 1907–30. Among the famous members of this group were Virginia Woolf, E.M. Forster, Lytton Strachey, John Maynard Keynes, and Arthur Waley. In the area are the University of London and its University College and Birkbeck College, as well as the London School of Hygiene and Tropical Medicine, the Royal Academy of Dramatic Art, the Slade School of Art, the Wellcome Museum of Medical Science, and the Courtauld Institute Galleries.

The atmosphere of Bloomsbury

The British Museum. The jewel in this cultural crown is the British Museum, which was founded there in 1753. Its present building was constructed in the period 1823–52, and its library became part of the British Library in 1973. The Reading Room, in which Karl Marx wrote *Das Kapital,* was formed by roofing over the central courtyard. The museum's print room is justly famed, as is the manuscript collection, which ranges from Aristotle's *Constitution of Athens* to modern British classics. The museum collections are extraordinarily varied, especially in antiquities of Egypt, western Asia, the Orient, Greece, Rome, and Britain.

To the west of Bloomsbury, Cavendish Square was laid out in 1717 by the Duke of Newcastle, whose daughter married Edward Harley, 2nd earl of Oxford and Mortimer, and whose granddaughter married the Duke of Portland. Oxford Street is named after the Earl, whose family name was given to one of the world's most celebrated concentrations of medical specialists, Harley Street.

The south side of Oxford Street belongs to the sin-

gle largest private landowner in London, the Duke of Westminster (the Church of England, the crown, and the boroughs of London are thought to be the three biggest institutional landlords). The lands stretch from Oxford Street through Mayfair, across Hyde Park Corner through Belgravia and Pimlico right down to the Thames. Originally Westminster Abbey property, by 1677 it was the dowry of Mary Davies when she wed Sir Thomas Grosvenor—Grosvenor being the family name of the dukes of Westminster. Grosvenor Square was built about 1700 and the Mayfair district surrounding it became immediately fashionable, as it remains today. On the neighbouring estate to the east, Berkeley Square was built about the same time, Bond, Dover, and Stafford streets having been built up some 30 years earlier by Sir Thomas Bond and the earls of Dover and Stafford. The gardens around which Grosvenor Square was built were given to the nation in 1946 as a memorial to U.S. president Franklin D. Roosevelt by the 2nd Duke. Park Lane, where the town houses of the Grosvenors and other wealthy Londoners stood, has become a street of hotels and offices.

The last Grosvenor land to be developed, before the creation of the dukedom of Westminster in 1874, included Belgravia and Pimlico. The original layout, made in 1815, was somewhat altered by the three builders, who took long leases on sections of the property and began construction in 1824. The owners wanted a mixed development—hence the palatial mansions of Belgrave Square (now mostly embassies and institutions), the tall Parisian town houses of Eaton Square, the opulent family houses in Chester and Wilton streets, and the more modest houses in Bourne Street, Graham Terrace, and Caroline Terrace. Although these houses may have differed originally in the kind of resident they attracted, today their occupants all tend to be wealthy. This imposing piece of town planning remains largely intact. Pimlico (by the Thames), until the 1970s ill-served by public transport, had become shabby by the end of World War II, but those parts that survived the severe bombing were subsequently restored to something close to their original dignity.

Originally all the squares in all the developments, planted as gardens or parks on private property, were closed to all save house owners with keys to the gate. Leicester Square, a common grazing ground usurped by the Earl of Leicester to grace his town house, was surrounded by houses by 1680 but became public only in 1870, and Lincoln's Inn Fields remained private until 1894. The majority of the gardens are still fenced and available only to key holders from surrounding buildings. Among the buildings surrounding the garden in St. James's Square (built by Henry Jermyn, earl of St. Albans, 1664) are Chatham House (home of three British prime ministers, present headquarters of the Royal Institute of International Affairs), the London Library (a private subscription library), several private clubs, and some discreet corporation headquarters. There are some modern intrusions, but most of the houses are 17th- and 18th-century. Fiscally, physically, and psychically, the square and its environs continue to constitute the conservatory of the "British Establishment," that astonishingly hardy perennial of island ecology.

Trafalgar Square. Possibly the most famous of all, Trafalgar Square, has always been public and has no garden. Seven major arteries pump automobiles around the great paved space, in which stands a statue of the hero of the Battle of Trafalgar (1805), Horatio Nelson, on a victory column. Two fountains, four stone lions, live pigeons, and assorted memorial military statuary keep him company. The openness of the space attracts numbers of mass meetings every year and is a favourite resting place for tired sight-seeing visitors.

The name Trafalgar Square was bestowed on the site in 1832, after which hundreds of hovels were cleared from the area. Nelson's column was erected in 1842. The National Gallery (into which architect William Wilkins in 1838 was obliged to incorporate the colonnade from George IV's razed palace, Carlton House) occupies most of the north side of the square, with the National Portrait Gallery next to it. To the east of the galleries is the church of St. Martin-in-the-Fields (1722–26) by James Gibbs, replacing a 1544 church on a site where a house of worship was first mentioned in 1222. The site shows off the church superbly. Just below, the Strand runs off to the City, one-half mile east, where it changes its name to Fleet Street.

South of Nelson's monument is a statue of Charles I, erected on the site of Charing Cross where the funeral cortege of Eleanor of Castile, wife of Edward I, rested in 1290. Several of the Cromwell supporters who voted the execution of Charles I were themselves executed there after the Restoration, after which the rediscovered 1633 statue was erected. To the west is the 1910 Admiralty Arch, which frames the entrance to the Mall, leading to Buckingham Palace.

Whitehall. Directly south from Trafalgar Square–Charing Cross runs a street that was once the power centre of the world, Whitehall. It remains the seat of the government's executive branch. Interspersed among typical Victorian and post-World War II structures are several graceful 18th-century buildings still in government service. On the east side is Inigo Jones's calm, lovely Palladian Banqueting House (1619–23), the only important part of Whitehall Palace to survive intact the fire of 1698. Across the street is the Horse Guards Parade, a 1760 replacement of an earlier headquarters. Every day the plumed troopers of the Household Cavalry ride in on black horses from their 35-story Knightsbridge Barracks (1970, Sir Basil Spence) to post guards at 11:00 AM.

The seat of government

Farther south is the Old Treasury, built around the remains of Henry VIII's tennis court after the fire, now occupied primarily by the Cabinet Office, connected by a corridor to the Cabinet Room at 10 Downing Street. Sir George Downing built this very short street of houses between 1663 and 1682 as a speculation, and behind the simple brick facades are 200 rooms reconstructed in the 1960s. The prime minister lives at No. 10, the chancellor of the Exchequer at No. 11, and the government's chief whip at No. 12. Past the Foreign and Commonwealth offices, the Home Office, and the all-powerful Treasury is Parliament Square, which has a small central garden with benches and statues. The square is the site of a bronze statue unveiled in 1973 of Sir Winston Churchill in a typically pugnacious pose. It leads on to two of the most important buildings in Britain, the Houses of Parliament and Westminster Abbey.

Parliament. Parliament dates its beginnings from 1265, when Henry III had to accept the Parliament called by Simon de Montfort. Meetings on fixed dates began under Henry VIII, and in Tudor times the House of Commons evolved its basic rules of procedure. Early meetings were held in the Abbey's Chapter House, which was then, from 1547, in the deconsecrated chapel of St. Stephen of the Palace of Westminster. The House of Lords met in the palace Chamber of Requests in Old Palace Yard. Originally built for Edward the Confessor (reigned 1042–66), the palace served Parliament until destroyed by fire in 1834. The crypt of St. Stephen, a section of 1526 cloister, and the medieval Westminster Hall are all that remain. The hall, completed in 1097, received its magnificent hammer-beam roof during 14th-century modifications. The roof was damaged by incendiary bombs in May 1941 but was later restored.

After the 19th-century fire, a competition for a new Palace of Westminster was won by Charles Barry, who was knighted for his work, and Augustus Welby Northmore Pugin, an ardent medievalist, who died mad. Full of variety and movement that lighten its massiveness, the prize-winning design functioned well for the Parliament of its day and became almost instantly a symbol of that institution's power and solidity. The Clock Tower at the eastern end houses among its bells Big Ben, which sounds the hours and quarters and is marked "13 tons 3 cwts 3 qtrs 15 lbs."

Big Ben

Westminster Abbey. Built by Edward the Confessor, Westminster Abbey was consecrated in 1065. To honour Edward, Henry III undertook to build a more magnificent church in 1245, in Gothic, leaving of the original building only the Norman Undercroft (since 1966 known as the Treasures Exhibition), the Chapel of the Pyx (formed from part of the Undercroft in the Middle Ages), and the Dark



Majestically dominating the London skyline are Big Ben, the Post Office Tower (left), and the Thorn Electric Building (right).
Fox Photos—Pictorial Parade

Cloister, all of which still exist. After 25 years the work stopped, not to be resumed until more than a century later, but in perfect imitation of the style in which it was begun. The octagonal Chapter House is one of the identifiable portions built in the mid-13th century. Henry VIII completed the fan-vaulted and richly carved chapel begun in 1503 by his father at the eastern end of the abbey. The restorations and additions of the 18th and 19th centuries were drastic; the twin towers on the west front are from 1735 to 1740, designed by Nicholas Hawksmoor. The church is styled a Royal Peculiar, with Dean and Chapter (its officials) depending directly on the monarch, head of the Church of England. Nearly all sovereigns since Harold II (who lost to William I the Conqueror in 1066) have been anointed and crowned there, and until 1760 they were buried there as well. National heroes of politics, literature, and war are interred there, although some of the most heroic (Shakespeare, Milton, Marlborough, Nelson, and Churchill, for example) lie elsewhere.

The abbey, its boys' school (Westminster, one of the nation's most distinguished), and its neighbour, the medieval Gothic church of St. Margaret's (parish church of the Commons), form a placid precinct on the edge of Victoria Street, cut through slums in 1851 to Victoria Station. Much of the bold Victorian architecture along the street was replaced in the 1960s with indifferent contemporary buildings, including a 22-story glass box, the Westminster City Hall, and just off Victoria Street, in a building of the same mock-courageous cut, is the New Scotland Yard, the headquarters of the Metropolitan Police, built in 1964. Scotland Yard, the 1891 Scottish baronial castle of convict-quarried granite, is half a mile east on Victoria Embankment.

St. James's Park. Behind Whitehall's ministries and Downing Street, westward toward Buckingham Palace, lies St. James's Park, the oldest and most ornamental of the six central Royal Parks that give the West End of London a heritage of beauty, repose, and sense of continuity.

The swampy ground belonging to the Sisters of St. James-in-the-Field was taken in 1532 by Henry VIII, who built nearby St. James's Palace, like a red-brick toy-soldier castle. To the court of St. James's the credentials of foreign diplomats are still addressed. The Stuarts admitted the public to the park, which Charles II had done up in the formal Versailles manner. The first Georges cut off one end of the ornamental water to form the Horse Guards Parade, where in early June the sovereign's official birthday is celebrated with the Trooping the Colour by the Brigade of Guards, to the music of the brigade's massed bands. Horse Guards Parade

George IV had the park redesigned in its present seductive form by John Nash in 1828. The formal "canal" was changed into a long, irregular lake graced with a small bridge from which the Whitehall buildings, topped with pinnacles and towers, form a rather "Arabian nights" silhouette. The weeping willows around the lake also lend an Oriental touch to what is an archetypal (and extremely artful) English landscape. The flower beds are the best in Central London. Among the many species of waterfowl to be seen are the pelicans originally introduced by Charles II. The northern border of the park is the Mall, a splendid ceremonial way that terminates at the bulky Queen Victoria Memorial, around which are planted each spring 40,000 tulips, succeeded by 14,000 geraniums.

Buckingham Palace. The memorial stands before Buckingham Palace, built in 1703 by the Duke of Buckingham and bought by George III in 1762. When Nash reshaped the park, he enlarged and remodelled the palace and designed Marble Arch, which served as the entryway until 1851, when it was removed to the northeast corner of Hyde Park. The Mall facade of the palace was redesigned again in 1913. When the monarch is in residence, the Royal Standard flies from the roof. The guard is changed in the forecourt every morning from April to September and every other morning the rest of the year whether or not the monarch is in residence. Victoria made Buckingham Palace the monarch's London residence, relegating St. James's to purely ceremonial use.

Green Park. Green Park, close by, is an unpretentious stretch of especially thick, rich grass and luxuriant stands of trees, the plainest of the Royal Parks. It is bordered on the east by the once-private homes that clustered around St. James's Palace: 19th-century Lancaster House, about which Queen Victoria remarked to the Duchess of Sutherland, "I have come from my house to your palace," now a government conference building; Bridgewater House, a 19th-century Florentine palazzo and one of the 280 buildings that are on the Grade I list of protected historic buildings; and the mid-18th-century Spencer House.

Despite the increasing number of modern buildings, the district embraced by St. James's Street and Pall Mall still retains its great distinction. More than a dozen of the most illustrious gentlemen's clubs are located in this corner. There are also shops, some of which have been in St. James's for more than 200 years.

The part of Piccadilly that runs along the northwest side of the park was from the 18th century the site of great town houses, some few of which survive today as clubs. No. 1 Piccadilly, once known simply as "No. 1, London," was the town house of the dukes of Wellington from 1815 until it was given to the nation in 1947 as the Wellington Museum. This mansion was built as Apsley House by Robert Adam in 1778, and the 1st Duke had the red brick clad in stone and added a big Corinthian portico. Wellington Arch (formerly called Constitution Arch), which stands on an island surrounded by roadway opposite the house, was designed by Decimus Burton as a memorial to Wellington.

Hyde Park and Kensington Gardens. Across the traffic of Hyde Park Corner is another work by the same architect, the triple-arched Ionic screen that was the original entrance to Hyde Park, for which Burton also designed the lodges at the other gates. From this corner a celebrated riding track, Rotten Row, starts westward under the arching boughs. Once the central part of Henry VIII's great forest hunting preserve watered by the West Bourne and Tyburn, Hyde Park and Kensington Gardens together have an area of 615 acres. The Tyburn Tree, a 12-foot-high triangular

Marble Arch and Speakers' Corner

gibbet with room for eight malefactors on each arm, stood at the northeast corner of Hyde Park until 1783. Marble Arch, which gave its name to the adjacent crossroads, is now isolated by traffic, but formerly it occupied the northeast corner of the park with Speakers' Corner, the traditional open-air, open-question forum. From there to the southwest corner of Kensington Gardens is a two-mile walk across grass and through groves of fine old trees. Only in very rare spots is the public requested to keep off the grass, under which, on the violated eastern border of the park, is a huge parking garage.

From an Italianate water garden in Kensington Gardens flows the Long Water, which, once it enters Hyde Park, becomes the Serpentine, where, depending on the weather and the hardiness of participants, there is boating and ice-skating or, at a strand called the Lido, swimming. William III nipped 26 acres off the western end of Hyde Park in 1689 to make a garden for Kensington Palace, enlarged by further acquisitions in succeeding reigns. Kensington Gardens was not opened to the general public until the mid-19th century and, until recently, still retained some feeling of elevated separateness, with nannies tending prams, although nowadays au pairs pushing strollers (push-chairs) are more usual. Expensive model boats are still sailed in the Round Pond and magnificent kites are still flown. The traditions of Hyde Park have always been—aside from the vanished magnificence of mass cavalcades in Rotten Row—more popular, from public hanging to public swimming. The Great Exhibition of 1851 was held in the Crystal Palace, Sir Joseph Paxton's remarkable iron-trussed glass building. The palace was erected in Hyde Park, almost, but not quite, in Kensington Gardens and later moved to South London. It was destroyed by fire in 1936.

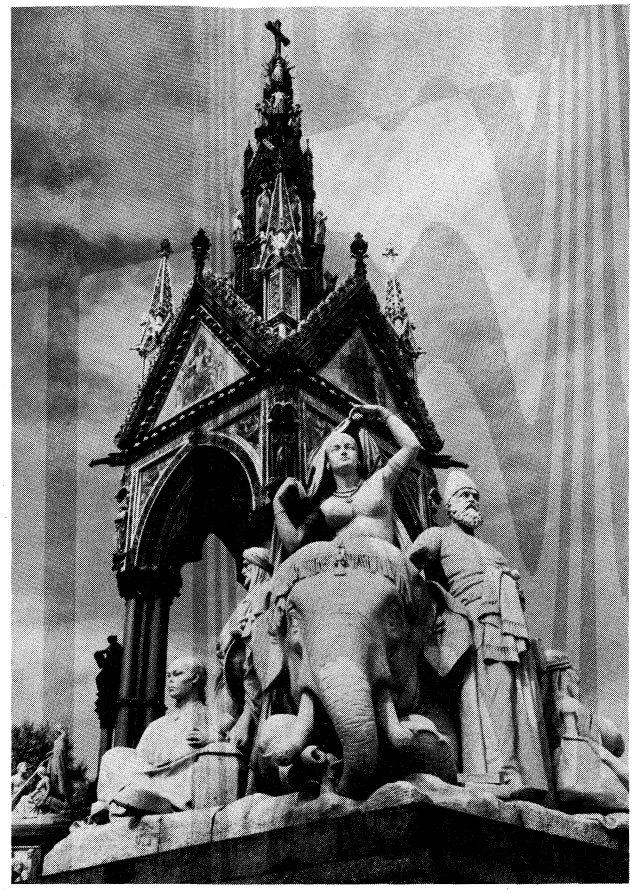
Unpretentious, domestic, dark red-brick Kensington Palace was the Earl of Nottingham's 1661 country house, enlarged by Sir Christopher Wren for William III in 1689 and again by William Kent for George I. About 1700, Kensington Square to the south was built for those who wished to be near the court. When she succeeded to the throne (June 1837), the 18-year-old Victoria was living in Kensington Palace with her mother. It is still inhabited by members of the royal family, and its staterooms are open to the public.

To the side of the palace runs a private road constructed in 1843 and lined with grandiose mansions of which a dozen are now embassies and some others are diplomatic residences. The southern end leads to the relatively opulent district of Kensington, and the northern end to the more heterogenous area of Notting Hill, both parts of the Royal Borough of Kensington and Chelsea. Although the borough has the smallest percentage of public-housing tenants of any in London, it has, in its northern reaches, districts of great squalor.

Albert Memorial and Royal Albert Hall

Kensington Gardens accommodates one of London's more remarkable monuments, the Albert Memorial, 175 feet high, graced with 175 sculpted figures somewhat larger than life. This Victorian Gothic improbability was concocted in 1863–71 by Sir George Gilbert Scott, who, at the same time (1868–74), wrought even more fanciful gothic conceits at St. Pancras Station, fortunately preserved. Out of the park and across the street is the domed, circular, Royal Albert Hall, part of a cultural centre for which the ground was bought with profits from the 1851 exhibition. There are five colleges and four museums within a quarter mile of Albert Hall (to the south): the Royal College of Organists, the Royal College of Music, the Royal College of Science, the Imperial College of Science and Technology, and the Royal College of Art, which was moved to Kensington Gore from Exhibition Road in 1962; the Science Museum, the Geological Museum, the Natural History Museum, and the Victoria and Albert Museum.

Regent's Park. Regent's Park, with Primrose Hill abutting, is about a quarter of an hour's walk north from Marble Arch. This parkland, too, was part of Henry VIII's hunting preserve. The elongated lake was made by damming the Tyburn. It was opened to the public in 1815, the culminating point of John Nash's Triumphal Way, which started from St. James's Park. Starting in 1811 and working until 1830, he built Waterloo Place (the column to the Duke of York was erected later), then lower Regent Street to Piccadilly. From there he was unable to continue Regent Street on a straight line, but he did obtain a right-of-way in a quarter-mile curving arc. To join this curve to the straight, he contrived the city's first traffic roundabout—Piccadilly Circus. Along the curved section, known as the Quadrant, he swung twin lines of arcaded buildings. (The arcades were ripped out before Nash was dead a dozen years.) He thrust on until he came to where his street had to be joined to Portland Place, already lined with buildings by Robert Adam (little of that work is left, either). There he constructed a church, All Souls Langham Place, with a circular portico, gently concealing the awkward junction of the two streets.



The Albert Memorial in Kensington Gardens.
Brian Seed—Aspect Picture Library

John Nash's Triumphal Way

The effect was marred later in the 19th century, when the ornate Langham Hotel was erected, and the vista changed still more drastically in 1931, when the British Broadcasting Corporation planted its Broadcasting House there.

Having already put Regent's Park in order, Nash massed around three sides of it one of London's most magnificent architectural showpieces, a series of terrace houses. The facades of the individual houses in each terrace, based on a continuous ground-floor arcade, were composed to form one grand entity with column-supported pediments and a unifying lattice of balcony across the front, so that each tenant seemed to occupy a portion of a palace. War damage and dry rot made the future of these terraces uncertain for years after 1945, but by 1970 they had been restored and put to educational, charitable, and residential uses.

The park has the amenities offered by most of the others, but two of its endowments are unique: the Royal Zoological Society's Zoological Gardens (a very superior middle-sized city zoo; the society's big zoo is at Whipsnade) and the Regent's (Grand Union) Canal. The western end of the canal is a district of richly maintained large houses called Little Venice.

Nash's Piccadilly Circus was no longer circular after 1877, when Shaftesbury Avenue was started there and hacked through the rookeries (slums) of Soho and St. Giles,



Piccadilly Circus.
© Karen I. Hirsch 1983

The atmosphere of Soho

eventually becoming the principal theatrical street of the West End. The Soho district has been the special resort of foreign colonies since Frenchmen found refuge there from the 16th-century wars of religion, followed in turn by Greeks in 1677 and, during the 19th century, the Italians.

These three groups with their ethnically flavoured restaurants and grocery shops still thrive in Soho amid the grimy nightclubs and the remnants of between-wars Bohemia. They have been joined by a growing number of Chinese from Hong Kong. The old Chinatown in Limehouse in dockland was virtually deserted after World War II. A sizable Greek community flourishes in Bayswater near the Greek cathedral of Hagia Sophia.

Piccadilly Circus, a seedy traffic chute by day and a neon-lit gathering place by night, still attracts visitors from all over the world. The youth of all nations sprawl on the steps of its stone island, which is crowned by the 1893 aluminum statue of Eros (in reality, the Angel of Christian Charity, a memorial to the reformer the 7th Earl of Shaftesbury).

The outer parks. Outside London and Westminster there are four other Royal Parks: Greenwich, Richmond, Hampton Court, and Bushy. Richmond Park has changed little since it was first enclosed in 1637 for the grounds of the vanished Richmond Palace. The ponds abound in fish, and the centuries-old oaks shelter several hundred deer. The Royal Ballet School occupies White Lodge in the middle of the park, which comes near to the Thames 15 miles upstream from Westminster Bridge. To the east are Wimbledon Common, Wimbledon Park, and Putney Heath. On the edge of the park is Alton West Estate, Roehampton, a low-rent housing group. In a garden setting, which includes some preserved 18th-century country houses, the community is served by its own schools, shops, and library.

Hampton Court

Further upstream past the tidal head at Teddington are Hampton Court and Bushy parks. Hampton Court palace was begun in 1515 by Cardinal Wolsey and was presented, along with its contents, to Henry VIII in about 1528 in a futile effort to stave off disgrace. The house, an imposing red-brick building with white-stone accents and typically Tudor twisted chimneys, was enlarged for Henry and, in the 17th century, enlarged further for William and Mary by Sir Christopher Wren, who planted the avenue of flowering chestnut trees that grow today in Bushy Park. This park is inhabited by deer, by horses belonging to the Royal Household, by the director of the National Physical Laboratory (in 18th-century Bushy House), and by the Admiralty Research Laboratory, in Charles II's Upper Lodge.

On the way back downstream is Kew Gardens, one mile below Richmond Park. Although it contains a former royal palace (1631), Kew is not a royal park. It is the site of the Royal Botanic Gardens, with four museums, a research and training centre, and gardens of trees and plants from all over the world.

Greenwich Park is farther downstream, five miles below London Bridge. The park was first enclosed by Humphrey, duke of Gloucester, in 1433. He built a castle on the hill, which Henry VIII enlarged and which Charles II demolished in 1675, to make way for Wren's Royal Observatory. The site of the Greenwich observatory became by international convention the origin of the standard meridian of longitude, and, thus of Greenwich Mean Time. The Royal Observatory itself was moved to Herstmonceux castle in East Sussex between 1948 and 1958. The Duke also built a palace, Placentia, on the riverbank, where Henry VIII and his daughters Mary and Elizabeth were born. On the rise behind the palace, Charles I had Inigo Jones build for Henrietta Maria the Queen's House, now the National Maritime Museum. Charles II tore down Placentia and had John Webb start a new palace. William and Mary made the building into a naval hospital and had Wren design further buildings. Nicholas Hawksmoor and Sir John Vanbrugh were architects for later additions. This so-called English Baroque group in Greenwich Park, best seen from the river, is one of the best sited and most finely balanced architectural compositions in London.

The park that rises behind them is graced with Spanish elms set out in the mid-17th century. Like the other Royal Parks, it is spacious, soothing, and deceptively "natural." Just to the south lies historic Blackheath, where rebel armies gathered on several occasions down the centuries, where James I introduced the Scottish game of golf to England, and where highway robbers lurked in the 18th century.

Hampstead Heath

Far to the opposite side of the city is another open space much prized by Londoners, Hampstead Heath. The residents of the wealthy and picturesque village of Hampstead, on the highest hill in London, organized into a protective

society and fought in the courts from 1856 to 1872 to assure free access to an integrally preserved natural open space. The heath, the citizens' group, and a good many village aspects still prevail in Hampstead today.

Highgate Wood, just north of the heath, belongs to the City of London, as do Queen's Park and West Ham Park. Farther out, other parks sustained by the City of London include Epping Forest in Essex, Coulsdon Commons in Surrey, and Burnham Beeches in Buckinghamshire. Hainault Forest, in Essex, is maintained by the London Borough of Redbridge. Both Hainault and Epping forests are remnants of what was a much larger forest, the Royal Forest of Essex.

Most parks in and around Central London provide entertainment (band and symphony concerts, drama, poetry readings, ballet, and opera) and facilities for sports from bowling on the green to motor racing. Coaching is given in lawn tennis and skiing, and championships open to all Londoners are held every year in seven different sports. On Sunday mornings in the season, Hackney Marshe park offers the spectacle of 2,500 footballers at play.

THE PEOPLE

Within the Greater London conurbation a decline in population has occurred in the City of London since 1851, in inner London since 1901, and in outer London since 1951. Encircling the conurbation area is a broad fringe of almost continuous population increase that embraces satellite towns beyond the Green Belt areas and extends along every major road and rail route from London.

As hundreds of thousands of people move out annually, a smaller number of "new Londoners" move into those city neighbourhoods that have often lost their "village" character to World War II bombs or urban renewal. Modern construction, especially of hotels and offices, has gnawed into residential districts. Most enduring of all has been a change in the outlook of the mass of the population: the working class has been emancipated from belief in its own inferiority. Better housing, improved education, a complete scheme of health care, and social security have done much to terminate mute acceptance of the old social system.

Erosion of the "working-class mentality"

The class system persists, and a much weakened "old boy network" of former public (private) school pupils may still marginally aid its members, but society—if perhaps not imbued with quite the easy egalitarianism of Australian or North American cities—is more open, and the traditional "good family, good school, good regiment" endorsements are not essential to a career. Great tracts of Central London still belong to ducal families, and some corporations still decorate their boards of directors with the nobility. At the opposite end of the scale, inhabiting the last of the substandard houses and finding the least desirable employment, are the blocs of unabsorbed immigrants, mostly nonwhites from former colonies. Legislation now protects them against the worst effects of racial discrimination, and special efforts have been made to accelerate their integration into London life.

ADMINISTRATIVE CONDITIONS

Greater London covers more than 600 square miles, but there are, administratively speaking, a dozen different Londons. For electricity and gas supplies and hospital services, Greater London is a part of an administrative area for four electricity boards, three gas boards, and four metropolitan regional hospital boards. The City of London has its own police force and its own government headed by the lord mayor. The London school area is the responsibility of the Inner London Education Authority; the outer London boroughs have their individual school systems. The London postal area is 570 square miles. The Metropolitan Police district, which does not serve the City of London, covers 786 square miles. The area serviced by the London Transport Board is 900 square miles, and that serviced by the Metropolitan Water Board is 570 square miles.

Administrative overlap and its effects

In 1965 the Greater London Council (GLC), which administered an area of 610 square miles, replaced the London County Council, which had taken over from the Metropolitan Board of Works as the city's government in 1889. The GLC was responsible for—among other things—overall planning, traffic control, sewers, courts, and historic buildings. Also in 1965 the separate local authorities were abolished, and 32 London borough councils (plus the City of London) were set up as the primary units of local government, little "cities" of 150,000 to 340,000 population. The elected borough councils fix the annual rate payable on local property and have charge of such activities as housing, public libraries, refuse collection, and borough planning. In 1986, officially as an attempt to save money, the GLC was abolished, and many of its services were taken over by the elected borough councils. Certain GLC responsibilities, such as civil defense, that require authority over areas larger than a borough have been transferred to joint authorities whose members are borough councillors.

History

THE EARLY PERIOD

Foundation and early settlement. The history of London begins effectively with the Roman period. Gracechurch Street, which leads from London Bridge up to Cornhill, led from the bridge in Roman times to the hub of Londinium, where the Romans built a basilica. Seventeen years after the Romans had installed themselves, Icenian tribesmen under Queen Boudicca (Boadicea) revolted and sacked the city. From traces of the fires they set, it can be determined that the settlement had already begun to spread westward across Walbrook toward the hill where St. Paul's Cathedral would later be built. After the sack, a bigger basilica was built: an aisled hall 500 feet long. On the same spot today stands Leadenhall Market, an 1881 effusion of cast iron and glass. To protect the city, Cripplegate Fort was constructed at the northwestern edge, but walls were not built until the end of the 2nd century AD. Remains of the wall can still be seen at the edge of Barbican (the street is called London Wall) and in a small park on Tower Hill. In medieval times the walls were rebuilt and extended, necessitating construction of additional gates but retaining the Roman gates: Aldgate, Aldersgate, Bishopsgate, Cripplegate, Ludgate, and Newgate. The legions were recalled to Rome early in the 5th century, and what happened to the city in the next two centuries is a matter of conjecture, though the Roman street pattern was not obliterated. (Cheapside and Cannon Street were the medieval successors to Roman east–west roads.)

No records tell how or when London fell into Saxon hands, but it was still, or had once again become, a city of importance by AD 597, the year that Pope Gregory I the Great sent St. Augustine to England from Rome. Aethelberht I, king of Kent, founded St. Paul's Cathedral, and Mellitus was installed as bishop there in 604. In the following century the Venerable Bede described London as a mart of many people arriving by land and by sea. After that the city vanished from Anglo-Saxon chronicles for more than a century, until the time of Alfred the Great (849–899) and the wars with the Danes.

Medieval London. The city's future importance as a financial and military—and therefore political—power was made clear at the time of the Norman Conquest (1066). One of the first acts of William I the Conqueror was to accord a charter promising the citizens that they should enjoy the same laws as under Edward the Confessor and that he would suffer no man to do them wrong. It was a further mark of respect for the city's latent puissance that he planted just outside its walls the intimidating Norman keep, the White Tower, central stronghold of the fortress-castle known as the Tower of London. A roughly square (118 by 107 feet) structure 90 feet high, it has towers at each corner of the walls, which are 15 feet thick at the base and 11 feet thick at the battlemented top. King Richard I returned from the Crusades with a new concept of fortification and began surrounding the keep with concentric systems of curtain walls with towers at intervals, a work completed by Henry III. Virtually every reign since then has added its bit, and the Tower incorporates almost every period of English architecture. An official royal residence until James I, it has also housed the Royal Mint,

The importance of the Tower of London

the Royal Menagerie, the public records, an observatory, and an arsenal. Although it is popularly thought of as a traditional site of executions, only six multiple executions took place within the confines of the Tower of London in its millennium-long history. All the others were carried out on Tower Hill, which is technically outside its walls. It has always served as a prison for state offenders. The first prisoner was the Bishop of Durham in 1101, and the last was the deputy party leader of Nazi Germany, Rudolf Hess, during World War II. The Crown Jewels are on display there, as is a superb collection of English arms and armour.

When Henry I died in 1135, Londoners, exercising their "right" to elect the king of England, chose Stephen, who was crowned three weeks later at Westminster Abbey. About 1136 a major fire destroyed many of the city's wooden houses and St. Paul's itself. In the rebuilding, houses of stone and tile began to appear, and streets were paved and at least partially cleansed by open sewers and conduits. By the end of the 12th century, the large colony of Danish merchants was outnumbered by the Germans, who had their own trading enclave, the Steelyard on the riverfront, until 1598. Other important trading groups were the Gascons, the Flemish, and the northern Italians. When the last named were firmly established as bankers, the Jews (in 1290) were despoiled, imprisoned, and then banished, not to return for centuries.

By astute purchase from needy monarchs, the guilds—110 of them by 1400—were able to buy increasing freedom from royal intrusion in their affairs and further their self-government. The first mayor of London, Henry Fitzailwin, probably took office in 1192. The first evidence of a Common Council dates from 1332. Since disorder in the realm provoked unrest in the City, London usually supported strong, orderly government, especially in such crises as the deposition of Edward II (1327) and Richard II (1399), the English Peasants' Revolt in 1381, and the rebellion headed by Jack Cade (1450).

Tudor London. Toward the middle of the 16th century, London had a startling growth of trade and population. From 1530 to 1600 the number of Londoners trebled. The excess population at first found living space in the grounds of the religious institutions seized during the Reformation by Henry VIII. To fill the void left by the cessation of the religious charities, the city organized poor relief in 1547, providing grain in times of scarcity and instigating the foundation or reconstitution of the five royal hospitals: St. Bartholomew's, Christ's, Bethlehem (the madhouse known as Bedlam), St. Thomas's, and Bridewell. Many of the private charities founded at this time still function. Trade was further promoted by the establishment of monopolies such as those held by the Muscovy Company (1555), the Turkey (later Levant) Company (1581), and the East India Company (1600).

Tudor charitable and trading foundations

By 1605 the City population had reached an incredible 75,000, and, despite laws that attempted to restrain its size, a further 150,000 lived outside the walls. Southwark, at the far end of the original London Bridge, became the City's 26th ward. New industries, including silk weaving and the production of glass and majolica pottery, were established, often outside the gates and thus outside the restrictive regulations of the livery companies. The establishment of Henry VIII's naval dockyard at Deptford on the south bank was accompanied by a straggle of waterfront hovels on the north bank at Wapping.

When Henry VIII in 1529 began to convert Cardinal Wolsey's York Place into the royal palace of Whitehall and to build St. James's Palace across the fields, the City of Westminster began to take more definite shape around the court. Between Westminster and the City the great houses of nobles began to appear, with gardens down to the river and each with its own water gate. Along the Strand opposite these houses were distinguished lodgings for gentlemen who were in town during legal sittings. The name London was before long to embrace both the City of London and the City of Westminster and the built-up land between them.

The reign of Elizabeth I (1558–1603) marked the apogee of the City's domination of England. The Queen based her strength on its militia, its money, and its love. It provided one quarter of the men for service abroad in 1585 and 1589 and formed its armed "trainbands" to defend against the threatened Spanish invasion.

17th-century London. The trainbands remained a force to be reckoned with, and Charles I, who had damaged the City's trading interests and flouted its privileges as cavalierly as he had Parliament's, was deterred from attacking London in 1642 by their presence at Turnham Green. Hostility toward the King made the fortified City the core of parliamentary support, and Parliament's success in the Civil War was due in good part to City allegiance.

In 1664–65, the Plague, a frequent invader since the Black Death of 1348, killed 75,000 Londoners. The following year the Great Fire burned from September 2 to September 5 and consumed four-fifths of the City. From the unscorched corners in the northeast and extreme west, the rebuilding began. A rational street plan was rejected, but streets along the old traces were wider and a bit straighter. Between 1667 and 1671 most of the new houses (brick—no more half-timbering was allowed) were rebuilt. Many of the tiny parishes were combined, and a few churches had escaped the fire, so only 50 were rebuilt, in addition to a new St. Paul's. Sir Christopher Wren, mathematician, astronomer, physicist, though only informally trained as an architect, was given the formidable task of designing them and supervising their construction.

The Great Fire and its effects

There is a Wren inscription in St. Paul's Cathedral, *Lector, si monumentum requiris, circumspice* ("Reader, if you seek a memorial, look about you"). The whole City is in fact Wren's memorial. His churches, ranging from the homely Dutch to the Gothic but mostly in his own expression of the classical style, are all touched with the same felicity and springing vigour, a series of virtuoso variations on basic architectural concepts. St. Paul's dome is one of the most perfect in the world and, like the rest of the cathedral, is classical in theme with Baroque grace notes. The memorial column for the Great Fire was adapted from a Wren design and erected near Pudding Lane, where the fire had started in the house of the king's baker. Wren constructed four other churches outside the City, built the Royal Hospital located in Chelsea, and parts of Kensington Palace, Greenwich Hospital, the Royal Observatory of Greenwich, and Hampton Court Palace.

Under Charles II royal abrogation of City rights was resumed, and although James II restored forfeited City charters before his flight in 1688, it was in Guildhall under protection of the trainbands that the lords spiritual and temporal met to declare allegiance to William, prince of Orange (henceforth known as William III of Great Britain).

To support the war against France in 1694, 40 City merchants formed the Bank of England, and thenceforth the City's money market became a prime factor in the affairs of state. Another aspect of the City's power in the nation was the centring of the national press in Fleet Street (*The Times,* founded in 1785 off Blackfriars Lane, only moved to new premises in 1974).

The foundation of the Bank of England

EVOLUTION OF THE MODERN CITY

By 1820, when George IV succeeded to the throne, the whole character of London had changed. Its population had increased to more than 1,100,000. Villages and hamlets that in 1666 had been the objects of summer outings from the heart of the city had become part of the built-up area. Some of the building had been well-planned works of great landowners; some, however, the sorry work of the small and greedy. The Bedford, Portman, and Foundling estates produced streets and squares that embellished the town. On the other hand, to the east, parts of Stepney and Bethnal Green were constructed with ill-built cottage terraces. Agar Town, which lay near the modern King's Cross and St. Pancras railway stations, was a scandal.

The changes brought by the years 1689–1820 had followed no conscious plan. Inside the City its government was in full control and reasonably active. Beyond its boundaries, unchanged since the Middle Ages, government services and communications for the new areas came piecemeal. The important developers obtained local acts

of Parliament enabling them to levy rates out of which to finance paving, lighting, cleansing, and the watch (a group of persons charged with protecting life and property). The popularity of the developers' streets depended in part on such services, and they were usually adequately administered. Lesser men left a legacy of slums and neglect for later generations to clear.

Socially, commercially, and financially, London was the hub of the kingdom. As a corollary to its great wealth, fed by the profits of the East and West Indies' trade and by trade with most of the known world, it reigned supreme in Britain in matters of the theatre, literature, and the arts. It was the London of David Garrick, Oliver Goldsmith, Samuel Johnson, and Sir Joshua Reynolds, of the great furniture makers and silversmiths, and of renowned foreign musicians, both resident and visiting. But its size brought increasing problems.

Organization, innovation, and reform. Although new dispensaries and new or enlarged hospitals were reducing mortality, the former riverside town required new forms of government, of communications, and of sanitation if it was to continue to grow. These were slowly and painfully evolved in the London of 1820–1914. Against a background of statistics that showed the population of the built-up area rising from 1,225,694 (1821) to 6,586,269 (1901), the innovations came piecemeal. In 1829 a centralized Metropolitan Police Force was provided, under the ultimate control of the home secretary, in place of the uncoordinated watchmen and parish constables. The lighting of streets by feeble oil lamps was revolutionized by the introduction of gas, and soon the Gas Light and Coke Company (1812) was followed by similar companies scattered throughout London. Omnibuses (1829) began a revolution in road transport, and carriage by rail came less than 10 years later. In 1845 an inquiry into public health was made, with the exposure of London's worst deficiencies, followed by legislation in 1852 ensuring a purer water supply. A statute in 1855 (the Metropolis Management Act) combined a number of the smaller units of local government and replaced the medley of franchises with a straightforward system of votes by all ratepayers. Major works, such as main drainage, were put in the hands of a Metropolitan Board of Works.

The Victorian era

The momentum of these changes, established by such diverse reformers as Bishop C.J. Blomfield, Sir Robert Peel, Edwin (later Sir Edwin) Chadwick, and the Earl of Shaftesbury, continued throughout the century. New churches, new schools, better law and order, main drainage, and care for the outcasts were some of the reformers' legacy; Trafalgar Square, the Embankment, and roads, such as Shaftesbury Avenue and Charing Cross Road, driven through the worst of the slums are their most obvious monuments. The changes in government continued, if not so drastically. The London County Council superseded the Metropolitan Board of Works in 1889, the vestries were transformed into metropolitan boroughs by the London Government Act (1899), and the various water companies combined in 1902 into a publicly owned Metropolitan Water Board.

Public and private works continued to transform the face of London. The opening of the Metropolitan, a steam railway, in 1863 and the making of Holborn Viaduct in 1869 were accompanied by the building of new Thames bridges and the rebuilding of Battersea, Westminster, Blackfriars, and London bridges. After years of discussion and agitation, the road bridges outside the City passed into public ownership, and the tollgates disappeared. Most of the main southern railways carried their lines northward across the Thames into London, to Victoria, Charing Cross, Blackfriars, and Cannon Street stations. It was an era in which an abundance of initiative and of capital was joined to abundant labour to make the widest use of new skills, cheap transport, and copious raw materials.

Technical progress continued gradually to alter the lives of Londoners and the face of the town. Cheap suburban trains enabled the artisan or clerk to live farther and farther from his work. The London School Board, established under the Education Act of 1870, set about the task of providing elementary education for all. Trains or streetcars (horse-drawn), after an unsuccessful beginning in 1861, became important in the 1870s and a major factor in metropolitan transport as their electrification developed in the first years of the 20th century. By then electricity was being used as the motive power for traffic below ground, the Prince of Wales opening the world's first electric underground railway, from King William Street to Stockwell, on November 4, 1890. With the arrival, before 1914, of the gasoline-driven omnibus, the outline of transport in modern London was complete and the way opened for still faster development of suburbia.

Innovations in transport in the 19th century

Inevitably this was accompanied by rising land values in the central zone, by the construction of ever larger offices, factories, and warehouses in place of small houses, and by a continuous outlay of public and private funds on better housing and street improvements. World War I, in which air raids inflicted some 2,600 casualties on London, brought only a temporary pause, and after the war development resumed on a mounting scale. As a national and in some respects a world capital, London required institutions capable of meeting its needs. An era of amalgamation and expansion ensued. From banks to hospitals, and from telephone exchanges to power stations, almost all was expanded. Street congestion, well-known in the 1850s, was worse in the 1930s, despite the rationalization of traffic authorities. By 1939 the population of the Greater London conurbation exceeded 8,000,000.

Reconstruction after World War II. World War II, with evacuation and heavy damage by air raids, brought the greatest setback in the history of modern London. Air attacks killed more than 30,000, injured more than 50,000, damaged most public buildings, and, in such areas as the City and Stepney, obliterated whole sections of the street system. Westminster Abbey and the Houses of Parliament were damaged but saved, as were St. Paul's and Guildhall in the City. Ordinary houses and the docks were severely bombarded.

The end of hostilities brought a gradual return of many evacuees and a housing shortage made worse by the ravages of dry rot. Reconstruction, hampered by the shortage of most materials, began at once—and with the advantage of plans and surveys already put forward. The Town and Country Planning Act (1944), followed by acts of 1947, 1954, 1959, and 1968, gave unprecedented powers of purchase, direction, and control. The urgency of the Festival of Britain (1951) produced Lansbury (a redevelopment in Poplar) as well as the Royal Festival Hall, but the most significant postwar features have been the vast investment in new houses, the restoration of services and the Port of London, the general acceptance of a planned urban economy, and the sustained effort to divert industries to new or expanded towns outside London.

BIBLIOGRAPHY

Topography and architecture: VICTOR S. PRITCHETT, *London Perceived* (1962; new ed., 1974), an evocative literary work on contemporary London; DAVID PIPER, *Companion Guide to London,* 6th rev. ed. (1977), a good introduction and guide; BLAKE EHRLICH, *London on the Thames* (1966), a popular account of the many "Londons" that make up the metropolis; STEEN E. RASMUSSEN, *London the Unique City,* 3rd ed. (1948, reprinted 1982), a Danish architect's study of the development of London into a metropolis unlike any other; NIKOLAUS PEVSNER, *The Buildings of England: London,* 2 vol. (1952–57), comprehensive, popular surveys of buildings and their architectural features; WILLIAM KENT (ed.), *An Encyclopaedia of London,* rev. ed. (1951), a useful source for a variety of information; ERIC DE MARE, *London's Riverside* (1958), a general account of the Thames and how London has developed along its banks; JOHN SUMMERSON, *Georgian London,* rev. ed. (1962), details of a distinguished architectural period, accompanied by maps and photographs with a list of Georgian buildings in each district; *Royal Commission on Ancient Monuments (Interim Reports),* no. 8–9, 11–13 (1924–30), a well-illustrated inventory of historical monuments in London; LONDON COUNTY COUNCIL, *Survey of London* (1900 *et seq.*), each volume deals in detail with a different area of London, describing the principal buildings and their situation, with maps, plans, photographs, and drawings; WALTER H. GODFREY, *A History of Architecture In and Around London,* new ed. (1962), finds examples to illustrate the history of architecture within a 40-mile radius of London.

Physical characteristics: S.W. WOOLDRIDGE and D.L. LINTON,

Structure, Surface and Drainage in South-East England (1955), a comprehensive scientific treatment, with charts and maps; S.W. WOOLDRIDGE and G.E. HUTCHINGS, *London's Countryside: Geographical Field Work for Students and Teachers of Geography* (1957), on the geologic build of the London region and aspects of its physiography and vegetation; ANGUS D. WEBSTER, *London Trees* (1920), an account of the trees that succeed in London; R.C. HOMES (ed.), *The Birds of the London Area Since 1900* (1957), by a committee of the London Natural History Society, on birds seen in their various habitats; T.J. CHANDLER, *The Climate of London* (1965), a thorough study of the climatic characteristics of the region with particular emphasis on the contrasts between the built-up area and the surrounding rural districts, including a chapter on atmospheric pollution; R. CLAYTON (ed.), *The Geography of Greater London: A Source Book for Teacher and Student* (1964); and JOHN T. COPPOCK and HUGH PRINCE (eds.), *Greater London* (1964), a study by a group of geographers of contemporary London from the viewpoints of physical, social, and economic geography, with photographs, maps, and tables.

Government: SIR IOAN G. GIBBON and R.W. BELL, *History of the London County Council, 1889–1939* (1939), a knowledgeable study of the development of the London County Council and its services; *Report of the Royal Commission on Local Government in Greater London,* 2 vol. (1960), a clearly written study of the development of Greater London as influenced by social and economic factors, which treats the nature and purpose of local government and planning; SYDNEY K. RUCK, *London Government and the Welfare Services* (1963), an account of the scope and variety of the services available, which also includes a description of the reorganization of the services in the new London boroughs; P.G. HALL, *London 2000* (1963), on the town-planning problems of the London region, including a program for regional development and urban restructuring (thoroughly documented); and SYDNEY K. RUCK and GERALD RHODES, *The Government of Greater London* (1970), a study that discusses the background and workings of the system of governing Greater London that was in effect from April 1, 1965, to March 31, 1986.

Commerce and industry: P.G. HALL, *The Industries of London Since 1861* (1961), a study of the locations of various industries in London and how these change, with an investigation of the new industries in the metropolis; JAMES BIRD, *The Geography of the Port of London* (1957), an analysis of the port, its docks and adjacent industrial development, and a study of the river traffic, markets, and trade; J.G. BROODBANK, *History of the Port of London,* 2 vol. (1921), based partly on Port of London Authority records; HENRY MAYHEW, *London Labour and the London Poor* (1851; 4 vol., 1861–62, reprinted 1969), a classic report; CHARLES BOOTH, *Life and Labour of the People of London,* 9 vol. (1892–97), a description of London life during a period of momentous changes—a pioneer work in the field of social investigation; LONDON SCHOOL OF ECONOMICS, *The New Survey of London Life and Labour,* 9 vol. (1930–39), a continuation along the lines of the Booth work, with the investigated region widened to take in all of Greater London; CENTRE FOR URBAN STUDIES, *London: Aspects of Change* (1964), views of historians and sociologists on the changing communities of the metropolis; DONALD L. FOLEY, *Controlling London's Growth* (1963), a study of planning for London since World War II and its results; and MARY CATHCART BORER, *The City of London: Its History, Institutions, and Commercial Activities* (1962), on the commodity markets, Baltic Exchange, Stock Exchange, and other aspects of mercantile London.

Transportation: IVAN MARGARY, *Roman Roads in Britain,* 3rd rev. ed. (1973), a study of these roads, most of them leading out of London; VERNON SOMMERFIELD, *London Transport* (1934), a record and a survey of the system, including a history of the events leading up to the creation of the London Passenger Transport Board that unified services and evolved into London Transport; and (ed.), *London's Buses* (1933), a brief but informative account of their 100-year history, with many illustrations; HOWARD LINECAR, *British Electric Trains,* 2nd ed. (1949), which includes a study of the building of London's early tubes and the development of the London Transport lines; E.A. COURSE, *London Railways* (1962), a history of the development of the railways that provides insight into London's growth in the 19th century; and H.P. WHITE, *A Regional History of the Railways of Great Britain,* vol. 3, *Greater London* (1963), which investigates the reasons for placement of the railways and studies their relationship to the economic and social life of London.

History: RALPH MERRIFIELD, *London: City of the Romans* (1983), a discussion of the latest archaeological discoveries; W.T. HILL, *Buried London* (1955), an early history of London in the light of the discoveries of the London Excavation Council, chiefly concerning the Roman period; GORDON C. HOME, *Roman London,* rev. ed. (1948), founded upon archaeological investigations; WILLIAM PAGE, *London: Its Origin and Early Development* (1929), a political history to the end of the 12th century; REGINALD R. SHARPE, *London and the Kingdom,* 3 vol. (1894–95), a well-documented history of the City in its relations with the monarchy and parliaments; GWYN A. WILLIAMS, *Medieval London, from Commune to Capital* (1963), an account of London during the rise of the merchants, based on City archives; JOHN STOW, *A Survey of London,* ed. by C.L. KINGSFORD (1908, 1956), reprinted from Stow's text of 1603 with additional notes tracing his source of information; N.G. BRETT-JAMES, *The Growth of Stuart London* (1935), a well-annotated work based principally on original sources; T.F. REDDAWAY, *The Rebuilding of London After the Great Fire* (1940), on the economic and social forces that shaped the rebuilding; M.D. GEORGE, *London Life in the XVIIIth Century,* new ed. (1966), which chiefly records the life and work of the poorer classes, using quotations to provide contemporary points of view; ROBERT CARRIER and O.L. DICK, *The Vanished City* (1957), a general account of the London of the 17th and 18th centuries that was obliterated by the changes of the 19th, with illustrations; and HENRY WHEATLEY and P. CUNNINGHAM, *London, Past and Present,* 3 vol. (1891), a useful reference work describing various major buildings and localities. (B.E./Ed.)

Los Angeles

A semitropical southern California metropolis of palm trees and swimming pools, television studios and aerospace factories, Los Angeles has become the second most populous city and metropolitan area (after New York) in the United States. The city sprawls across some 464 square miles (1,202 square kilometres) of a broad coastal plain agreeably situated between the San Gabriel Mountains on the east and the Pacific Ocean on the west. Its hallmark is an architecturally dramatic network of freeways. The automobile so dominates life in this uniquely mobile community that Reyner Banham, an English observer who took his cue from scholars who study Italian in order to read Dante, is said to have learned to drive a car so he could "read Los Angeles in the original."

The city is the seat of Los Angeles County, which contains more than 80 other incorporated cities, including Beverly Hills, Pasadena, and Long Beach, within its 4,083 square miles. The county also encompasses two channel islands, Santa Catalina and San Clemente; a mountain peak, Mt. San Antonio, familiarly known as Old Baldy, 10,080 feet (3,072 metres) high; more than 900 square miles of desert; and 74 miles (119 kilometres) of seacoast. The metropolitan area has paid for its spectacular growth by acquiring such present-day urban attributes as smog-filled skies, polluted harbours, clogged freeways, crowded classrooms, explosive ghettos, and annual budgets teetering on the brink of bankruptcy. Since the city and the county are so intertwined physically and spiritually, any consideration of Los Angeles must move back and forth between the two entities.

This article is divided into the following sections:

Physical and human geography 307
- The landscape 307
  - The city layout
  - Climate
  - Natural phenomena
- The people 309
- The economy 309
  - Agriculture
  - Industry
  - Motion pictures
  - Transportation
  - The communications media
- Administrative and social conditions 310
  - Government
  - Education
- Recreation and cultural life 310
  - Parks
  - Sports
  - Historic landmarks
  - The arts

History 311
- The growth of the metropolis 311
  - The Spanish–Mexican town and city
  - The American city
- The modern city 312

Physical and human geography

THE LANDSCAPE

The city layout. The city is grotesquely shaped, like a charred scrap of paper, with independent municipalities such as Beverly Hills and Culver City as well as unincorporated county land lying within its boundaries. Elevation averages about 275 feet (84 metres), ranging from sea level to 5,082 feet (1,549 metres) at Mt. Lukens (also called Sister Elsie Peak). The Santa Monica Mountains, covering an area of 92 square miles and reaching heights of 3,000 feet, bisect the city, separating Hollywood, Beverly Hills, and Pacific Palisades from the southern boundary of the San Fernando Valley, a 220-square-mile area with such suburban communities as Burbank, Glendale, North Hollywood, Studio City, Sherman Oaks, Encino, Tarzana, Woodland Hills, and the mission city of San Fernando.

The valley's principal east–west artery, Ventura Boulevard, is a 17-mile bazaar of specialty shops, ethnic restaurants, banks, medical buildings, shopping malls, automobile agencies, and realtors' offices. The French writer-diplomat Romain Gary pronounced it "one of the most remarkable streets in the world." In the 1920s it was a dirt road. The post-World War II boom turned it into the main street of what would now be the country's sixth largest city if the valley were an independent entity.

Once the sanctuary for middle-class white families fleeing the city's congestion and racial tensions, the valley is now breaking up its rural estates to make room for condominiums and shopping centres. Walnut orchards and truck gardens have given way to housing for blacks, Hispanics, and Asians who have gone to work in new plants ranging from basic industry to high technology. Burbank, long the butt of television comedians ("Burbank has a low suicide rate, because living in Burbank makes suicide redundant"), now proclaims itself the country's entertainment centre. It is home for recording companies, the National Broadcasting Company, and three major motion-picture studios (Walt Disney Productions, Warner Bros., Inc., and Columbia Pictures Television Division).

Hollywood, eight miles northwest of the central city, was laid out in 1887 by Horace Wilcox, a Prohibitionist, who intended his subdivision to be a sober, God-fearing community. In 1910, when its water supply ran low, Hollywood was gobbled up by Los Angeles. The following year Blondeau Tavern, at the intersection of Sunset Boulevard and Gower Street, was turned into Hollywood's first motion-picture studio—to be abandoned 60 years later when Columbia Pictures moved to Burbank. By then the stars had long since left Hollywood, many of them moving into secluded hillside mansions above Beverly Hills, the most famous of which was Pickfair, built by Douglas Fairbanks for Mary Pickford in 1919. (After Pickford's death in 1979 it was sold to a sports entrepreneur.)

Climate. Coastal mountain ranges to the north and east act as buffers against extremes of summer heat and winter cold. Even in the hottest months the humidity tends to be mercifully low and the nights cool. "Night and morning low clouds" is the most common summer forecast, with the sun breaking through in the afternoon. Pronounced climatic differences occur in different sections of the city. The San Fernando Valley is generally several degrees cooler in winter and warmer in summer than communities on the opposite side of the Santa Monica Mountains. The city's mean temperature is about 64° F (18° C). The record high was 110° F (43° C) on September 1, 1955, the record low 27.9° F (−2.3° C) on January 4, 1949. The average annual rainfall is 14 inches (356 millimetres), with most of it falling in the winter months.

Environmental blessings and blights

Natural phenomena. *Smog.* Juan Cabrillo, California's first European explorer, reported "many smokes" in 1542 hanging over "a large bay," which was probably the future harbour of Los Angeles. Four centuries later, in September 1943, Angelinos had their first massive doses of air pollution. Southern California's highly publicized sunshine, they learned, was creating an inversion, a warm layer of air that clamped a lid on the cooler air of the saucer-like basin below.

The sun continues to cook the noxious vapours rising from factories and oil refineries, and, most of all, from the tail pipes of the county's millions of cars, vans, and trucks. After many years and the expenditure of billions of dollars, Angelinos have had to settle for measures de-



The city of Los Angeles and vicinity.

signed not to make the air really clean but to keep it from getting appreciably worse. No significant curtailment of smog can be achieved, they have been told by the head of their urban air-pollution management agency, "without a major overhaul of the industrial and transportation structure, severe limits on growth, and a radical change in life-style."

Earthquakes. The great San Andreas Fault is 33 miles from downtown Los Angeles at its closest point, but more than 40 known lesser faults crisscross the metropolitan area. No earthquakes of major intensity—*i.e.,* higher than 7 on the Richter scale of 10—have hit southern California since the 1850s, but aside from the periodic minor jolts that Angelinos take for granted, destructive temblors have struck Santa Barbara (1925), Long Beach-Compton (1933), and the San Fernando Valley (1971).

Winds. The city is occasionally buffeted by a Santa Ana, a hot, dry wind named for the canyon through which it often blows. Santa Anas occur when air rushes down from the high inland plateaus and is heated by compression. Angelinos are noticeably restless during a Santa Ana. The mystery-story writer Raymond Chandler once wrote, "Meek little wives feel the edge of the carving knife and study their husbands' necks."

Fires. Dry winds whipping through narrow canyons on hot days heighten the ever-present danger of fire in the brush-covered mountains. The city's most costly brushfire occurred on November 6, 1961, when 484 homes were destroyed in Bel Air, an affluent west side community. Hillside homeowners run the risk not only of fire but also of mud slides when winter rains pelt canyon walls denuded by flames.

THE PEOPLE

When the city celebrated its bicentenary in 1981, whites had become a minority, as they had been when the city was founded. In the 1980 census, which regards "Spanish origin" as an ethnic, not a racial group, whites made up roughly half of the population, Hispanics about one-fourth, and blacks one-sixth. Asians, Pacific Islanders, and American Indians were smaller minorities.

The ethnic minorities

Blacks. Although 26 of the city's 44 founders were of African ancestry and the granddaughter of a black tailor at one time owned the land now occupied by Beverly Hills, there were only 38,894 black Angelinos in 1931. Fifty years later, streams of black immigrants, mostly from the rural South, had raised the figure to 500,000. Housing patterns have consigned their children to the separate and unequal schools of de facto segregation. A long, bitter court battle to preserve the "neighbourhood school" ended in victory for the segregationists in 1981, sealing off the city's multiethnic children from one another at the very time that Angelinos were celebrating the 200th anniversary of the city's founding by black and brown parents and their 22 children.

Watts, the 2.5-square-mile core of the city's south central black ghetto, exploded on the hot summer night of August 11, 1965. The flames of Charcoal Alley gave whites a glimpse of the misery and anger of rural blacks trapped in an urban slum. After six days of looting and burning, 34 people were dead and 1,032 wounded, and 3,952 had been arrested. Fifteen years later blacks in Watts found themselves competing with Mexican-Americans for jobs and housing.

Mexican-Americans. Young Mexican-Americans, calling themselves Chicanos, grow up in gang-ridden *barrios* (neighbourhoods) where jobs are scarce, housing substandard, schools overcrowded, streets drug-infested, and welfare checks inadequate. Unlike their black contemporaries, who also roam dead-end streets, they have the additional handicap of speaking an alien tongue in the land of their forebears.

Los Angeles County contains nearly half of the state's Spanish-speaking population. Los Angeles has a Mexican metropolitan centre second in size only to that of Mexico City, and the *barrios* have been uniting to demonstrate their growing political strength.

Asians and Pacific Islanders. "We're fighting with the blacks and Chicanos for the same piece of the pie," a Chinese-American social worker pointed out in the late 1970s, when the city's burgeoning Asian and Pacific Island population was helping to resettle refugees from Vietnam, Laos, and Kampuchea (Cambodia). Little Tokyo, a few blocks from City Hall, has been the centre of the country's largest mainland concentration of Japanese-Americans since the early 1900s. During World War II some 40,000 southern Californians of Japanese ancestry were placed in isolated detention camps. For two decades following their release, Little Tokyo languished. By the 1980s it was once again bursting with life, redeveloped largely with capital from Japan.

THE ECONOMY

Once a land of vineyards, orange groves, and dairy farms, Los Angeles has become the nucleus of an industrialized urban area covering a 60-mile radius from the city's bustling new financial district. If this "Golden Circle" were a separate country, its gross national product would be among the highest in the world.

Agriculture. While its population rose almost 50 percent in the 1950s, Los Angeles County sacrificed 3,000 acres (1,200 hectares) of farmland a day to the bulldozer (freeways alone consumed about 40 acres per mile), and it ceased to be the nation's wealthiest agricultural county, a position it had held since 1910.

Industry. Before World War I, San Francisco was the state's manufacturing hub, but since the 1920s it has been far outstripped by Los Angeles. Major products include aerospace equipment, petroleum and refining, processed food, electronics, pharmaceuticals, glass, rubber, and cement.

Aviation. The first international air meet in the Western Hemisphere was held on the outskirts of Los Angeles in 1910. A dozen years later Donald W. Douglas was turning out an airplane a week in his Santa Monica plant. One-third of the nation's World War II military aircraft were built in the Los Angeles area. The city's preparation for the 1984 Olympics included a major refurbishing of its International Airport.

Oil. Los Angeles sits on one of its major resources, as Edward L. Doheny demonstrated in 1892 when he started selling local crude oil for industrial fuel. In 1921, after the automobile had created a ravenous appetite for petroleum products, the world's richest deposit in terms of barrels per acre turned up on Signal Hill. In its first half-century its 2,400 wells produced 859,000,000 barrels of oil.

Motion pictures. Two generations of moviegoers have looked up to Los Angeles as the world's film capital. It still plays a leading role in the motion-picture, television, radio, and recording industries, although in recent years some major studios have been disposing of their valuable real-estate holdings. The back lot of the 20th Century-Fox Film Corporation was converted to Century City, a multimillion-dollar "instant city." The filmmaking activities at Universal Studios in the San Fernando Valley have turned into a popular attraction for the area's millions of visitors each year.

Transportation. The vast majority of Angelinos working in the metropolitan area drive to their jobs in a car, van, or truck, while only a small percentage use public transportation. Ironically, much of the city's astonishing growth in the early 1900s was due to the superb interurban transit service provided by the big red electric cars of the railroad magnate and art collector Henry E. Huntington. The system became a freeway casualty in the early 1940s.

Demise of the urban transit system

The communications media. The city's first newspaper, the *Star,* began weekly publication in 1851. Thirty years later the *Los Angeles Times* published its first issue. Acquired the following year by Gen. Harrison Gray Otis, it became the bible for the city's boosters, conservative Republicans, and antilabour forces. When it passed into the hands of Otis' grandson, Otis Chandler, in the 1960s, it

James Randklev—Shostal Assoc.



The campus of the University of California at Los Angeles.

took a more liberal and worldly stance. It became a breakfast table institution; its chief competition in the 1980s came less from its afternoon rival, the Hearst Corporation's *Herald-Examiner,* than from the suburban press, including the San Fernando Valley's restyled *Daily News.* Three of the area's radio stations and one of its television stations serve the Spanish-speaking community. There is also a television channel for programs in Japanese, Chinese, and Korean.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The city's mayor and its 15 council members are elected to four-year terms, along with the city attorney, controller, and the seven-member board of education. The county is run by a five-member board of supervisors serving four-year terms. They preside over a jurisdictional jungle of overlapping city–county agencies.

Institutions of higher learning

Education. There are about 20 junior colleges in the county and five state colleges. The area's two oldest institutions of higher learning are the University of Southern California (USC; 1880) and Occidental College (1887). USC is noted especially for its schools of law, medicine, dentistry, engineering, and performing arts. The state-supported University of California at Los Angeles (UCLA; 1919) has a wide range of undergraduate and graduate offerings, with a particularly heavy commitment to the life and geophysical sciences and the arts. California Institute of Technology, founded in 1891 as Throop Polytechnic Institute, moved to its present location in Pasadena in 1910. Researchers there conduct probes of outer space in conjunction with the Jet Propulsion Laboratory and the Mt. Wilson Observatory.

RECREATION AND CULTURAL LIFE

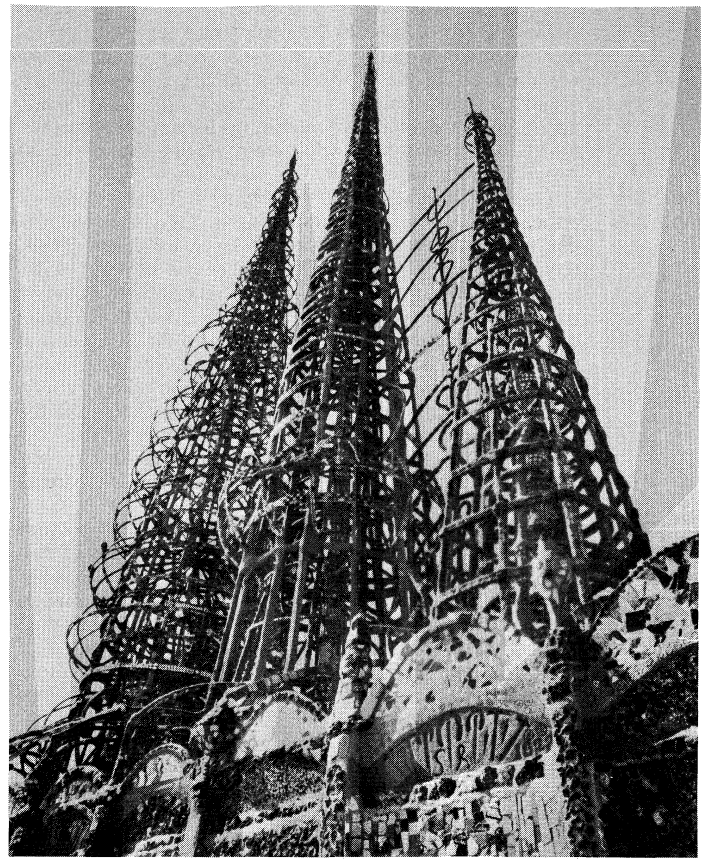
Parks. Hancock Park, the 23-acre site of the Rancho La Brea pits, was given to the county in 1916 by G. Allan Hancock, an oil magnate. Excavations between 1906 and 1913 of the black, bubbling pools had revealed a unique repository of fossilized skulls and bones of long-extinct mammals trapped in the seepage of *brea,* or "tar." In addition to the tar pits, the park features life-size figures of such creatures as the imperial mammoth, the American mastodon, the sabre-toothed tiger, the ground sloth, and the short-faced bear.

Griffith Park is spread across 4,064 rugged mountainous acres, an area larger than Beverly Hills. Mule deer wander down from the park's hillsides to peer at some of the 3,000 more exotic creatures in the city's zoo. The ranches of humorist Will Rogers in Pacific Palisades and silent film cowboy-actor William S. Hart in Newhall have been preserved as public parks. Within easy reach of the city by freeway are such commercial ventures as Disneyland, Knott's Berry Farm, Marineland, and Magic Mountain.

Sports. Intercollegiate athletics is highlighted by the intense USC–UCLA rivalry. The annual New Year's Day Rose Bowl football game is played in Pasadena. Angelinos also support professional major-league teams in baseball, football, basketball, and ice hockey. There is horse racing at Santa Anita Park and Hollywood Park.

Historic landmarks. Angelinos have held onto little of their Spanish-Mexican past, but largely through the efforts of Charles F. Lummis, a colourful turn-of-the-century editor and writer, the missions of San Gabriel Arcángel (1771) and San Fernando Rey de España (1797) have been preserved. Olvera Street, a narrow, block-long string of Mexican shops, cafés, and the Ávila family's adobe townhouse (*c.* 1818), has been a popular tourist attraction since its opening in 1930. It is now part of El Pueblo de Los Angeles State Historical Park, an area of some 40 acres, which includes the plaza where the city got its start and the first church its residents built (dedicated in 1822, rebuilt in 1861–62).

The three Towers of Simon Rodia (99, 97, and 55 feet high), better known as the Watts Towers, were built of broken tiles, dishes, bottles, and seashells over a 33-year period by Rodia, an unschooled Italian immigrant who later explained, "I had in mind to do something big, and I did." When his work was completed in 1954, he gave the property to a neighbour and left, never to return.



The Watts Towers by Simon Rodia, completed in 1954, in Los Angeles.
Tom McHugh—Photo Researchers

The city's turn-of-the-century Mission-style architecture gave way to the wood-sheathed California bungalow modelled on the work of Charles and Henry Greene, whose 8,000-square-foot Gamble House, built in Pasadena in 1908, is now used as a laboratory for student architects and designers. Frank Lloyd Wright received several commissions in southern California in the early 1920s following completion of Hollyhock House in what is now Barnsdall Park. R.M. Schindler helped supervise the project before setting out to build his own masterpieces. Irving Gill is generally regarded as the city's most neglected architect and Richard Neutra its most influential during the last half-century. His steel-frame Lovell House, built in the late 1920s, stands as a monument to the International Style.

One of the astonishing delights of downtown Los Angeles is the Bradbury Building (1893). Sunlight warms and illuminates the five-story inner court with its delicate French ironwork, Belgian marble, and Mexican tile. Union Passenger Terminal (1939), among the last of the country's railroad cathedrals, is a charming interpretation of Spanish Mission architecture. The city's Central Library (1926) was the last building designed by Bertram Goodhue, whose low, buff-coloured, stuccoed building recalls the work of Irving Gill, although the basic form was Beaux-Arts. It has survived its celebrated neighbour, the Richfield Building, a superb example of 1920 Art Deco, which disappeared in the late 1960s to make room for the twin skyscrapers of a bank and an oil company.

The arts. Los Angeles has from time to time sheltered outstanding composers and writers such as Igor Stravinsky, William Faulkner, F. Scott Fitzgerald, and Aldous Huxley. The city has been the subject of innumerable novels, the most durable of which appear to be Fitzgerald's *The Last Tycoon* (1941), Huxley's *After Many a Summer Dies the Swan* (1940), Evelyn Waugh's *The Loved One* (1948), and Nathanael West's *Day of the Locust* (1939), in addition to those by Raymond Chandler and his successor, Ross Macdonald (a pseudonym for Kenneth Millar). The city is also well known to the world's book collectors

for the works of such fine-printers as Saul Marks, Grant Dahlstrom, and Ward Ritchie, as well as for Jake Zeitlin, the dean of Los Angeles antiquarian booksellers, whose shops became gathering places for local and visiting writers, artists, photographers, and bibliophiles.

Los Angeles has developed a lively marketplace for works of art. Many of its leading galleries are scattered among the fine restaurants, antique shops, and rare book stores on La Cienega and Melrose boulevards in West Hollywood.

The performing arts

The performing arts are housed at the Music Center, which includes the elegant Dorothy Chandler Pavilion, home of the Los Angeles Philharmonic Orchestra; the Ahmanson Theatre, used for plays, musical comedies, and light operas; and the cake-shaped Mark Taper Forum, designed for experimental stage productions. The Music Center has no resident opera company, and it depended on New York City, London, Copenhagen, and Moscow for its ballet until 1983, when Robert Joffrey moved his company from New York City to the Dorothy Chandler Pavilion.

The Hollywood Bowl, a natural amphitheatre in the Hollywood Hills, offered its first production, *Julius Caesar*, on May 19, 1916, and initiated its programs known as Symphonies Under the Stars in 1922. The out-of-doors Greek Theatre in Griffith Park provides a wide range of summer entertainment, mostly musical. The Universal Amphitheatre in Universal City has offered year-round entertainment since it reopened in 1982 with a dome covering its seats. The Huntington Hartford Theatre is in Hollywood, near the site of the barn in which *The Squaw Man* was filmed in 1914. The Shubert Theatre in Century City is a few blocks from the massive 1880s New York set built for the motion-picture version of *Hello, Dolly!* (1969). Small, innovative theatres continue to spring up in the motion-picture capital.

Museums and libraries

The delightful California State Museum of Science and Industry is located in Exposition Park, not far from the county's Museum of Natural History, the fourth largest natural history museum in the United States. Its unique fossil collection has been transferred to the George C. Page Museum of La Brea Discoveries in Hancock Park. It shares the historic acreage with the Los Angeles County Museum of Art, which seems to float in the black pools of tar. In the early 1980s Arata Isozake, a Tokyo-based architect, was preparing plans for the Museum of Contemporary Art, to be located near the Music Center. The Southwest Museum in Highland Park, devoted to Native American art and artifacts, has one of the country's five richest collections of textiles, pots, baskets, photographs, and books in this field. A touch-and-play Children's Museum has enriched and enlivened the Civic Center, and a Junior Arts Center is part of the Barnsdall Park complex built around the Municipal Art Gallery.

Pasadena's Norton Simon Museum houses a notable art collection spanning 2,000 years. In nearby San Marino the Henry E. Huntington Library, Art Gallery, and Botanical Garden is a hospitable, parklike retreat, where visitors are invited to enjoy its roses, camellias, desert plants, and Japanese Garden as well as its wealth of rare books, manuscripts, and British art works. The J. Paul Getty Museum, perched above the beach at Malibu, is a replica of an ancient seaside villa. It became the world's richest museum in 1976 when it inherited $700,000,000 from its founder. In addition to local libraries in such cities as Pasadena, Santa Monica, Burbank, and Beverly Hills, Angelinos draw on the resources of the city's system and the county's regional and community libraries. UCLA has one of the country's finest academic libraries. The off-campus William Andrews Clark Memorial Library is noted for its John Dryden collection.

History

THE GROWTH OF THE METROPOLIS

The Spanish–Mexican town and city. On August 2, 1769, a Spanish expedition headed by Gaspar de Portolá, searching for mission sites, camped near a river they named the Porciúncula, in honour of Our Lady the Queen of the Angels of Porciúncula (Nuestra Señora la Reyna de los Angeles de Porciúncula). The area's first Europeans exchanged gifts with the peaceful, Uto-Aztecan-speaking Indians of the nearby village of Yang-na and left on the following morning. Despite three earthquakes during his overnight stay, Father Juan Crespi noted in his diary that "this delightful place among the trees on the river" had "all the requisites for a large settlement."

The early settlers

Two years later the Mission San Gabriel Arcángel was established about nine miles northeast of the campsite. A decade went by before Gov. Felipe de Neve succeeded in colonizing the fertile river basin with 44 recruits from Mexico, most of them of Indian and African blood. The illiterate settlers assembled on the west bank of what is now the Los Angeles River on September 4, 1781, to claim the land they had been promised. Little is known about the events of that momentous day, but mythmakers have cloaked the city's founding in a ceremonial splendour worthy of its destiny and its high-sounding name, which has long been confused with the name given the river. It is now generally agreed that the city's correct name is El Pueblo de la Reyna de Los Angeles (The Town of the Queen of the Angels).

El Pueblo, as it was commonly called, remained so isolated from the United States during its formative years that Joseph Chapman, the first Yankee to become an Angelino, was thought of as an Englishman ("El Inglés"). An engaging pirate from Boston, Chapman landed in 1818 with a black fellow privateer, Thomas Fisher. The first outsider to arrive by way of the arduous overland route was a fur trapper, Jedediah Smith, who turned up in 1826, four years after an independent Mexico had hoisted its flag above El Pueblo.

Los Angeles, with a population of nearly 1,250, had become a *ciudad* (city) in 1835 when Richard Henry Dana looked in on it. "In the hands of an enterprising people," he mused in *Two Years Before the Mast* (1840), "what a country this might be." By the time war broke out between the United States and Mexico in 1846, the California capital was so overrun with enterprising people that Gov. Pío Pico felt helpless against "the hordes of Yankee immigrants," who were "cultivating farms, establishing vineyards, erecting mills, sawing up lumber, building workshops, and doing a thousand other things which seem natural to them, but which Californians neglect or despise." When American forces under Capt. John C. Frémont and Commo. Robert F. Stockton entered the city on August 13, 1846, not a shot was fired. A revolt was put down the following January, and on July 4, 1847, Los Angeles celebrated its first Independence Day.

The American city. Los Angeles was incorporated on April 4, 1850, and designated the seat of Los Angeles County. The lawless, adobe cow town—"gambling, drinking and whoring are the only occupations," grumbled a pioneer physician in 1849—prospered in the wake of the Gold Rush when hungry miners in San Francisco and Sacramento gorged on beef from southern California. A catastrophic drought (1862–65) following several years of declining cattle prices brought an end to the era of the ranchos. Vast Spanish and Mexican land grants, mortgaged and bankrupt, their owners ignorant of *Yanqui* laws and *Yanqui* interest rates, were broken up, fenced, and planted by a new breed of Angelino. By 1860 the city had become so Americanized that it had banned bullfighting and formed a baseball club.

With the arrival of the railroads (the Southern Pacific in 1876, the Santa Fe in 1885), Los Angeles began to ship its oranges back east and, by means of a massive advertising campaign, to lure immigrants westward to the New Eden. Aided by a railroad rate war, the boom of the 1880s more than quadrupled the city's population, from 11,183 in 1880 to 50,395 in 1890. Father Crespi's campsite was overrun by shrewd, aggressive Yankee boosters determined to build a city in their own image and requiring only two things denied them by a bountiful providence: a harbour and an adequate water supply.

The search for a port and water

Unlike San Francisco and San Diego, Los Angeles has no natural harbour. A narrow, artfully gerrymandered "shoestring strip" connects the inland city with its port, 23 miles south of City Hall. Los Angeles acquired its two

harbour communities, San Pedro and Wilmington, by consolidation in 1909 following an epic struggle between the powerful Southern Pacific Railroad interests and the Free Harbor League. Work began on the harbour in 1899. The opening of the first municipal wharf in 1914 coincided with the completion of the Panama Canal, which put the ports of the Atlantic seaboard some 8,000 miles closer to Los Angeles. Its harbour became the busiest on the West Coast.

In 1904, casting about for new sources of water to sustain the city's relentless growth, William Mulholland, water-bureau superintendent, explored the Owens Valley some 250 miles northeast of Los Angeles and returned with a bold plan for an aqueduct to carry melted snow from the southern slopes of the Sierra Nevada to Los Angeles faucets. The plan outraged Owens Valley ranchers, enriched two syndicates of Los Angeles speculators, gave rise to rumours of wrongdoing and, highly fictionalized, ended up on the motion-picture screen as *Chinatown* (1974).

"There it is; take it," Mulholland told the thousands of Angelinos who assembled on November 5, 1913, to watch the Owens River water come cascading into a San Fernando Valley spillway. The 233-mile-long aqueduct, with 142 separate tunnels totalling 52 miles in length, has been supplemented by a 105-mile extension into the Mono Basin. The system supplies 80 percent of the city's water needs. The remainder comes from local wells, the California Aqueduct, and the Colorado River.

A severe drought in 1976–77 gave Angelinos a foretaste of the time (1985) when their Colorado River water would, by court order, be shared with Arizona. Meanwhile, with rain and snow dumping about three-fourths of California's usable surface water in the north, while more than half of its people live in the south, the "water wars" between powerful corporate farmers and resourceful urban environmentalists continue to be fought out in the legislature, the courts, and the voting booth.

THE MODERN CITY

In the first decade of the 20th century, while San Francisco tidied up the rubble of its 1906 earthquake, Los Angeles tripled its population, from about 100,000 to nearly 320,000. One local entrepreneur opened the first motion-picture theatre in the United States, in 1902, and another built the city's first garage to accommodate its growing number of automobiles, but the parasol-shaded girls from the red-plush brothels run by Pearl Morton and Cora Phillips still drove about in open carriages, much to the distress of the retired druggists, dentists, and wheat farmers from the Middle West who kept streaming into Los Angeles.

Emergence of the bizarre in the city

"Virtue has become virulent," reported a writer in *Smart Set* (March 1913), who described an "overgrown village" swarming with "spiritualists, mediums, astrologists, phrenologists, palmists and all other breeds of esoteric windjammers." Successive waves of migratory writers, taking much the same tack, have stereotyped the city as an open-air institution for the eccentric.

"Los Angeles represents the ultimate segregation of the unfit," Bertrand Russell declared, and Frank Lloyd Wright agreed. "It is as if you tipped the United States up and all the commonplace people slid down there into Southern California," he remarked in 1940, but within the next few years Los Angeles had become so overrun with such gifted European refugees as Heinrich and Thomas Mann, Arnold Schoenberg, Bertolt Brecht, Bruno Walter, Franz Werfel, Lion Feuchtwanger, and Alfred Neumann that the city had come to be dubbed "The Fourth Weimar Republic."

Angelinos live in enclaves walled off from one another by ethnic, cultural, and economic differences. They go their own way, surfing, riding, skiing, yachting, hiking, playing golf and tennis. Nowhere in the world is the pursuit of happiness more unabashedly hedonistic, and, perhaps, no city in modern times has been so universally envied, imitated, ridiculed, and, because of what it may portend, feared. As early as 1927 it was recognized by Bruce Bliven as "a melting pot in which the civilization of the future may be seen bubbling darkly up in a foreshadowing brew."

BIBLIOGRAPHY. There is no single comprehensive book on Los Angeles, but for an introduction to its history see JOHN and LAREE CAUGHEY (eds.), *Los Angeles: Biography of a City* (1976); and JOHN D. WEAVER, *Los Angeles: The Enormous Village 1781–1981* (1980). Three standard works, although dated, are still indispensable: JAMES M. GUINN, *A History of California and an Extended History of Los Angeles and Environs,* 3 vol. (1915); CAREY MCWILLIAMS, *Southern California Country: An Island on the Land* (1946; reissued as *Southern California: An Island on the Land,* 1973); and WRITERS' PROGRAM, CALIFORNIA (comp.), *Los Angeles: A Guide to the City and Its Environs,* 2nd ed. rev. (1951; 1941 ed. reissued 1972). The first 40 years as a city are covered in two classic studies: ROBERT GLASS CLELAND, *The Cattle on a Thousand Hills: Southern California, 1850–1880,* 2nd ed. (1951, reissued 1969); and GLENN S. DUMKE, *The Boom of the Eighties in Southern California* (1944, reissued 1970).

Three early-day Angelinos have left invaluable recollections: HORACE BELL, *Reminiscences of a Ranger* (1881, reissued 1965–67), which should be read with caution; HARRIS NEWMARK, *Sixty Years in Southern California, 1853–1913,* 4th ed. rev. (1970); and BOYLE WORKMAN, *The City That Grew* (1936). One of the most useful records of pioneer days is JOHN ALBERT WILSON's compendium, *History of Los Angeles County, California* (1880; reissued as *Reproduction of Thompson and West's History of Los Angeles County, California,* 1959). WILLIAM L. KAHRL, *Water and Power: The Conflict over Los Angeles' Water Supply in the Owens Valley* (1982); and ABRAHAM HOFFMAN, *Vision or Villainy: Origins of the Owens Valley–Los Angeles Water Controversy* (1981), are excellent studies of this key subject. REYNER BANHAM, *Los Angeles: The Architecture of Four Ecologies* (1971), is the work of a perceptive English visitor. DAVID BRODSLY is the first child of the freeway generation to make a serious appraisal of his off-ramp heritage, in his *L.A. Freeways: An Appreciative Essay* (1981).

(J.D.W.)

The Low Countries

The Low Countries—also called the Benelux countries from the initial letters of their names (Belgium, The Netherlands, and Luxembourg)—are in west-central Europe, bordering the North Sea on the west and north, West Germany on the east, and France on the south. Enjoying a favoured position on the North Sea coast and interlaced by such rivers as the Rhine, the Meuse, and the Scheldt, which are navigable by seagoing vessels deep into the interior, the Low Countries have long been economically important. The addition of a network of canals and other artificial waterways and the establishment of a system of dikes and pumps to hold back the seas and reclaim the land have provided an impressive history of human control of the environment.

In 1944, during World War II, the governments-in-exile of Belgium, The Netherlands, and Luxembourg signed an agreement to establish a postwar customs union, which became operative in 1948. Progressive abolishment of economic barriers led to the signing in 1958 of the comprehensive Treaty of the Benelux Economic Union, which took effect on November 1, 1960.

The article is divided into the following sections:

Physical and human geography 313
History of the Low Countries (to 1579) 313
  Prehistory 314
    Paleolithic remains
    Mesolithic food gatherers
    Danubian farmers
    Spread of Neolithic life
    The Bronze Age (1600–600 BC)
    The Iron Age
  Ancient and early medieval times 314
    The Roman period
    Frankish rule
    The development of the territorial principalities and the rise of the towns (925–*c.* 1350)
  Burgundian and Habsburg rule (1384–1579) 321
    Constitutional developments
    The Burgundians
    The Habsburgs
    The revolt and the formation of the republic (1567–79)
The Netherlands 325
  Physical and human geography 326
    The land
    The people
    The economy
    Administrative and social conditions
    Cultural life
  History 332
    The Union of Utrecht
    Foreign intervention
    The formation of a new government
    Ascendency of the Dutch economy
    The Twelve Years' Truce
    War with Spain, 1621–48
    The first stadholderless period
    William III
    Dutch civilization in the Golden Age (1609–1713)
    The 18th century
    The period of French rule (1795–1813)
    The Kingdom of The Netherlands (1814–1918)
    The Netherlands since 1918
Belgium 342
  Physical and human geography 343
    The land
    The people
    The economy
    Administrative and social conditions
    Cultural life
  History 349
    The Spanish Netherlands (1579–1713)
    The Austrian Netherlands (1713–1795)
    French rule
    Kingdom of The Netherlands
    Independent Belgium
Luxembourg 355
  Physical and human geography 355
    The land
    The people
    The economy
    Administrative and social conditions
    Cultural life
  History 356
    Ancient and medieval periods
    Habsburg and French rule
    Personal union with The Netherlands
    Independent Luxembourg
Major former principalities 357
  Brabant 357
  Flanders 358
    The counts of Flanders in the Middle Ages
    The Burgundians and the Habsburgs
    Institutions
    Economic life
    Arts and sciences
  Frisia 360
  Holland 360
    First line of counts
    House of Avesnes and Wittelsbachs
    Burgundian dynasty
    Habsburgs and House of Orange
    Member of the United Provinces
    French regime and Kingdom of The Netherlands
  Limburg 362

Physical and human geography

The name Low Countries is more a political or economic rather than a strictly geographic concept. It is true that, along the coast behind an irregular band of dunes and along a series of estuaries, there is a broad area of flat country mostly below sea level—a country of fens, or polders, stretching from Friesland in the north to Zeeland and Flanders in the south. Elsewhere, however, the land of the Low Countries is not so low or flat. Farther south and east in Belgium and Luxembourg and in the Limburg tip of The Netherlands, the relief becomes one of foothills; frequently they are forested. Some of these highlands form part of the historic plateaus of the Ardennes and Lorraine, which are northern reaches of the central European mountain ranges that continue eastward and southward into Germany and France.

The climate of the lowlands is maritime. Winters are damp and relatively mild, with frequent fogs. The hilly regions, because of their position and altitude (from 1,000 to 2,000 feet [300 to 600 metres], or more), have a climate akin to continental conditions. Southeastern Belgium and Luxembourg can have relatively severe winters and a greater precipitation than the lowlands.

Ethnically, the Low Countries form a transitional zone between Germany and France, between the Germanic languages on one side and the Romance languages on the other. Dutch and its Flemish dialects are spoken north of a latitudinal line running more or less regularly from Verviers in the east, through Waterloo, to the coast at Dunkirk. French and its Walloon dialects are spoken south of this line. German and its dialects are found in the frontier communes of the east and in Luxembourg.

History of the Low Countries (to 1579)

For historical purposes, the name Low Countries is generally understood to include the territory of the Kingdom

of The Netherlands and what is now Belgium as well as parts of northeastern France and Luxembourg. Belgium, however, though it was not constituted as an independent kingdom until 1831, had a distinct history of its own from 1579, when the southern provinces of the Low Countries began to be separated from the northern. Here, therefore, the history of the Low Countries will be surveyed as a whole to 1579, or slightly beyond. The histories of the separate Netherlands, Belgium, and Luxembourg are treated individually in the sections on those countries, below.

PREHISTORY

Paleolithic remains. The oldest remains of human industry in the Low Countries are Clactonian and Acheulean flints from gravels in the Mons (Hainaut) region of Belgium dating from the third (Riss) glaciation of the Pleistocene Epoch. Middle Paleolithic Mousterian industry is very well represented in Belgian alluvial and cave deposits (Hainaut, Liège, and Brabant). Human (Neanderthal) fossils have also been found, especially in the cave at Spy. The Upper Paleolithic inhabitants (*Homo sapiens*) of the northern provinces of the Netherlands were reindeer hunters, fishermen, and food gatherers, who appeared in the final phase of the last glaciation (*c.* 10,000 BC). They left flint implements closely related to the Hamburgian industry of northwest Germany. The Upper Paleolithic cultures in Belgium represent the northwesternmost limit of the classic Aurignacian, Perigordian, and Magdalenian industries as known in France, Spain, Switzerland, and southern Germany. The Low Countries have no Upper Paleolithic cave art; but some objects and stone slabs are decorated in typical French style.

During the final Paleolithic phase (*c.* 9000–8000 BC) there were three cultures in the Low Countries, of which two appeared both in the Netherlands and in Belgium—namely the Tjonger group (so called from the Tjonger river in Frisia [Friesland]); the Ahrensburgian, which derives from the Hamburgian; and, in southern Belgium only, the epi-Magdalenian.

Mesolithic food gatherers. During Mesolithic times (8000–*c.* 4000 BC), three groups of cultures are found. The first, the Maglemosian people, lived near lakes and rivers in the Netherlands, Flanders, and most of the north European plain. Their cultural remains consist largely of bone and antler points and harpoons. The second group, the Tardenoisians, lived on high plateaus and in sandy regions throughout the Low Countries. They were hunters of small game, and their remains consist of arrowheads and parts of composite implements made of geometrical microliths. Lastly there were the people who lived in the forests, whose characteristic remains are heavy tools suitable for woodworking, such as axes and adzes. These have been found in Frisia (northern Netherlands) and around Liège (Belgium).

Danubian farmers. Farming people from central Europe settled in the Dutch province of Limburg and on the Hesbaye plateau (Belgium) at the end of the 5th millennium BC. They brought the Neolithic way of life, based on agriculture and stock breeding. Excavation has shown that their village consisted of large, wooden longhouses, stables, and barns. They cultivated cereals on loess soil cleared of forest by the slash-and-burn method and bred pigs, sheep, and cattle. Among their most typical remains are shards of pottery decorated with wavy and angular incised lines and adzes made of polished hard or volcanic rock.

Spread of Neolithic life. After a few centuries the Danubians departed and were replaced by a second wave of central European immigrants whose arrival originated two civilizations: the Funnel-neck Beaker culture in the Netherlands and the Michelsberg in Belgium. Both date from the 3rd millennium BC. The Belgian Michelsberg culture is much the same as the German. Among its most important remains are the flint mines of the Mons region (Spiennes), the products of which were used over a wide area. Its most characteristic feature is its pottery, with tulip beakers and flat baking disks. The remains of the Funnel-neck Beaker culture in the Netherlands are found mainly in the eastern provinces, especially Drenthe. There are numerous large megalithic passage graves; and pottery remains consist of funnel beakers, collared flasks, buckets, and bowls decorated with deep horizontal and vertical grooves forming geometrical patterns. Excavations along the North Sea coasts have revealed a series of farmers' and fishermen's settlements. These seem, however, to belong to a Mesolithic population that had adopted the Neolithic way of life; their remains show strong Funnel-neck Beaker and western Neolithic influences.

The end of the Neolithic period in the Netherlands is marked by a series of invasions and migrations. Warlike herdsmen arrived in the Netherlands from the east about 2200 BC. Avoiding the Funnel-neck Beaker area, they settled in the Veluwe (Guelders) and at the western limit of the northern sandy areas. They buried their dead in single graves under barrows and left battle-axes of hard stone and cord-impressed beakers. About 2000 BC groups of prospectors of the Bell-Beaker civilization arrived, settling along the Belgian and Dutch coasts and in the provinces of Gelderland (Guelders) and Drenthe in the eastern Netherlands. They introduced the first metal objects. Some of them travelled on to the British Isles.

The Bronze Age (1600–600 BC). This was a quiet period for the Low Countries because they lay outside the main trade routes. The period is chiefly memorable for its complicated timber-structured grave barrows. At the end of the Early Bronze Age, people came from the British Isles to settle in the northern regions of Belgium and the Netherlands. They cremated their dead and put the bones into urns, which they placed under barrows.

During the Late Bronze Age, bronze weapons and implements began to be imported into the Low Countries from many parts of central and northern Europe, including the British Isles. Urnfield people (probably Celtic), so called from their custom of putting the cremated bones of their dead into urns, came from Switzerland and the Rhineland to settle in northern Belgium and the southern Netherlands—*i.e.*, south of the Meuse, Waal, and Rhine rivers. Their equipment and grave goods were poor. In Flanders and central Belgium the genuine Urnfield tradition of flat graves was kept, while toward the north barrows were adopted from the autochthonous population. North of the rivers came a large Germanic Urnfield immigration originating from northwest Germany.

The Iron Age. This period (600 BC to the Roman conquest, 12 BC onward) was very quiet north of the three rivers; the population of agriculturists and herdsmen continued the traditions of Urnfield times. About the beginning of the 3rd century BC the lower coastal stretch of Frisia was inhabited for the first time, by a population of cattle breeders. In order to protect themselves and their cattle from the high tidal floods, these new people built their villages on artificial mounds, or terps. The systematically excavated terp of Ezinge has yielded much information about their life. The area south of the rivers, however, presents a different picture. In the 7th century BC a group of Celtic (Hallstatt) warriors, probably from Bavaria, conquered central Belgium and established their rule over the autochthonous population. The Hallstatt influence was also felt in northern Belgium and the southern Netherlands; here there are a number of graves of chieftains containing rich grave goods, including golden trinkets, mostly imported from the Mediterranean. During the La Tène period there were three cultural groups south of the great rivers. In the northernmost, the Urnfield traditions were still carried on, but the pottery was strongly influenced by the French Marne civilization. The strongest La Tène influence shows in the pottery, metal objects, and gold ornaments of Hainaut. The Ardennes formed part of the Eifel-Hunsrück civilization. These last people built numerous hill forts (*oppida*) surrounded by stone and earthen walls.

(S.J.de L.)

ANCIENT AND EARLY MEDIEVAL TIMES

The Roman period. At the time of the Roman conquest (1st century BC), the Low Countries were inhabited by a number of Celtic tribes to the south and west of the Rhine and by a number of Germanic tribes to the north. Cultural and ethnic influences in both directions, however,

make it difficult to draw the line between Celtic and Germanic peoples. On the coast of northern France and in Flanders lived the Morini; to the north of them, between the Scheldt River and the sea, the Menapii; in Artois, the Nervii; between the Scheldt and the Rhine, the Eburones and the Aduatuci; and, in what is now Luxembourg, the Treveri. North of the Rhine, the Frisii (Frisians) were the principal inhabitants, although the arrival of the Romans brought about a number of movements: the Batavi came to the area of the lower reaches of the Rhine, the Canninefates to the western coastal area of the mouth of the Rhine, the Marsaci to the islands of Zeeland, the Toxandri to the Campine (Kempenland), the Cugerni to the Xanten district, and the Tungri to part of the area originally inhabited by the Eburones (who were wiped out by Julius Caesar).

The Roman conquest of Gaul, which was completed by Caesar in 59–52 BC, stopped short at the Rhine. The emperor Augustus' attempt to extend Roman military power over the Elbe failed, and the area occupied by the Frisians, north of the Rhine, was therefore never under Roman rule. In the Rhine delta and to the south and west of the Rhine, the Romans set up the same administrative organizations as those found in other parts of Gaul. The Low Countries formed part of the provinces of Belgica and Germania Inferior (later Belgica Secunda and Germania Secunda), which themselves were subdivided into *civitates:* in Belgica, those of the Morini, Menapii, Treveri, Tungri, and possibly the Toxandri; in Germania Inferior, those of the Batavi, Canninefates, and Cugerni. Because of the later adoption by the church of the division into *civitates,* a number of centres of the *civitates* have become the seats of bishoprics; among these are Thérouanne, Tournai, Tongeren (Tongres), and Trier (Trèves).

In the northern regions, fishing and cattle raising were the principal means of support, while to the south of the Rhine it was possible, with Roman protection, to engage in farming in the Gallo-Roman fashion, using villas as centres. There must also have been a certain amount of trading and other traffic, even if only to provision the Roman military and civil services.

Decline of Roman power

In the mid-3rd century AD, Roman power in the Low Countries began to weaken, and the forts were abandoned. This was the result not only of a resurgence of the Germanic tribes but also probably of the encroachment of the sea, which probably brought about a drastic change in the area's economy. A temporary recovery began at the end of the 3rd century. In particular, Julian, caesar of Gaul, waged several wars in the Low Countries between 355 and 360 and was able to put new strength, for a time, into the Rhine border. A great invasion by Germanic tribes in 406–407, however, ended the Roman occupation of the Low Countries. The Romans had already tolerated the Germanic penetration of their territory and had given some tribes the task of protecting the borders of the empire. The Franks, who had settled in Toxandria, in Brabant, were given the job of defending the border areas, which they continued to do until the middle of the 5th century.

The Franks were probably influenced considerably by Roman culture, becoming familiar with the Roman world and way of life, although the expansion of their own race and their growing self-confidence were barriers to complete romanization. Around 450 they moved southward, founding a new Frankish kingdom centred on the region between the Somme and the Loire rivers. The number of Franks in this Merovingian-Frankish kingdom has been vigorously debated: French and Belgian scholars believe that only a small number of ruling families made the move; German scholars claim a mass migration of the whole tribe. It is certain, however, that the Franks never achieved a majority in the population of the kingdom but probably settled in the conquered area as colonists and farmers. Their cultural influence was too slight to oust the Romance languages; thus, the linguistic frontier between the Romance and Germanic languages has remained the same as the limit of Roman culture in the north—crossing the Low Countries, in particular the later principalities of Flanders, Brabant, and Liège, in an east–west line.

Frankish rule. The first kings of the Merovingian Franks—Childeric (died 481/482) and Clovis (ruled 481/482–511)—ruled the regions of Tournai and Cambrai before the great conquest of Gaul had even begun. Despite this, however, the Low Countries originally formed a border region in which the might of the Frankish state was hardly felt. The area occupied by the Frisians in the north was completely outside the Frankish sphere of influence, but the Rhine delta and even what is now Noord-Brabant also appear to have retained the virtually independent status they had possessed during the Roman era. It is possible, though not certain, that Nimwegen (Nijmegen) was a Merovingian outpost.

Resistance to the Franks

The area occupied by the Frisians was part of a North Sea culture that formed a distinct foil to Frankish power. The Frisians played an important role in trade, which sought routes along the Rhine and the Meuse and across the North Sea. Industrial products were imported from northern France, the Meuse plain, and the Rhineland, where Merovingian power was more firmly established and where centres of commerce (*e.g.,* Dinant, Namur, Huy, and Liège) developed. The more or less independent area on the North Sea coast, however, found itself threatened during the 7th century by the rise of the Frankish nobles. In particular, the family of the Pepins, who came from the centre of Austrasia (the Ardennes and upper Meuse), was able to secure land in Limburg. Moreover, encouraged by the Frankish king Dagobert (ruled 623–639), the Frankish church began an offensive that led to the foundation of the bishopric of Thérouanne (the *civitas* of the Morini) and of several monasteries.

This collaboration between church and nobles prepared the way for an expansion of political power to the north, which was carried out under the leadership of the Pepins, who as "mayors of the palace" in Austrasia had virtually taken over power from the weakened Merovingian kings. After defeating the mayor of the palace of Neustria (the western part of the Frankish Merovingian kingdom) at Tertry (687), Pepin II turned to the north, where he won a victory over the Frisian king Radbod near Dorestad (689). The Anglo-Saxon Willibrord was then ordained archbishop of the Frisians in Rome and, with Pepin's help, was given Utrecht as see of the bishopric and centre for his missionary activities. Even so, Frankish power expanded northward only gradually, making use of such bases as Utrecht and several former Roman forts. Charles Martel, a bastard son of Pepin II, who managed after several years' fighting (714–719) to grasp supreme power over the whole Frankish Empire, succeeded in 734 in forcing his way through to the northern centres of the Frisians and gaining a victory near the River Boorne. His victory was later consolidated by Pepin III (who in 751 became King Pepin I) and Charlemagne (ruled 768–814), his son. The whole area of the Low Countries thus effectively formed part of the Frankish Empire, which was then ruled by the Pepin, or Carolingian, dynasty.

Life under the Franks

Government. The administrative organization of the Low Countries during this period was basically the same as that of the rest of the Frankish Empire. Supreme authority was held by the king, who, aided by servants of the palace, toured the country incessantly. The Carolingian kings naturally made several visits to the Low Countries, where they had old palaces or built new ones (Herstal, Meerssen, Nijmegen, Aix-la-Chapelle) and where they also possessed extensive crown estates. Their authority (*bannus*) was delegated to counts who had control of counties, or *gauen* (*pagi*), some of which corresponded to Roman *civitates.* Among these counties in the Low Countries were the *pagus* Taruanensis (centred on Thérouanne), *pagus* Mempiscus, *pagus* Flandrensis (around Bruges), *pagus* Turnacensis (around Tournai), *pagus* Gandensis (Ghent), *pagus* Bracbatensis (between the Scheldt and the Dijle rivers), *pagus* Toxandrie (modern Noord-Brabant), and, north of the great rivers, Marssum, Lake et Isla, Teisterbant, Circa oras Rheni, Kinnem, Westflinge, Texla, Salon, Hamaland, and Twente. In the north, however, it is frequently not possible to determine with certainty whether the word *gau* in fact denoted a region controlled by a count who exercised the king's authority or indicated simply a region of land without reference to its government. Smaller admin-

istrative units were the *centenae,* or hundreds, and districts called *ambachten.* These last were mainly in what are now the provinces of Vlaanderen, Zeeland, and Holland.

Religion. The conversion to Christianity of the southern Low Countries, which took place largely during the 7th century, led to the foundation of further bishoprics at Artois, Tournai, and Cambrai, which were part of the ecclesiastical province of Rheims (the former Roman province of Belgica Secunda). Germania Secunda contained the ecclesiastical province of Cologne, in which the *civitas* of Tongres seems to have had an uninterrupted existence as a bishopric since Roman times; the centre of this bishopric was moved for a time to Maastricht (6th and 7th centuries) until, in about 720, Liège became the seat of the bishopric. Christianity was brought to the north of the Low Countries mainly by Anglo-Saxon preachers, by Frisians influenced by them, and by Franks. This Anglo-Saxon Christianity was particularly important in the missionary bishopric of Utrecht, which at first, because of its missionary character, had no precisely defined borders. True, the city of Utrecht had been named as the see of the bishopric, but, as in England, the monasteries played an important part in the missionary work; among these was the monastery of Echternach in Luxembourg. The country between the Meuse and the Waal rivers and the area around Nijmegen belonged to the bishopric of Cologne, while certain districts in the north and east were part of the bishopric of Münster (founded by Charlemagne). The bishoprics were subdivided into relatively large parishes, about the foundation of which only scanty information exists. With an increase in population during the 10th, 11th, and 12th centuries, the parishes were split into smaller ones, based on small chapels that at first had only limited rights. In the north and, above all, in Friesland, the initiative for founding parishes often came from the faithful themselves, who built a church and provided it with the necessary income in the form of estates (*dos*). The influence of laymen and, in particular, of the aristocracy was often extremely strong; and they usually collected the tithes and other revenues. On the other hand, this system of proprietary churches formed a distinct stimulus for the conversion of the country areas.

Justice. The administration of justice was the duty of the king and his appointed counts. They pronounced judgment principally in important cases (*causae majores*) or in cases involving prominent people: theirs was known as the higher court of justice. To fulfill this task, Charlemagne appointed groups of permanent judges, called *scabini* (French *échevins*), who under the direction of the count "found" the law and applied it as a verdict. At lower levels, it seems that justice must have been administered in the smaller legal units of hundreds and *ambachten.* Only to a small extent is the content of the law included in the codes of Germanic law, such as the Lex Salica, which consist largely of criminal law (robbery, theft, homicide, and rape) and property law (land and cattle). On Charlemagne's initiative, a large part of Frisian law was put down in writing as the Lex Frisionum, as well as the *ewa quae se ad Amorem habet,* "the law that exists on the Amor" (Amor being the name of an unidentified river, probably in the Waal and Meuse district). Charlemagne amplified the Lex Salica and Lex Saxonum in various of his *capitularia* (ordinances). The law was thus materially diverse, while formally the situation was complicated even further by the immunity that was frequently granted to churches or monasteries, which put their property beyond the reach of the legal and administrative powers of the count and his subordinates. The secular jurisdiction that in this way became the prerogative of bishops or abbots was the basis of later, more extensive secular powers, which in the Low Countries, more than anywhere else, were to play an important role in the foundation of the ecclesiastical territorial principalities.

The people

Social classes. The social structure of the Low Countries in the Frankish era included a number of classes. At the top was an elite that probably already operated on a hereditary system and of which the members were bound to the king as vassals and rewarded by fiefs (*beneficia*). Next were the freemen (*liberi, ingenui*), bound to the king by an oath of allegiance and traditionally under an obligation to serve in the army and in the law courts. A freeman's *Wergeld*—the sum that had to be paid to his family if he was killed—was in principle 200 shillings (*solidi*); but the *ingenui Franci,* or *homines Franci* (found in the region of the great rivers; probably descended from native nobles who had early placed themselves in the service of the Franks in their policy of conquest), had a much higher *Wergeld.* At the bottom of the ladder were the bondsmen, who were closely dependent on a lord (often an important landowner), in whose service they stood, in most cases working on his estates. It may be supposed that the position of the bondsmen was relatively favourable in the coastal areas of Holland and Friesland, where there were no large estates and, moreover, where the struggle against the sea required as much manpower as the community was able to offer.

Economy. Economically, the structure of the Low Countries in the Frankish period was principally agrarian. Particularly in the south and east, it was common practice to exploit the land from a central farmhouse (*villa,* or *curtis*), using the services of dependent subjects (bondsmen), who were duty bound to work on the domain of the lord and to this end received small farms from him. The nature of the land in the west and north, however, probably to a large extent precluded this classical type of exploitation of the domains; there was scattered, even fragmentary, ownership of land, and the *curtis* was no more than a gathering place to which the bondsmen had to take a part of their produce. In Holland and Friesland, fishing and the raising and selling of cattle were of importance. This Frisian trade, of which Dorestad was a centre, was greatly stimulated by absorption into the Frankish Empire, and it reached its zenith under Charlemagne and Louis the Pious (ruled 814–840). Moreover, by virtue of its becoming part of the Frankish Empire, Friesland obtained an important hinterland in the southern regions of the Meuse and Rhine and was thus in a position to develop export and through trade to Denmark, Norway, and the Baltic countries. The importance of Frisian trade may be seen in the Carolingian coins found in Dorestad, where there was a toll and a royal mint. This trade was supplied by the southern Low Countries. Thus the cloths that were sold as Frisian cloths were produced in the area of the Scheldt (later called Flanders). Quentovic, at the mouth of the Canche, was another trading centre; it too had a toll and a mint. Smaller trade settlements (*portus,* or *vicus*) emerged at Tournai, Ghent, Bruges, Antwerp, Dinant, Namur, Huy, Liège, and Maastricht—a clear indication of the commercial importance of the Scheldt and the Meuse.

Decline of the Frankish Empire. The great Carolingian Empire passed into a decline as early as the reign of Louis the Pious, and the process was accelerated after his death, in 840. Repeated wars broke out under his sons, leading eventually to the partition of the empire: by the Treaty of Verdun (843), the empire was divided into an eastern, western, and an extensive middle kingdom. The last of these, allotted to Louis's eldest son, Lothair, who also wore the imperial crown, was, after his death, further divided into three parts—the northern part, Lotharingia (Lorraine), which stretched from the Jura Mountains to the North Sea, falling to his son Lothair II, after whom it was named. After it had belonged alternatively to the western and eastern Frankish kingdoms, Lorraine eventually (925) became part of the eastern kingdom, by then the German kingdom, because there were no longer any ruling Carolingians. The nobles of Lorraine had themselves taken an important part in this changing political game. Giselbert, a member of one of the most noble families, had turned the scales in favour of joining the German kingdom, for which service King Henry I presented him with what then became the duchy of Lorraine.

Partition of the empire

The dissolution of the Carolingian Empire had also been helped by Viking, Magyar, and Saracen attacks—the Viking attacks being of greatest import for the Low Countries. The attacks had begun immediately after the death of Charlemagne (814) in the form of plundering raids, the magnitude and danger of which soon increased. (Dorestad, for example, was destroyed four times between

834 and 837.) Churches and monasteries, with their rich treasures, were the principal targets for the Vikings, who soon took to spending the winter in the Low Countries. In order to ward off the danger, attempts were made to throw up walls around towns and monasteries or even to drive off the Vikings by fierce counterattacks—a procedure that enjoyed some success—so that the counts of Flanders, for example, were able to lay a firm foundation for their own power. Another method of defense was to admit the Vikings on the condition that they defend the areas given them against other Vikings. The danger diminished after 900, though it was not until about 920 that Bishop Balderic risked the journey from Deventer to Utrecht, whence his predecessor had fled.

The development of the territorial principalities and the rise of the towns (925–c. 1350). Politically speaking, the period between 925 and about 1350 is characterized by the emergence, growth, and eventual independence of secular and ecclesiastical territorial principalities. The rulers of these principalities—both secular and spiritual—had a feudal relationship with the German king (the Holy Roman emperor), with the exception of the count of Flanders, who held his land principally as the vassal of the French king, with only the northern part of his county, Imperial Flanders, being held in fealty to the German king. While the secular principalities came into being as a result of individual initiative on the part of local rulers and of their taking the law into their own hands, to the detriment of the king's authority, the development of the spiritual princes' authority was systematically furthered and supported from above by the king himself. The secular principalities that arose in the Low Countries and whose borders were more or less fixed at the end of the 13th century were the counties of Flanders and Hainaut, the duchies of Brabant and Limburg (after 1288 joined in personal union), the county of Namur, the county of Loon (which was, however, to a large degree dependent on the bishopric of Liège), the county of Holland and Zeeland, and the county (after 1339, duchy) of Guelders. The Frisian areas (approximately corresponding to the modern provinces of Friesland and Groningen, excluding, however, the city of Groningen itself) had no sovereign authority. The spiritual principalities were Liège and Utrecht. The secular authority of the bishop of Utrecht was exercised over two separate areas: the Nedersticht (what is now the province of Utrecht) and the Oversticht (what are now the provinces of Overijssel and Drenthe and the city of Groningen).

Origin of the principalities

This article will sketch the emergence of these principalities, their relations with one another, their dissociation from the German king, and the increasing influence upon them of France and England. It must be noted that, although these principalities displayed common characteristics in their economies, social structures, and cultures, it was a pure coincidence—the intrusion of the Burgundian dynasty—that brought about a certain degree of political unity, which in turn furthered economic, social, and cultural unity and even led to the beginnings of a common national feeling (which was nevertheless too weak to prevent partition in 1579).

The secular principalities. The foundations upon which the secular princes built up their power were basically the following: In the first place, the count still exercised the rights that had for centuries been attached to the Carolingian office of count, denoted by the term *comitatus*. They included the administration of justice, various military powers, and the right to levy fines and tolls. To these rights fiefs were attached, which during the passage of time were expanded by the counts, who eventually owned such large estates that they were by far the greatest landowners in their territories. Soon the term *comitatus* covered not only the office, or duty, but also the whole area over which that office was exercised; thus it could be said that the count held his county in fief of the king. An important element of the count's authority was supervision over the county's religious foundations, especially the monasteries. In the 10th century, the counts sometimes even assumed the function of abbot (lay abbot); but they later contented themselves with the control of appointments to ecclesiastical offices, through which they often had great influence over the monasteries and profited from the income from monastic land. Thus, monasteries such as St. Vaast (near Arras), St. Amand (on the Scarpe), St. Bertin (near St. Omer), and St. Bavon and St. Pieter (in Ghent) became centres of the power and authority of the counts of Flanders; Nivelles and Gembloux, of the dukes of Brabant; and Egmond and Rijnsburg, of the counts of Holland.

At the end of the 9th and in the 10th centuries, during the Viking attacks and while connections with the empire were loosening, the local counts built up their power by joining a number of *pagi* together and building forts to ensure their safety. The counts of Flanders amalgamated the *pagi* Flandrensis, Rodanensis, Gandensis, Courtracensis, Iserae, and Mempiscus, the whole being thenceforth called Flanders; they fortified this area of their power with new or surviving Roman citadels. In the northern coastal regions, the Viking Gerulf was granted in about 885 the rights over a number of counties between the Meuse and the Vlie (Masalant, Kinnem, Texla, Westflinge, and a district known as Circa oras Rheni, which was, as the name implies, on both sides of the Rhine); his descendants consolidated their power there as counts of west Frisia and, after 1100, took the title of counts of Holland. In Brabant and Guelders, the amalgamation of fragmentary and dispersed estates took place later than in Flanders and Holland.

Expansion of counts' power

These secular princes were bound by a feudal relationship to the German or French kings, and they in turn used a similar relationship to bind the nobles to them. As supreme lords, they stood at the head of their vassals, who were duty bound to carry out services of *auxilium* and *consilium* (to serve in the army and at court). These services were rewarded with fiefs, and in most of the principalities a powerful feudal nobility emerged, to which the prince had to give the fullest consideration. At first, in the 10th and 11th centuries, the prince and his subjects had no direct link, and the princes had no specific duties to carry out, as bearers of authority, toward their subjects. On the contrary, the princes' activities during this period were almost exclusively motivated by greed for power and money.

During the 10th and 11th centuries, the German kings of the Saxon and Salian dynasties—particularly Otto I (ruled 936–973), Henry II (ruled 1002–24), Conrad II (ruled 1024–39), Henry III (ruled 1039–56), and Henry IV (ruled 1056–1106)—attempted to impose their authority on the increasingly powerful secular principalities by the appointment of dukes. In Lorraine, a duke was appointed to exercise imperial power over the local counts; but the plan failed, except for a short period (953–965) during the reign of Otto I, when the king's brother Bruno held the position (Bruno's success was largely due to the fact that he was also archbishop of Cologne). Taken all in all, the dukes of Lorraine had too little authority and no effective means to exercise their power to subdue the local counts. Lorraine was soon split into two dukedoms—Upper and Lower Lorraine. In Lower Lorraine, the title of duke was given to the counts of Louvain and the counts of Limburg—the former at first called themselves dukes of Lorraine but soon assumed the title of dukes of Brabant; the latter were known as the dukes of Limburg.

The spiritual principalities. That the German kings failed to integrate Lorraine into the Holy Roman Empire as a duchy ruled by a viceroy may be attributed to the fact that the kings soon developed another way to strengthen their power, not only in Lorraine but throughout the empire, by systematically investing bishops and abbots with secular powers and making them pillars of authority. This procedure, developed by Otto I and reaching its summit under Henry III, was carried out in phases and led eventually to the establishment of the Imperial Church (Reichskirche), in which the spiritual and secular principalities played an important part. The most important ecclesiastical principalities in the Low Countries were the bishoprics of Liège, Utrecht, and, to a lesser degree, Cambrai, which, though within the Holy Roman Empire, belonged to the French church province of Rheims. The secular powers enjoyed by these bishops were based on the right of immunity that their churches exercised over

Rise of bishops and abbots

their properties, and that meant that, within the areas of their properties, the counts and their subordinates had little or no opportunity to carry out their functions. This situation, which had existed during the Carolingian and Merovingian periods, enabled the bishops to exercise jurisdictional and military powers in these areas, helped by lay officials known as advocates (*advocati*), whom they themselves appointed. The German kings strengthened this immunity and put it to their own service by enlarging its jurisdictional powers while at the same time investing the bishops' advocates with the *bannus* (public authority) so that they formed a link between the king and the immunities. Next, the kings gave certain bishops jurisdictional authority in areas outside their own property, such as in the towns where they resided; at the same time, the bishops' properties were extended by gifts of crown land, which were then also covered by the powers of immunity. Further, the kings transferred such regal powers (regalia) to the bishops as the right to found markets, raise tolls, and strike coins. Through the exercise of these regalia, the bishops became for the first time bearers of royal authority and, thus, of the public authority; and they also derived from these rights a considerable personal income. The last phase in the bishops' growth occurred when the kings decided to transfer to the bishops the powers of counts in certain areas that were not covered by immunity (the power thus acquired is referred to as ban-immunity). Certain bishops, such as those of Liège and Utrecht, were able to combine their rights of immunity, certain jurisdictional powers, regalia, and ban-immunities into a unified secular authority, thus forming a secular principality called a *Sticht* (as distinct from the diocese) or—where the power structure was very large and complex, as in the case of the bishop of Liège—a prince-bishopric. As princes, the bishops were vassals of the king, having to fulfill military and advisory duties in the same way as their secular colleagues. The advantage of this system to the kings lay in the fact that the bishops could not start a dynasty that might begin to work for its own ends, and its smooth running stood and fell with the authority of the kings to nominate their own bishops—an authority that they maintained firmly.

Thus the spiritual-territorial principalities of the bishops of Liège and Utrecht emerged—the prince-bishopric of Liège and the *Sticht* of Utrecht. In Liège this development was completed in 972–1008 under the guidance of Bishop Notker, appointed by Otto I. As early as 985 he was granted the rights of the count of Huy, and the German kings made use of the bishopric of Liège to try to strengthen their positions in Lorraine. Utrecht, which lay more on the periphery of the empire, developed somewhat later; although the foundations of this power had already been laid in the 10th century, it was principally the kings Henry II, Conrad II, and Henry III who strengthened the secular power of the bishops through privileges and gifts of land.

Struggle for independence. Thus, the Low Countries during the 10th and 11th centuries saw the development of the pattern of a number of more or less independent feudal states, both secular and ecclesiastical, each of which was struggling for more freedom from the king's authority, the enlargement of its sphere of influence, and the strengthening of its internal power. Flanders led the way. In the 10th and 11th centuries it needed to pay only scant attention to the weak French kings of the Capetian dynasty and was thus soon able to exercise its power farther south—in Artois—and was even able to play an important part in a political power struggle around the French crown. In 1066 the Count of Flanders lent his support to the expedition to England of his son-in-law, William, duke of Normandy. The counts of Flanders built up a strong administrative apparatus—the *curia comitis,* based on central officials and on local rulers called burgraves, or castellans (*castellani*), who were in charge of districts known as castellanies, where they had extensive military and administrative powers. The reclamation of land from the sea and from marsh and wasteland in the coastal area, which began in the 11th century, enlarged the estates and the income of the counts and brought about the need for a rational administrative system. The nobles were a power to be reckoned with, but Count Robert I (ruled 1071–93) and his successors were able to find support and a balancing force in such developing towns as Bruges, Ghent, Ypres, Courtrai, and Cassel. The murder by a family of former bondsmen of the powerful and highly respected Count Charles the Good (ruled 1119–27), who was childless, plunged Flanders into a crisis that involved not only the nobles and the towns but also, for the first time, the French king.

Around 1100 such other territories as Brabant, Hainaut, Namur, and Holland began to expand and form principalities, helped by the weakening of the German crown during the Investiture Contest (a struggle between civil and church rulers over the right to invest bishops and abbots). The Concordat of Worms (1122) ruled that bishops were to be chosen by the chapter of canons of the cathedral; thus, the German king was obliged to transfer the secular powers to an *electus,* who was then usually ordained bishop by the metropolitan. Although the king still exercised some influence over the elections, the local counts were able to make their voices heard the loudest in the chapter, so that Utrecht, for example, soon had bishops from the families of the counts of Holland and Guelders. This was the end of the strong influence that German imperial power exercised through the bishops in the Low Countries. Thenceforth, the spiritual and secular princes stood together, although the death of a bishop still tended to plunge the principality into a crisis—a fact that did nothing to maintain the stability of government.

Decline of the German kings

French and English influence. As their power declined, the Holy Roman emperors could do little more than involve themselves almost incidentally in the affairs and many conflicts of the Low Countries. The German decline went hand in hand with the increasing influence of the French and English kings, particularly after 1200; this applied especially to French power in Flanders. A struggle for the throne that broke out in Germany at the death of Henry VI (1197) found the two powerful factions—the Ghibellines and Guelfs—on opposite sides; in the Low Countries, a game of political chance developed for the territorial princes, in which the Duke of Brabant (Henry I) played an important role, alternatively supporting both parties. The French king Philip Augustus and his bitter opponent, King John of England, both interfered in the conflict, which polarized into an Anglo-Guelf and a Franco-Ghibelline coalition, each looking for allies in the Low Countries. A decisive victory won by the French king at the Battle of Bouvines, east of Lille (1214), put the Count of Flanders at his mercy.

Throughout the 13th century, the French kings increased their influence in Flanders, which was joined to Hainaut by personal union; by 1300, the annexation of Flanders was almost complete. Resistance by Count Guy, which was supported by the crafts in the towns, culminated in a resounding victory by the Flemish army (which consisted largely of citizens of the towns fighting on foot) over the French knights at Courtrai (the Battle of the Spurs, 1302) and prevented total annexation. French power in Flanders remained very strong, however, until the outbreak of the Hundred Years' War (1337). French political influence in Brabant was less powerful; an attempt at political infiltration was foiled by a coalition of the Duke of Brabant and the Count of Hainaut-Holland. English influence and money kept the balance of power in the Low Countries from passing into French hands; England had a strong commercial interest in relations in the Low Countries, particularly in the important ports and shipping of the coastal areas of Flanders, Zeeland, and Holland.

Social and economic structure. To obtain some insight into the social structure of the Low Countries between 900 and 1350, it is important to realize that, although the territorial princes wielded supreme power, the people were in fact directly dependent on an elite that, by virtue of owning land and possessing certain powers of jurisdiction and administration, had formed seigneuries, in which they held considerable effective power. These lords could control their dependents by demanding agricultural services, exercising certain rights over dependents' inheritances, levying monies in return for granting permission to marry,

Social classes

and forcing them to make use of the lords' mills, ovens, breweries, and stud animals. In the main, the owners of these seigneuries were treated as nobles and were often, though not always, bound to the territorial prince by feudal ties. A separated class was formed by the knights, who in the 12th century were usually *ministeriales* (servants who had originally been bondsmen) and were used by their lords for cavalry service or for higher administrative duties, for which they received a fief. It was not until the 13th century and, in many places, even later that the feudal nobility and ministerial knights became unified in a single aristocracy. Apart from these nobles, there were also freemen who owned their own land (*allodium*), but little is known about them; they were present, however, in large numbers in the cattle breeding regions of Flanders, Zeeland, Holland, and Friesland, where the numerous rivers and streams must have split up the land into many small farms. The descendants of noble families who were no longer able to live as richly as the nobles and who were known as *hommes de lignage* (in Brabant), *hommes de loi* (Namur), or *welgeborenen* (Holland), must have been very close in status to the freemen. In the agricultural areas of Hainaut, Brabant, Guelders, and the Oversticht were dependents whose legal status is difficult to determine, though they may be classed as bondsmen because of their being liable for various services and payments.

A factor of great, if not decisive, importance for social and economic relations, not only in the Low Countries but in all of western Europe, was the growth of the population. There is no direct statistical information but only a certain amount of indirect knowledge—after about 1050, it can be seen in the internal colonization (in the form of reclamation of such waste ground as woods and bogs), in the building of dikes and polders, in the expansion of agricultural land, and in the growth of the villages (new parishes) and towns.

Expansion of the land

The opening up of extensive areas of wood and heathland led to the foundation of new settlements (known in the French-speaking areas as *villes neuves*), to which colonists were attracted by offers of advantageous conditions—which were also intended to benefit the original estates. Many of these colonists were younger sons who had no share in the inheritance of their fathers' farms. The Cistercian and Premonstratensian monks, whose rules prescribed that they must work the land themselves, played an important part in this exploitation of new land. In the coastal regions of Flanders, Zeeland, and Friesland, they were very active in the struggle against the sea, building dikes both inland and on the coast itself. At first these dikes were purely defensive, but later they took on an offensive character and wrested considerable areas of land from the sea.

Especially important was the reclamation of marshland in the peat-bog areas of Holland and Utrecht and in the coastal regions of Flanders and Friesland. The Frisians had specialized in this work as early as the 11th century; Flemings and Hollanders soon adopted their methods, even applying them in the Elbe plain in Germany. The system, which consisted of digging drainage ditches, lowered the water table, leaving the ground dry enough for cattle grazing and later even for arable farming. The colonists, who were freemen, were given the right to cut drainage ditches as far back from the common watercourse as they wished. Certain restrictions were later imposed by the lords, however, who regarded themselves as the owners of these waste areas and demanded tribute money as compensation. Reclamation work was organized by a contractor (*locator*), who was responsible to the count and often carried out the function of local judge.

Thus, in the 12th and 13th centuries, a large area of land in the Holland–Utrecht peat-bog plain was made available for agriculture, facilitating the rise of nonagricultural communities (*i.e.*, the towns). In Flanders, Zeeland, Holland, and Utrecht, this struggle against the sea and the inland water was particularly noteworthy in that it led to the foundation of water boards, which in the 13th and 14th centuries were amalgamated to form higher water authorities (the *hoogheemraadschappen*). Mastery over the water had to be carried out on a large scale and in an organized fashion; the building of dikes required a higher authority and coordinated labour. Thus, various organizations emerged, acting independently in the field of canal and dike building and maintenance and responsible only to the government itself. These were *communitates*, with their own servants and their own managements (dike reeves and *heemraden*) and empowered to take necessary measures to maintain the waterworks, administer justice, and issue proclamations. The *hoogheemraadschappen* of Rijnland, Delfland, and Schieland were all in the county of Holland; the dike reeves at the head of these organizations were also the count's bailiffs and were delegated by the count the duties of supreme judge and administrator within their bailiwicks.

Formation of water authorities

The increase in the population and the reclamation of land from the sea and marshes, as well as the fight to keep the sea out, all helped change the social and economic structures of the Low Countries. For centuries, the southern and eastern areas had been agricultural, often making use of the domain system. Fishing and cattle raising had also been important, especially in the water-rich areas of Flanders, Zeeland, Holland, and Friesland (Frisian trade, with Dorestad as its centre, had not been weakened by the Viking invasions). The increasing population and the growing importance of such regions as Flanders and Holland–Zeeland, together with increases in agricultural production, stimulated trade further, though new routes had to be found.

Dorestad itself fell into decay not so much as a result of Viking raids (it was rebuilt after each one) as of a change in the course of the river upon whose banks the town was situated. Dorestad's leading position in trade was then taken over by Tiel, Deventer, Zaltbommel, Heerewaarden, and the city of Utrecht. Wheat was imported from the Rhine plain, salt from Friesland, and iron ore from Saxony, and, before long, wine, textiles, and metal goods were brought along the Meuse and Rhine from the south. The IJssel in Guelders also began to carry trading traffic through Deventer, Zutphen, and Kampen and, on the coast of the Zuiderzee (now IJsselmeer), through Harderwijk, Elburg, and Stavoren.

Growth of Flanders

In the south, Flanders became a second important trading and industrial area. For centuries, sheep farming had produced the wool needed in the cloth industry; but to meet an increased demand wool was imported from England, for which purpose merchants from various Flemish towns joined together in the Flemish Hanse in London. To a large extent, Flanders became economically dependent on England, a fact that must have had considerable political repercussions, particularly because it presented Flanders with the dilemma of having to choose which side to support in conflicts between France and England—on one hand, Flanders owed allegiance to its lord, the king of France, and, on the other, its economic interests were best served by taking the English side. These difficulties led to a reduction in Flemish commercial activity in England. Many Flemish craftsmen emigrated to England, while the English themselves took over the transport of wool. Nevertheless, Flemish cloth remained a high-quality product sold in large quantities at the fairs in the Champagne district and on the banks of the Rhine, to which it was transported by the people of Ghent.

On the whole, however, Flemish trade took on an ever more passive character. Although Flemish ships still sailed to Spain, Portugal, and Bordeaux to buy wine, the merchants of Bruges limited themselves to allowing colleagues from abroad to trade in the city. In this way, Bruges became the centre of trade with the Baltic countries, England, Spain, Portugal, the south of France, and—through the Champagne fairs—Italy. Grain, wood, tar, pitch, and furs were all brought from the east by the merchants of the German Hanse, who had formed a colony of their own in Bruges, where they bought Flemish cloth, wine from France and the Rhineland, and salt from France. Around 1300, the large Italian banking and trading houses set up branches in Bruges and began to send their fleets through the Strait of Gibraltar to Flanders; also, the English kings often sent their wool exports to Bruges, and the city became an international centre of trade and finance.

For the cloth industry itself, Ghent and Ypres were the most important towns. In Ghent, the production process was run by drapers (*drapiers*), who bought the raw material, had it treated by spinners, weavers, fullers, and dyers, and eventually sold the final product. A drop in wool imports from England could therefore cause immediate social and political upheavals in the city, particularly from the large weavers' quarter.

The area of the Meuse also carried on considerable trade and industry; merchants from Liège, Huy, Namur, and Dinant are named in 11th-century toll tariffs from London and Koblenz. This trade was supplied mainly by the textile industry of Maastricht, Huy, and Nivelles and by the metal industry of Liège and Dinant. Trade in Brabant, actively supported by the dukes, used the road, or system of tracks (medieval road systems were not advanced), that ran from Cologne through Aix-la-Chapelle, Maastricht, Tongres, Louvain, and Brussels to Ghent and Bruges. The counts of Brabant wanted to control as much of this trade route as possible by expanding eastward, though there was also considerable north–south trade along the Dijle.

Since prehistoric times, fishing, particularly for herring, had been important in the coastal regions of Zeeland and Flanders. This fishing industry was given added stimulus by the shift of the herring shoals from the coast of Schonen (Sweden) to the North Sea. The ships, however, were increasingly placed at the disposal of general trade and, in particular, of the wool trade with England. The merchants of the German Hanse also turned their attention to Holland, where Dordrecht became the most important centre. Because of its central position in the rivers area, this town offered the counts the chance to raise tolls on all traffic in the neighbourhood; moreover, all cargoes had to be unloaded and offered for sale—wine, coal, millstones, metal products, fruit, spices, fish, salt, grain, and wood—an arrangement that led to further economic expansion of Dordrecht. Not until the 14th and 15th centuries did trade in other towns begin to catch up with that of Dordrecht. Among these were Delft, Rotterdam, and Amsterdam, while Leiden had been one of the first towns with a textile industry; none of them, however, outstripped Dordrecht itself.

The towns. The towns gave the Low Countries a special character of their own. Apart from some towns that had existed even in Roman times, such as Maastricht and Nimwegen, most towns arose in the 9th century; in the 11th and 12th centuries, they expanded and developed considerably. The emergence of the towns went hand in hand with the population increase and the extension of cultivable land, which made possible higher production. The population centres that emerged were not primarily agrarian but specialized in industry and trade.

Industrial and trade centres

The oldest towns were in the regions of the Scheldt and Meuse. Near existing counts' castles or walled monasteries, merchants formed settlements (*portus,* or *vicus*). These gradually merged with the original settlements to form units that both economically and in their constitutions took on their own characters with respect to the surrounding country, characters that were later manifested by defensive ramparts and walls. In this connection, mention must be made of Valenciennes, Bruges, Ghent, and Ypres (around the Scheldt) and of Dinant, Namur, Huy, Liège, Maastricht, Nivelles, and Gembloux (around the Meuse). A later group (though not much later) was formed by the northern towns of Deventer and Tiel, while Utrecht had long been a town in the sense of a commercial centre. Zutphen, Zwolle, Kampen, Harderwijk, Elburg, and Stavoren are other examples of early towns. Much younger (13th-century) are the towns of Holland—Dordrecht, Leiden, Haarlem, Alkmaar, and Delft.

All the towns formed a new, nonfeudal element in the existing social structure, and from the beginning merchants played an important role. The merchants often organized into guilds, the origins of which may be traced as far back as the Carolingian era. From a manuscript dated about 1020, it appears that the merchants of Tiel met regularly for a drinking bout, had a common treasury, and could clear themselves of a charge by the simple expedient of swearing an oath of innocence (a privilege they claimed to have been granted by the emperor). Thus, there and elsewhere, the merchants constituted a horizontal community formed by an oath of cooperation and with the maintenance of law and order as its goal.

In contrast, therefore, to the vertical bonds in the feudal world and within the manors, horizontal bonds emerged between individuals who were naturally aiming at independence and autonomy. The extent to which autonomy was achieved varied greatly and depended on the power exercised by the territorial prince. Autonomy often developed spontaneously, and its evolution might have been accepted either tacitly or orally by the prince, so that no documentary evidence of it remains. Sometimes, however, certain freedoms were granted in writing, such as that granted by the Bishop of Liège to Huy as early as 1066. Such town charters often included the record of a ruling that had been the subject of demands or conflicts; they frequently dealt with a special form of criminal or contract law, the satisfactory regulation of which was of utmost importance to the town involved. Indeed, the first step a town took on the road to autonomy was to receive its own law and judicial system, dissociated from that of the surrounding countryside; a natural consequence of this was that the town then had its own governing authority and judiciary in the form of a board whose members were called *échevins,* headed by an *écoutète* (*scultetus, major, vilicus*). The *écoutète* was appointed by the territorial prince and had the prince's interests at heart. The *échevins* also had to take an oath before the prince, but they were always recruited from the inhabitants of the town itself and therefore more often favoured autonomy than did the *écoutète.* As the towns grew, the work of the board of *échevins* gained in importance, and functionaries appeared who had to look after the town's finances and its fortifications. They were often called burgomasters (*burgimagistri*).

The struggle for autonomy

The development of a town's autonomy sometimes advanced somewhat spasmodically as a result of violent conflicts with the prince. The citizens then united, forming *conjurationes* (sometimes called communes)—fighting groups bound together by an oath—as happened during a Flemish crisis in 1127–28 in Ghent and Bruges and in Utrecht in 1159. The counts of Flanders from the house of Alsace (Thierry, ruled 1128–68, and Philip, 1168–91) kept careful watch, supporting and aiding the towns in their economic development but otherwise keeping the process in check.

In their struggle for autonomy the towns had to fight for financial freedom, such as for the lessening or abolition of the taxes and tolls they had to pay to the prince but also and principally for the right to impose their own taxes, usually in the form of indirect taxation (*e.g.,* excise duties), in order to raise money for necessary public works. Especially important to them was the right to frame their own laws; this legislative right (the *keurrecht*) was in most towns originally restricted to the control of prices and standards in the markets and shops but was gradually extended to cover civil and criminal law. The extent of a man's obligation to serve in the prince's armed forces was often fixed or limited or both (sometimes by the provision for payment in lieu, sometimes by a legal definition of the number of foot soldiers or manned ships to be made available).

Thus, the town in the Low Countries became a *communitas* (sometimes called *corporatio* or *universitas*)—a community that was legally a corporate body, could enter into alliances and ratify them with its own seal, could sometimes even make commercial or military contracts with other towns, and could negotiate directly with the prince. Land within the town's boundaries usually became its property, and the town's inhabitants were usually exempt from any dependent relationship with outsiders.

Size of towns

The size of the population naturally varied considerably from town to town, but, in general, a normal town in the northern Low Countries had a population of between 2,000 and 6,000; Dordrecht, with 10,000 inhabitants, was a very large town for the region, surpassed only by Utrecht. In comparison, the Flemish towns were exceptionally large: in the mid-14th century, the population of Ghent was about 56,000; that of Bruges, 35,000. On the

other hand, Antwerp, which belonged to Brabant, had at that time only 5,000 inhabitants.

A town's population usually had a distinct social structure. The merchants, the oldest and leading group, soon emerged as a separate class (the patriciate); they generally managed to gain control of the offices of *échevin* and burgomaster and thus controlled the town's finances. Sometimes the *homines novi,* a new class of up-and-coming merchants, tried to become part of the patriciate, as in Dordrecht and Utrecht. Beneath the patriciate a lower class formed, called the *gemeen* ("common," in the strict sense of the word), which embraced the artisans and organized into crafts such tradesmen as butchers, bakers, tailors, carpenters, masons, weavers, fullers, shearers, and coppersmiths. These crafts originally developed out of charitable organizations of people in the same profession and had to adhere to regulations laid down by the authorities. Gradually, however, they tried to obtain their independence, exercise influence in politics, cut themselves off from outsiders by means of compulsory membership, and introduce their own regulations regarding prices, working hours, quality of products, apprentices, journeymen, and masters. Early tension between crafts and patriciate developed in the Flemish towns and in Liège, Utrecht, and Dordrecht. In Ghent, Bruges, and other Flemish towns, the crafts were able, with the help of the count, to gain a victory over the French and their patriciate sympathizers.

Relations between towns and territorial princes

In Flanders and in the bishopric of Liège, the towns rapidly attained such power that they constituted a threat to the territorial prince, which hindered stable development in these principalities and often resulted in violent conflicts. In contrast to this, relations between the prince and the towns of Brabant were extremely harmonious; the political interests of the prince and the economic interests of the towns usually coincided. In Holland, however, the towns did not really develop until as late as the 13th century, when they were deliberately helped by the counts William II (ruled 1234–56) and Florent V (ruled 1256–96), who granted them charters that included extensive rights.

During this period, foundations were laid for the dominant role the towns later played in the Low Countries in political, social, economic, and cultural fields. A decisive change also took place in the authority of the territorial prince. Originally he regarded his powers mainly as a means of increasing his income and of extending the area over which he could exercise power. He felt little duty toward his subjects or desire to further the welfare of the community as a whole. At best there were religious as well as material motives in his dealings with the churches and monasteries. There were no direct relations between the prince and all his subjects, for he was primarily lord of his vassals. The political, social, and economic developments discussed above, however, brought a change in this situation. In the first place, the prince's increasing independence meant that he himself began to behave like a king or sovereign lord. His authority was then referred to as *potestas publica* ("public authority"), and it was believed granted by God (*a Deo tradita*). The area over which he ruled was described as his *regnum* or *patria.* This implied not only the duty of a lord toward his vassals but also that of a prince (*princeps*) toward his subjects. This duty included as its first priority the maintenance of law and order (*defensio pacis*) by means of laws and their administration. He had further to protect the church (*defensio* or *advocatio ecclesiae*), while his involvement in land reclamation and in the building of dikes and with the development of the towns brought him into direct contact with the nonfeudal elements of the population, with whom his relations were no longer those of a lord toward his vassals but took on a more modern aspect—that of a sovereign toward his trusted subjects. He became, according to the 14th-century lawyer Philip of Leiden, the *procurator rei publicae* ("he who looks after the matters of the people"). Contact with his subjects was through the representatives of the *communitates* of the water boards and *heemraadschappen* and through the towns and nonurban communities, which were legally corporate bodies not only in dealings with outsiders but also with the prince. Sometimes the towns expressly placed themselves under the protection of the prince and declared themselves committed to loyalty to him. Such a town was Dordrecht, which, in a document dated 1266, expressed its loyalty and at the same time described the count of Holland as *dominus terrae* ("lord of the land"). These new notions of a sovereign lord and sovereign power point to a more modern conception of a state, to a growing awareness of territoriality, and to new possibilities of collaboration between prince and subjects.

BURGUNDIAN AND HABSBURG RULE (1384–1579)

Among the many territorial principalities of the Low Countries, Flanders, Brabant, Hainaut-Holland, and Guelders in the mid-14th century had a dominating military and diplomatic position. Flanders had already arrested the course of French domination, and its feeling of territoriality was strengthened by this and by many minor wars between the principalities. The feeling found expression when the prince's subjects increasingly set themselves up against him as organized estates—those of the clergy or nobles and those of the towns. In the 14th century, this led almost everywhere to the formation of permanent organs that consisted of representatives of the estates. These organs represented the common interests of the estates against or, sometimes, in agreement with those of the territorial prince. At the same time, committees and law courts were set up to keep a watch on the prince's officials.

Organized estates

Constitutional developments. In Brabant, the close cooperation between duke and towns resulted in the harmonious growth of estates, though this was also due to the continuing lack of financial resources from which the dukes suffered and which gradually made them more and more dependent on the increasingly powerful and rich towns. The towns, therefore, set the pace, especially during periods of regency, when a ruler was below the age of majority. Between 1248 and 1356, the towns of Brabant developed cooperation between estates, culminating in the Joyeuse Entrée, in which this cooperation was defined and laid down in a constitution: the unity of Brabant and Limburg was declared indivisible, and the prince was obliged to choose his officials exclusively from among the people of Brabant and to win the approval of the towns and of his county before embarking on a war.

In Flanders, the power of the three towns of Ghent, Bruges, and Ypres was too great to enable smooth cooperation with the prince and the two other estates—those of the clergy and the nobles. The count succeeded only in making the Franc of Bruges—the countryside surrounding the town—a fourth estate, but there was no question of a constitution, written or otherwise.

Better results were obtained in Holland, although only after a political conflict between two political factions—Kabeljauwen (Cods) and Hoeksen (Hooks)—in which the Kabeljauwen advocated cooperation between the estates and the prince but had too few noble families among their supporters. Guelders, however, was late in its development, partly because the powerful Duke Willem (ruled 1379–1402) had his own financial resources as a result of his military activities in the service of the English and, later, French kings; under Willem's successors, however, the knights and the towns became more powerful and finally gained permanent representation as estates. In Utrecht, too, there was cooperation between the prince (the bishop) and the estates; and the clergy, particularly the collegiate churches of the town of Utrecht, played an important part: the Land Charter of Bishop Arnold in 1375 is comparable to the Joyeuse Entrée of Brabant. In the prince-bishopric of Liège, cooperation between prince and estates had to be won by violent conflicts between the towns and the bishop and, within the towns, between the patriciate and the crafts. It was mainly to these territorial estates that the princes had to turn for financial help, which was often voted to them only on limiting conditions.

Beginning of Burgundian infiltration

The Burgundians. In the second half of the 14th century, the dukes of Burgundy (princes of the French royal house of Valois) began to penetrate these territorial principalities in the Low Countries, whose feelings of territoriality made them regard the dukes with suspicion. The marriage in 1369 of Philip of Burgundy (Philip the Bold)

to the heiress of the Count of Flanders (Margaret) signified the beginning of this Burgundian infiltration, which was repeatedly furthered by such tricks of fate as marriages and inheritances.

Through his marriage Philip gained possession, after the death of his father-in-law in 1384, of the counties of Flanders, Artois, and Rethel and also of the county of Nevers and the free county of Burgundy (Franche-Comté), the last of which was within the Holy Roman Empire. He thus not only gained a large and powerful part of the Low Countries but was also able to extend his Burgundian property. Though it seemed at first that French power might again become the dominant force in the Low Countries, it soon became clear that the Burgundian dukes, while happy to continue taking part in French politics, were extremely independent and more interested in forging a single powerful empire out of the Low Countries and Burgundy. Duke John the Fearless succeeded to all his father's lands in 1404, while his younger brother Anthony was given Brabant, where the childless Duchess Joanna had named him as her successor, which was accepted by the estates. Anthony's branch of the Burgundians died out as early as 1430, so that Brabant fell to the other branch under Philip the Good (ruled 1419–67), who also gained possession—through war, family relations, and purchase—of Hainaut-Holland, Namur, and Luxembourg. This Burgundian power structure was not a state but was founded on a personal union among the various principalities, each of which jealously guarded its own freedom and institutions. The Burgundian dukes did, however, attempt to set up central organizations to bridge the differences among the principalities and to keep the various regions under stricter control by appointing viceroys (stadholders); but this ambition ran aground on the forced and overly hasty centralization and expansion of power carried out by Charles the Bold (ruled 1467–77), who was able, nevertheless, to annex Guelders. Charles imposed increasingly high financial demands, which were put before the States General—an assembly that united the delegates from the various states at meetings called by the duke and held at regular intervals; he introduced a central law court (the Parliament of Malines) and a central exchequer and did all he could to gain the status and title of king. After his death, a movement for regional independence arose and won a series of privileges from his daughter Mary (ruled 1477–82) that virtually nullified the previous centralization. Moreover, the duchy of Burgundy itself was taken over by the French crown, so that the Burgundian union became a union without Burgundy. An important point was that both the States General and the territorial states demanded and were given the right to meet on their own initiative.

After Mary's position had become more firmly established by her marriage to Maximilian of Habsburg (the son and future successor to the Holy Roman emperor), the States General, because of its internal particularism, proved unable to provide a lasting administration. Gradually, a restoration took place, at first under the regency of Maximilian after Mary's death in 1482 but principally later, after their son Philip the Handsome took over the government. Philip resumed the centralization process by refounding the central law court (then known as the Great Council of Malines) and by setting up within the duke's council permanent commissions to discuss important political and financial questions.

Centralization

The Habsburgs. The fate of the Low Countries was already closely bound up with that of Austria by virtue of the Habsburg marriage; in 1504, this situation was intensified when Philip and his wife, Joanna, inherited the Spanish crown. From then on, the Low Countries were merely a part of a greater whole, and their fate was principally decided by the struggle of this Spanish-Austrian empire for European hegemony. They repeatedly had to make sacrifices for the many wars waged against France, particularly under Emperor Charles V, who in 1519 had added the German imperial crown to his many possessions. The Emperor, who was almost always out of the country, placed the Low Countries under the rule of governors general—first his aunt Margaret and later his sister Mary, who retained control and worked toward further centralization even when he was in the country.

The States General could do little more than offer passive resistance, principally by providing financial obstructions to centralization. As a meeting place for the regional deputies, the States General did have a certain influence and, by its opposition, strengthened a sort of negative feeling of unity. That the Emperor himself also saw the Low Countries as a unit can be seen in his measures to separate them (but not entirely) from the empire as "Burgundian Kreis" (1548) and in the Pragmatic Sanction (1549), which stated that succession would be regulated in identical fashion in all the regions of the Low Countries that he had included in his empire (among these were Utrecht, Guelders, Groningen, and Friesland). The Low Countries were thus prevented from being split up.

In the meantime, the process of centralization had reached a decisive phase with the foundation of the collateral councils (1531), which were separate from the Great Council. They were the Council of Finance, which had in effect already existed for some time; the Council of State, in which members of the high nobility could advise the governess; and the Secret Council, in which permanent officials dealt with everyday administration and composed ordinances without having to wait for advice. All the government organs, except for the central law court in Malines, were in Brussels, which may thus be considered the capital of the Low Countries. The States General and territorial states were no longer allowed to sit whenever they wished; the territorial states were still a stumbling block in the acquisition of financial resources, so that Charles V was never able to provide himself with a standing army.

Under Charles's son Philip II, who in 1555–56 succeeded as king of Spain and prince of the Netherlands, the policy of centralization was continued. It culminated in the introduction of a new ecclesiastical hierarchy. The Low Countries, which formerly had been, ecclesiastically speaking, merely an extension of the archbishoprics of Cologne and Reims, became by virtue of a papal bull of 1559 a directly governed region of the church under three archbishops and 15 bishops. Resistance to this was fierce by the high nobles, who saw the high positions in the church slip from their grasp; by the abbots, who feared the incorporation of their monasteries; and by a number of territories, which were afraid of greater centralization activities under new bishops. The high nobles, who were often excluded from the activities of the Secret Council, led the resistance under the capable Prince William of Orange (1533–84) and the popular count of Egmond. Resistance increased when the Burgundian Granvelle, who was bishop of Artois and virtually Margaret of Parma's prime minister, was appointed cardinal and then archbishop of Malines and primate of the Netherlands. The government gave way, and Granvelle was forced to leave the country; yet the high nobles themselves hardly knew how to run affairs. The initiative was thus transferred to the low nobility, who in 1565 united by bond of oath in the Compromise and in 1566 presented to the Governor a petition requesting the relaxation of edicts and ordinances against the Calvinists and other Protestants. At the same time, they adopted the name Geuzen (*gueux,* "beggars"), originally a term of abuse.

Resistance to a new ecclesiastical hierarchy

The resistance grew stronger, the Calvinists became more confident, and fanatics started a violent campaign against churches—the "breaking of the images" (August 1566)—against which the Governor took powerful measures and in the first few months of 1567 managed to restore peace. The King, however, whose information concerning these events was somewhat out of date because of slow communications and who was uneasy because of the "breaking of the images," decided to take stern measures. He sent his trusted general, the Duke of Alba, to the Netherlands. Alba's strict regime precipitated a revolt that eventually led to the partition of the Netherlands.

Economic structure. The economic structure of the country underwent far-reaching changes in the 14th–16th centuries. The growth in population, which in western Europe had begun in the 10th century, ceased with relative suddenness after 1300. Moreover, Europe in the mid-14th

century was afflicted by wars, famines, and epidemics, which also affected the Low Countries. The Black Death of 1347–51 spared the Low Countries to a certain extent, but the cut in the population of the other European countries must have reduced the market for the Low Countries' trade and industry. The economy, relatively free in previous centuries, was now more limited and regulated, especially where the crafts had come to power, as in Flanders and Liège. The cloth industry of the large towns in Flanders and Brabant suffered a severe setback. Nevertheless, the economy during the 14th and, especially, the 15th and 16th centuries sought and developed new areas for expansion. This led to a remarkable contrast between the apparent flourishing of trade and industry and the visible poverty of the mass of the people. The most noticeable development was the decline of trade in Bruges and its growth in Antwerp, which became the economic centre of the Netherlands in the late 15th century.

Antwerp, which as early as the 13th century had attained some importance, had attracted an increasingly large proportion of the English cloth imports during the 14th and 15th centuries. Conservative rule in Bruges and the insecurity created by the disturbances in Flanders caused many foreign merchants to move their business to Antwerp. The Portuguese, in particular, took their spices from the Indies to Antwerp, where they traded them for copper and silver brought by German merchants. Thus, Antwerp developed into an international port, where foreign merchants organized themselves into "nations" and where not only German but also Italian banking and trading houses had branches. New methods of trading and international finance developed in Antwerp, such as buying from samples and transferring money by means of bills of exchange. Even princes contracted loans in Antwerp.

Commercial capitalism

A commercial capitalism was developing that stimulated the entire economy of the Netherlands. Although the cloth industry declined in the towns, it flourished in the villages and began to specialize in lighter cloths of coarser quality (principally in the country districts of Flanders, with Hondschoote as its centre).

The towns battled these rural industries in vain, though in 1531 Holland issued an edict to restrict them throughout the county but with little success. Moreover, Holland itself began to play an increasingly important economic role—Leiden boasted a growing textile industry, Delft and Haarlem had export breweries, and in the area of the River Zaan shipbuilding developed; nevertheless, fishing, shipping, and trade were its main means of support apart from arable farming and cattle breeding. Dordrecht was rivalled by Rotterdam and Gorinchem and, by the 16th century, was even outstripped by Amsterdam, which cornered an increasing proportion of Baltic trade, as may be seen from the lists of the toll in the Sound (between Sweden and Denmark).

The regions along the Meuse and IJssel also maintained their commercial activity. In the bishopric of Liège there was even a metal industry with blast furnaces, paid for by capital raised by traders. Coal mining in the area between the Meuse and the Sambre was also organized according to modern capitalist methods.

The cultivation of commercially exploitable crops also developed in country areas—hemp for rope making, hops and barley for brewing, flax for the manufacture of linen. But all this was at the expense of wheat farming. Grain had to be imported in increasingly large quantities, and, whenever grain imports fell off, the people, particularly the lower classes, went hungry. The economic apparatus had become more versatile and brought greater prosperity, but at the same time, precisely because of this specialization, it had become more vulnerable. The distribution of prosperity was variable; the great mass of the people in the towns suffered the consequences and bore the main burden of the rise in prices occasioned by inflation. It is estimated that, of the total population of Brabant, Flanders, and Holland, between a third and a half lived in the towns, where the opportunities for earning a living were still greater than in the country and where charitable institutions concerned themselves with the lot of the poor.

Culture. The Netherlands played an important part in the artistic, scientific, and religious life of Europe. In the late Middle Ages, when prosperity was increasing and the princely houses, particularly that of the Burgundians, as well as the middle classes in the towns, were encouraging progress, the Netherlands began to make an independent contribution to cultural life. In the southern Netherlands, this was already the case with mysticism, which reached its zenith in the 13th and 14th centuries in the poems of Sister Hadewych and in the prose of the prior Joannes Ruusbroec (Jan van Ruysbroeck). Ruusbroec's writings were founded on a considerable knowledge of theology; it is not certain whether his work had a direct influence on the founding of the religious movement along the IJssel—the modern devotion (*devotio moderna*)—or whether mysticism merely created the intellectual climate in which the new school of thought could develop. The modern devotion was inspired by Geert Grote (Gerard Groote, 1340–84) of Deventer, who preached, as did many others, the ascetic and pious life and resistance to the secularization of the church. His message was well received, and many lay people found in themselves a desire to live in communities devoted to the service of God; these were the Brethren and Sisters of the Common Life, who later organized themselves into the Windesheim monasteries and convents, which followed Augustinian rules. Their communities were extremely important for both education and religion; they were industrious copyists and brought a simple piety to the lower classes. Their work, like that of the mendicant orders, was a typical product of life in the towns. The movement reached its peak in Thomas à Kempis, from Zwolle, whose *Imitatio Christi* (*The Imitation of Christ*) became quite widely read.

Artistic, scientific, and religious contributions

Within the modern devotion, where great importance was attached to good teaching, Dutch Humanism was able to develop freely. Of importance was the foundation in 1517 of the Collegium Trilingue in Louvain, where Latin, Greek, and Hebrew were taught. The greatest Dutch Humanist was Desiderius Erasmus (1469–1536), whose fame spread throughout the world and who had been taught in the schools of the Brethren of the Common Life. He drew his inspiration, as did many other Humanists, from antiquity and was famed for his pure Latin. He was in touch with the greatest minds of his time, visited England (Cambridge) and Italy, and worked for some years in Basel and in Freiburg. Erasmus' greatest achievement was to turn the science of theology, which had degenerated into meaningless Neoscholastic disputes, back to the study of sources by philological criticism and by publishing a new edition of the Greek New Testament. Although he vociferously criticized the church and even the princes, he avoided out of conviction a break with the church and pleaded for a religious tolerance that can be seen under many rulers of the later republic.

The Humanists were principally intellectuals, however, expressing themselves in literary and scientific treatises and having little impact on the broad masses of the people. Many of them, like Erasmus, desired no break with the church and did not accept that break when it became a fact by the appearance of Martin Luther. Instead, they wanted reformation within the church. It was otherwise for the reforming movements that brought turmoil to the Netherlands in the first half of the 16th century. Even Lutheranism had few followers, despite its early appearance in the Netherlands (Luther's dogmas were condemned by the University of Louvain as early as 1520). There was a Lutheran community in Antwerp; but otherwise, support for Lutheranism was limited to individual priests and intellectuals. It was precisely over the question of the Eucharist that the Sacramentarians differed with Luther; they denied the consubstantiation of Christ in the Eucharist. Although the Sacramentarians enjoyed little support from the people, their questioning of the value of consubstantiation was of considerable significance, and their teaching was propagated in heretical Bible translations and even in rhetoric competitions.

Reforming movements

Uproar was caused by the Anabaptists (so called because they rejected the baptism of infants and therefore had themselves rebaptized as adults), who refused to swear the oath of allegiance to the prince or to serve in the

armed forces or in government per se and who believed in a *lumen internum* ("inner light"). This baptist movement won great popularity in the Netherlands after 1530; from the very beginning there were two branches—the social revolutionaries and the "quiet baptists." The first of these was characterized by a lively enthusiasm and a willingness, once the external trappings of the church had been rejected, to organize itself into communities, which soon formed close ties with each other. Prophesies by the social-revolutionary branch of the imminent coming of Christ and of a New Jerusalem fascinated the masses, while their fanaticism and readiness to sacrifice themselves made a deep impression on a population suffering poverty and misery. In 1534 a section of the Anabaptists moved to Münster in Westphalia, where they supposed that the New Jerusalem would be built; and in 1535 an abortive attempt was made to take over the town hall in Amsterdam. After a long siege, the Bishop of Münster succeeded in reconquering his town, and the Anabaptists suffered terrible vengeance. Only the "quiet baptists" were able to continue under the leadership of the Frisian pastor Menno Simons (these Mennonites are even today strongly represented in the provinces of Groningen, Friesland, and North Holland).

The future of the movement for reformation in the Netherlands was assured, however, not by the biblical Humanists nor by the Anabaptists but by a movement less intellectual than the first and more realistic than the second—Calvinism.

Calvinism

The theology of John Calvin (1509–64) was radical, strict, logical, and consistent. Its central theme was the absolute might and greatness of God, which made man a sinful creature of no significance who hoped merely to win God's grace by honouring him in daily hard work. Calvinism found its way to the Netherlands by way of France, though there may have been some direct influence from Geneva, Calvin's town. Calvinist writings were known in Antwerp as early as 1545, while the first translation into Dutch of his *Christianae religionis institutio* is dated 1560, which was also the year in which support for him spread in the Netherlands, largely because the Calvinists preached their creed in public and held open-air services.

Calvinist teaching appealed not only to the lower classes but also to the intellectual and middle classes because of its glorification of work, its discipline, its organization into communities, and its communal singing of the psalms. The government saw the movement as a threat to its plans of unity and centralization, and it took stern measures against Calvinism. Calvinists forcibly removed their coreligionists from prisons and occasionally even attacked monasteries, and their rejection of icons, paintings, statues, and valuables in churches sometimes led them to remove them and hand them over to the town magistrates. But this idealism became corrupted, the leaders were unable to retain control of the movement, and it degenerated into the "breaking of the images" (1566), mentioned above. The nobility, and especially the lower nobility, which had been sympathetic to Calvinism, discarded the movement; the government besieged and captured the Calvinist centre, Valenciennes, by defeating a Calvinist army at Oosterweel, near Antwerp. The result was a great exodus of Calvinists. Nevertheless, Calvin's ideas had penetrated deeply, and his supporters, who had emigrated to England, East Friesland, and the Pfalz of Germany, were able to maintain their unity, which had been threatened by friction between the *rekkelijken* ("liberals") and the *preciezen* ("rigids"), and to support their coreligionists in the Netherlands. The Calvinists were to become the driving force behind the revolt against Spanish rule.

The revolt and the formation of the republic (1567–79). The forcible measures taken by the central government against the "breaking of the images" were followed by a brief period of peace. Fernando Álvarez de Toledo, duke of Alba (who became governor after the departure of Margaret of Parma on the last day of 1567), introduced stern measures at the express command of the King. These provoked a resistance to the government (often referred to as the "revolt") that triggered off the Eighty Years' War (1568–1648). The "breaking of the images" itself, which had raged across the country like a storm, had already shown a deep-rooted resistance that had many causes and was brought to a head by Alba's measures.

The revolt

It is impossible to label any of the causes of the revolt as the decisive factor. An important one, however, was a religious motive. Criticism of the structure of the church and the riches and worldly way of life of its prelates and the accompanying desire for reform had always been strong in the Netherlands; and Protestantism, through the teaching of Luther, the Sacramentarians, the Anabaptists, and, above all, the Calvinists, had gained a firm foothold. The measures taken against the resistance—harsh edicts, prison sentences, torture, and death sentences, carried out with great cruelty—fanned the flames all the more and among all classes. Social and economic causes, however, also lay behind the resistance, especially among the lower classes—the wars with France, the epidemics, poor harvests, hard winters, floods, and a frightening inflation and consequent rise in prices all combined to cause despair and misery among the masses and made them susceptible to radical ideas. At the same time, in the upper classes of the nobility and the urban patriciate, there was a sharply felt reaction against the absolutist centralization policy of the King, who lived far away in Spain and yet whose wish was law in the Netherlands. Towns felt their privileges being threatened, and the nobles found their independent status being undermined by the ever-increasing activities of the secret council. The mercenaries, who were often stationed in a town as a garrison and acted as occupying forces, also aroused hostility. The fact that the resistance did not present a united front may be ascribed to the particularism among the territories—Holland, with its commercial interests, could hardly be expected to be enthusiastic on behalf of typically agrarian feudal provinces such as Hainaut or Artois.

It is clear, however, that the terror organized by Alba burst like a bombshell in this political, social, economic, and religious climate. The Prince of Orange, with sharp political insight, had decided not to wait for Alba's arrival; he had managed to escape in time to his birthplace in Nassau-Dillenburg, leaving behind all his possessions, which were promptly confiscated. His son, Philip William, was taken prisoner to Spain. Alba sent his troops to the principal towns and set up the Council of Troubles (or Council of Blood), which imposed severe penalties, often including the death sentence or confiscation of property, sparing nothing and nobody, not even the most powerful—the counts of Egmond and Hoorne were publicly beheaded in Brussels in June 1568.

Alba also rushed through installation of the new ecclesiastical hierarchy, which had not been completed. Furthermore, he attempted to make the central government independent of the provincial states by means of new taxes on property, on the sale of land or building, and on the sale of goods. This met with violent resistance because the taxes were to be general and permanent, so that the separate states would no longer have the means to make conditions for the furnishing of taxes (although they themselves already levied taxes on the sale of goods) and, more important, because a permanent tax system would make the king independent of his subjects. The taxes were the final link in the policy of absolutism and centralization, which would lead to a unified state controlled by a prince with unlimited power.

New taxes

The severity with which Alba ruled was not able to prevent the immediate appearance of resistance. The Geuzen conducted pillaging raids in the country and piracy at sea, for which they had "authority" in the form of letters of marque issued by William of Orange in his capacity as sovereign of the principality of Orange. Attacks took place as early as 1568. A small force led by Louis of Nassau, William's brother, enjoyed a modest victory over the Spaniards at Heiligerlee (in the province of Groningen), considered the beginning of the Eighty Years' War; but shortly afterward Louis was defeated near Jengum in East Friesland. A greater setback, however, was the complete failure, due to lack of funds, of a campaign led by William himself in the south of the country. During the sombre years of 1568–72 the "Wilhelmus" was written—a song

of faith, hope, and trust that was to become the Dutch national anthem. Other songs written by the Geuzen kept up the spirits of the people during this period and in later years.

During these years, William negotiated for help from Germany, England, and, above all, the French Huguenots. A large-scale attack was planned for the summer of 1572. Before William could carry it out, the Geuzen seized the port of Briel (April 1, 1572), west of Rotterdam. This was a move of considerable strategic importance because the port controlled the mouth of both Meuse and Waal, and the Prince immediately supported the movement. The Geuzen then took Flushing, Veere, and Enkhuizen, so that William had useful bases in Holland and Zeeland. The help that the Geuzen received from the Calvinists in these towns was striking—the Calvinists, a radical minority, were again and again able to force the more conservative town magistrates either to cooperate or resign. Oudewater, Gouda, Dordrecht, Leiden, Hoorn, and Haarlem followed, only Amsterdam keeping the Geuzen out. The purposeful activities of the Calvinists also led to their gaining churches, often the principal church of a town, for their services; they closed monasteries, and Catholic services were soon forbidden.

The revolt was at first successful only in Holland because of its unique position. As a commercially oriented province, it had been more inclined to look after its own interests than to cooperate with other provinces. Trade had been seriously threatened by the Geuzen but was now free again. Moreover, the province lay in a strategically favourable position—difficult to reach from the central government in Brussels and almost inaccessible to the Spanish armies by virtue of its many rivers, lakes, drains, and bogs.

Legal basis

To give the revolt a legal basis, the fiction was invented that it had been a revolt not against the King but against his evil advisers, particularly the Governor. By their own authority, in July 1572 the states of Holland gathered in Dordrecht, where William of Orange was proclaimed stadholder of Holland and Zeeland. The Prince himself went to Holland and, realizing that the Calvinists had been the driving force behind the revolt, became a member of the Calvinist church. But he repeatedly expressly avowed his ideal of the United Netherlands, in which there would be room for Catholics and Calvinists alike.

Alba, disappointed by his failure to push through the tax reforms and about to return to Spain, learned of the fall of Briel and decided to stay and start a counteroffensive. The south was immediately brought under control with the occupation and plundering of Malines; then Zutphen and Naarden in the north were taken and likewise plundered. This provoked stronger resistance, and Haarlem was retaken only after a very long siege, which not only demoralized and decimated Alba's troops but also strengthened the other towns in their decision to offer resistance (1573). Thus, the Spaniards were unable to take Alkmaar, their fleet suffered a heavy defeat in the Zuiderzee, and a long siege of Leiden was relieved by flooding the surrounding country (1574). (As a reward, the town later was given a university, where Calvinistic theology was to be the principal subject for study.) Spanish troops were never again to force their way into Holland—a heavy blow for the most powerful monarchy in the world.

Alba left on December 18, 1573, and his successor, Don Luis de Requesens, was unable to prevent further secessions in the north. Even the south, which had been loyal to Spain until then, but where active Calvinist movements existed (especially in Ghent), became amenable to William's ambition for a united resistance to the Spanish regime. There were considerable problems involved, and one of the greatest points of difference was the question of religion—the more radical north demanded the total abolition of Catholicism in Holland and Zeeland and the acceptance of Calvinism by the southern provinces, though William was diplomatic enough not to make this demand. It was finally agreed that the States General would deal with the question later, and until such time the Calvinists would be masters only of their own house in Holland and Zeeland. A new governor (Requesens died in March 1576) was to be accepted only if he approved the pacification and sent away the foreign troops, who, because they had received no pay, were beginning to mutiny and plunder and were becoming an increasing nuisance. Another condition of his acceptance was that he govern with native officials and in close consultation with the states. On this basis, delegates from all the provinces came to an agreement, and on November 8, 1576, they signed the Pacification of Ghent. Their sense of unity was further strengthened by the news that on November 4 Antwerp had been invaded by mutinying Spanish troops, who had slaughtered 7,000 citizens in what came to be known as the "Spanish Fury."

The Pacification of Ghent

William's idealism, his desire for unity, and his tolerant ideas had apparently triumphed. Unity of thought, however, did not last long; and within three years signs of a split appeared between the northern and southern provinces (which later became a permanent split). It was immediately obvious that within the United Netherlands there were opposing powers of radicalism and reaction. For various reasons, they could not maintain equilibrium; the reactionaries tried to force their ideas on the country with the help of the new governor, Don Juan de Austria, a half-brother of the King, and the Calvinists continued their radical program to make theirs the official and only religion. Many factors caused these conflicts—deep-running religious differences between north and south, a deeply rooted particularism that hindered cooperation, structural and economic differences between Holland and Zeeland on the one hand (commerce and industry) and Hainaut and Artois on the other (agrarian economy and feudal possession of land). It is impossible to point to any one decisive factor. William did his utmost to save the Pacification, and he found support for his ideas of tolerance among the rich burghers; yet he was unable to bridge the differences between rich and poor, Catholics and Calvinists. Moreover, Don Juan died in 1578 and was succeeded by Alessandro Farnese (duke of Parma and son of the earlier governess Margaret), who was conspicuous for his military and diplomatic gifts, which made him a worthy opponent for William and to which may be ascribed the bringing back of the southern provinces to loyalty to the king.

Notable, too, was the appearance in the north and south of movements in the direction of "limited unions," which within the whole of the United Netherlands were to bring about greater community of interests between certain provinces. On January 6, 1579, the Union of Arras (Artois) was formed in the south among Artois, Hainaut, and the town of Douay, based on the Pacification of Ghent but retaining the Catholic religion, loyalty to the king, and the privileges of the estates. Although William himself did not at first reject the idea of a northern union and expressly stated that it must be open to all provinces, matters subsequently developed in a direction not to his liking. (C.v.d.K.)

The Netherlands

One of the smaller countries of Europe, the Kingdom of The Netherlands (Dutch, Koninkrijk der Nederlanden) is a parliamentary democracy under a constitutional monarch. The Netherlands has often been known in English as Holland (from Houtland, or Wooded Land). This name, which was originally given to the medieval core of what later became the state and which is still used for two of its 11 provinces (Noord-Holland and Zuid-Holland), accurately reflects the country's character. The irregular outline of The Netherlands, not unlike a piece of a jigsaw puzzle, encloses some 16,042 square miles (41,548 square kilometres) of extraordinarily flat land lying between the North Sea on the north and west, West Germany on the east, and Belgium to the south. The kingdom incorporates the former colony of the Netherlands Antilles as well.

Some 3,000 square miles of this land have been wrested from the sea by a process of careful water management, drainage, and reclamation that dates back to the 12th century, although the really large schemes were initiated in the 20th century. Much of the Zuiderzee, the great arm

The struggle against the sea

of the sea that formerly tongued into the country from the north, has been turned into *polderland,* land surrounded by dikes and artificially drained. The Waddenzee, the northern seaward sector, is now separated from the larger southern portion, known as IJsselmeer (Lake IJssel), by a huge barrier dam almost 20 miles (32 kilometres) long.

Despite a government-encouraged emigration after World War II, which sent some 500,000 persons out of the country (mostly to the English-speaking countries of North America and the traditionally white nations of the then British Empire and Commonwealth), the country is today one of the most densely populated nations in the world. Partly as a result of the earlier emigration, about one-tenth of the citizens are now over 65 years old, but this fact has not prevented the capital, Amsterdam, from becoming one of the liveliest centres of the international counterculture of Western youth. This independence of outlook, if it can be thus interpreted, has very strong roots in The Netherlands. Citizens of the 16th and 17th centuries rose time and again against foreign domination to found and secure their own republic. Dutch merchant ships sailed the world and helped to lay the foundations of a great trading nation characterized by a vigorous spirit of enterprise. The Netherlands contains the estuaries of the rivers Rhine, Maas (Meuse), and Scheldt, and through their seaports the Dutch control a large share of the world's waterborne traffic. All this is a far cry from the phlegmatic composure that other Europeans have sometimes ascribed to the Dutch, though something of the traditional image of the country may be preserved in the neat, brick-built little towns, the countless windmills, and the magnificent colours of the spring tulip fields, even if the famous peasant dress—wooden clogs, the full-bottomed trousers of the men, and the white caps and wide skirts of the women—is not so widespread today.

With Belgium and Luxembourg, The Netherlands is a member of the economic union known as Benelux, in which foods, goods, and people can freely circulate; this union, in fact, served as a model for the larger European Economic Community (EEC), of which the Benelux nations are members. The country is also a member of the North Atlantic Treaty Organization (NATO) and the Organisation for Economic Co-operation and Development (OECD). Of the former Dutch colonies, the Dutch East Indies became independent as Indonesia in 1949, and Suriname became independent in 1975.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* The landscape of The Netherlands is almost entirely flat. It is characterized by dunes and dikes protecting the most densely populated part of the country, which would otherwise be inundated. Villages and estates are surrounded by dikes and a system of drainage canals, while about half of the land area consists of polders, which intensify the flat appearance of the land.

There are several hundred polders, some made possible by the operation of windmills, which, despite their limited power, carried out the drainage in earlier times. But the larger polder areas, such as those of IJsselmeer, have been created by steam, diesel, and electric pumps. Fen, river, and sea polders lie on peat land along the rivers, but some polders above sea level, created by silting, are drained of excess water by sluice gates at ebb tide. The western part of The Netherlands lies below sea level, sometimes as much as 22 feet (6.7 metres) below. On such land, building is only possible after driving piles down to the sand layer.

Limburg, in the border zone of the Ardennes, is the only area that is not part of the lowland. Indeed, in the entire country the only relief contrasts are provided by the Torenberg, in the Veluwe region (340 feet high); a 315-foot elevation around Nijmegen; and the highest point in The Netherlands, the Vaalserberg, rising 1,050 feet in the south of Limburg.

The Zuiderzee

The Zuiderzee was originally an estuary of the Rhine. It then became a shallow inland sea, biting deep into the land, and eventually was hollowed into an almost circular shape by the action of winds and tides. Only in 1932 did the barrier dam, running northeastward to connect the province of Noord-Holland with that of Friesland, finally tame and divide the sea into the outer Waddenzee and the inland IJsselmeer. In the latter, four giant polders were constructed around a freshwater basin fed by the IJssel River and linked with the sea by sluices. Ninety percent of the land thus reclaimed is arable, and it is parcelled out in large plots that are rented to farmers. After World War II, additional development made the polders available for industrial, residential, and recreational purposes as well as for agriculture.

Farther south, the disastrous spring tide of February 1, 1953, which flooded 400,000 acres (162,000 hectares) of land and killed 1,800 people, called into existence the Delta Plan, which aims to close the sea inlets of the southwestern province of Zeeland. These huge delta works are designed to shorten the coastline by 450 miles, combat the salination of the soil, and allow the development of the islands of Zeeland. The Oosterschelde (Eastern Scheldt) is already dammed; large sluices dominate the Haringvliet (river), and the embankment dikes along the Westerschelde (Western Scheldt) have been strengthened.

Glacial, riverine, and marine deposits

Soils. After the Ice Age (about 10,000 years ago), a large area in the north of what is now The Netherlands was covered by moraine (the glacial accumulation of earth and stones). In the centre and south, the Rhine and Maas unloaded thick layers of sand and gravel from the European mountain chains. Behind the coastal dunes, sea clay was deposited in the sheltered lagoons, and peat soil often subsequently developed in these areas. If the peat soil was washed away, sea clay was deposited again, and the alluvial layer had to stabilize through drainage and settlement.

Climate. The climate is temperate, with gentle winters, cool summers, and rainfall in every season. Southerly and westerly winds predominate, and the sea moderates the climate through onshore winds and the effect of the Gulf Stream.

The position of The Netherlands—between the area of high-pressure air masses centred on the Azores and the low-pressure region centred on Iceland—makes the country an area of collision between warm and polar air masses in summer, thus creating unsettled weather. In springtime, on the other hand, northern winds often cause night frosts. Winds meet with little resistance over the flat country, though the hills in the south diminish by more than half the 13-mile-per-hour wind velocity that prevails along the coast. Frost averages 60 days per year. July temperatures average about 63° F (17° C), and those of January average 35° F (2° C). The rainfall averages 31 inches (790 millimetres), with only about 25 clear days per year. The average rainfall is highest in summer (August) and autumn and lowest in springtime. The country is known—not least through the magnificent landscapes of Dutch painters—for its heavy clouds, and on an average day three-fifths of the sky is clouded.

Plant and animal life. The fauna of The Netherlands has been greatly affected by the activities of man; thus, the badger, hawk, river lobster, and marsh turtle have disappeared, while various fish species and the boar have been introduced, and the red-grouse population has been nurtured.

The Netherlands possesses the characteristic flora of the Atlantic district of the Euro-Siberian vegetation province to which it belongs. The flora is dominated by plants of the dunes (salt plants and water plants) and by plants of the woods, of cultivated land, and of meadows. In addition, the cultivation of ornamental plants, flowers, and bulbs has become a major industry.

The growth of urban centres

Settlement patterns. Urbanization in The Netherlands has taken place mainly in the 20th century. In 1900, more than half the population was still living in villages or towns of fewer than 10,000 inhabitants. By the late 20th century this figure had decreased to about 10 percent. The population of the cities and towns is growing much faster than the population of the rural areas, for migration to the towns more than exceeds the rural areas' higher birthrate. There has nevertheless been a decrease in the downtown population of the large cities. These inner-city regions are now becoming economic and cultural centres, their populations having spread outward in search of newer housing and greater living space in the suburbs. Urban-

MAP INDEX

Political subdivisions

Drenthe 52·50n 6·35e
Friesland 53·10n 5·50e
Gelderland 52·00n 5·50e
Groningen 53·15n 6·45e
Limburg 51·15n 6·00e
Noord-Brabant 51·40n 5·00e
Noord-Holland 52·35n 4·55e
Overijssel 52·15n 6·30e
Utrecht 52·00n 5·15e
Zeeland 51·25n 3·45e
Zuid-Holland 52·00n 4·40e

Cities and towns

Aalsmeer 52·16n 4·45e
Aalten 51·56n 6·35e
Alkmaar 52·37n 4·44e
Almelo 52·21n 6·39e
Alphen aan den Rijn 52·07n 4·40e
America 51·26n 5·59e
Amersfoort 52·09n 5·24e
Amstelveen 52·18n 4·51e
Amsterdam 52·22n 4·54e
Anna Paulowna 52·52n 4·50e
Apeldoorn 52·13n 5·58e
Appingedam 53·18n 6·52e
Arnhem 51·59n 5·55e
Assen 52·59n 6·34e
Axel 51·16n 3·55e
Baarn 52·13n 5·16e
Balk 52·54n 5·34e
Bergen 52·40n 4·41e
Bergen op Zoom 51·30n 4·17e
Berlikum 53·15n 5·39e
Beverwijk 52·28n 4·40e
Blerick 51·23n 6·10e
Bolsward 53·03n 5·31e
Borger 52·55n 6·46e
Borne 52·18n 6·45e
Boxmeer 51·39n 5·57e
Boxtel 51·35n 5·20e
Breda 51·35n 4·46e
Breukelen 52·10n 5·00e
Brunssum 50·56n 5·59e
Bussum 52·16n 5·10e
Cadzand 51·22n 3·25e
Callantsoog 52·49n 4·41e
Castricum 52·33n 4·39e
Coevorden 52·40n 6·45e
Culemborg 51·56n 5·13e
Cuyk 51·44n 5·52e
De Bilt 52·06n 5·10e
Dedemsvaart 52·36n 6·28e
Delft 52·00n 4·21e
Delfzijl 53·19n 6·46e
Den Helder 52·54n 4·45e
Deventer 52·15n 6·10e
Dieren 52·03n 6·06e
Dirkshorn 52·45n 4·45e
Doetinchem 51·58n 6·17e
Dokkum 53·19n 6·00e
Domburg 51·34n 3·30e
Dongen 51·37n 4·56e
Doorn 52·03n 5·21e
Dordrecht 51·49n 4·40e
Drachten 53·06n 6·05e
Driebergen 52·03n 5·16e
Echt 51·06n 5·52e
Ede 52·03n 5·40e
Egmond aan Zee 52·36n 4·37e
Eindhoven 51·26n 5·28e
Emmeloord 52·43n 5·45e
Emmen 52·47n 7·00e
Enkhuizen 52·42n 5·17e
Enschede 52·12n 6·53e
Epe 52·21n 6·00e
Ermelo 52·17n 5·37e
Etten 51·34n 4·38e
Flushing, see Vlissingen
Franeker 53·11n 5·32e
Geldrop 51·25n 5·33e
Geleen 50·58n 5·52e
Gemert 51·34n 5·40e
Genemuiden 52·37n 6·07e
Goes 51·30n 3·54e
Goirle 51·32n 5·04e
Goor 52·14n 6·35e
Gorinchem 51·50n 5·00e
Gouda 52·01n 4·43e
Grijpskerk 53·15n 6·18e
Groenlo 52·03n 6·38e
Groesbeek 51·47n 5·55e
Groningen 53·13n 6·33e
Haaksbergen 52·09n 6·44e
Haarlem 52·23n 4·38e
Hague, The, see The Hague
Harderwijk 52·21n 5·36e
Haren 53·10n 6·35e
Harlingen 53·10n 5·24e
Hattem 52·28n 6·04e
Haulerwijk 53·04n 6·20e
Heemstede 52·21n 4·37e
Heerenveen 52·57n 5·55e
Heerlen 50·54n 5·59e
Heiloo 52·36n 4·43e
Hengelo 52·15n 6·45e
Hillegom 52·18n 4·35e
Hilversum 52·14n 5·10e
Hoek van Holland (Hook of Holland) 51·59n 4·07e
Hoensbroek 50·55n 5·55e
Hollum 53·26n 5·37e
Holwerd 53·22n 5·54e
Hoogeveen 52·43n 6·29e
Hoogezand 53·09n 6·47e
Hook of Holland, see Hoek van Holland
Hoorn 52·38n 5·04e
Huizen 52·17n 5·14e
IJmuiden 52·27n 4·36e
IJsselstein 52·02n 5·03e
Kampen 52·33n 5·54e
Katwijk aan Zee 52·13n 4·24e
Kerkrade 50·52n 6·04e
Kootwijk 52·12n 5·45e
Krimpen aan den IJssel 51·54n 4·35e
Krommenie 52·29n 4·45e
Laren 52·15n 5·14e
Leerdam 51·54n 5·05e
Leeuwarden 53·12n 5·46e
Leiden (Leyden) 52·09n 4·30e
Lelystad 52·31n 5·27e
Lemele 52·27n 6·25e
Lemmer 52·50n 5·42e
Leyden, see Leiden
Lisse 52·15n 4·33e
Losser 52·15n 7·00e
Maassluis 51·55n 4·15e
Maastricht 50·52n 5·43e
Made 51·41n 4·46e
Medemblik 52·46n 5·06e
Meerseen 50·53n 5·45e
Meppel 52·42n 6·11e
Middelburg 51·30n 3·37e
Middelharnis 51·45n 4·11e
Midsland 53·22n 5·16e
Monster 52·02n 4·10e
Naaldwijk 52·00n 4·12e
Naarden 52·17n 5·09e
Nijkerk 52·13n 5·30e
Nijmegen 51·50n 5·50e
Nijverdal 52·22n 6·27e
Noordwijk-Binnen 52·13n 4·27e
Nunspeet 52·23n 5·46e
Oegstgeest 52·10n 4·29e
Oisterwijk 51·35n 5·12e
Oldenzaal 52·19n 6·56e
Oosterbeek 52·00n 5·50e
Oosterburen 53·24n 6·10e
Oosterend 53·05n 4·52e
Oosterhout 51·38n 4·51e
Ouddorp 51·18n 3·56e
Oudenbosch 51·35n 4·31e
Oss 51·46n 5·31e
Pieterburen 53·24n 6·27e
Purmerend 52·31n 4·57e
Putte 51·22n 4·23e
Putten 52·15n 5·36e
Renesse 51·44n 3·46e
Renkum 51·58n 5·45e
Rheden 52·01n 6·02e
Ridderkerk 51·52n 4·36e
Rijssen 52·18n 6·30e
Rijswijk 52·04n 4·20e
Roermond 51·12n 6·00e
Roodeschool 53·25n 6·45e
Roosendaal 51·32n 4·28e
Rotterdam 51·55n 4·28e
Sassenheim 52·13n 4·31e
Schiedam 51·55n 4·24e
Schijndel 51·37n 5·25e
's-Gravenhage, see The Hague
's-Hertogenbosch 51·41n 5·19e
Sittard 51·00n 5·53e
Sliedrecht 51·49n 4·45e
Slochteren 53·12n 6·50e
Smilde 52·56n 5·27e
Sneek 53·02n 5·40e
Soest 52·09n 5·18e
Stadskanaal 53·00n 6·55e
Steenwijk 52·47n 6·08e
Stein 50·57n 5·46e
Tegelen 51·21n 6·09e
Terneuzen 51·20n 3·50e
The Hague ('s-Gravenhage) 52·06n 4·18e
Tiel 51·54n 5·25e
Tilburg 51·34n 5·05e
Urk 52·39n 5·36e
Utrecht 52·05n 5·08e
Valkenswaard 51·21n 5·28e
Veendam 53·06n 6·58e
Veenendaal 52·02n 5·34e
Veghel 51·37n 5·33e
Veldhoven 51·24n 5·24e
Velp 52·00n 5·59e
Venlo 51·24n 6·10e
Venraij 51·32n 5·59e
Vlaardingen 51·54n 4·21e
Vlissingen (Flushing) 51·26n 3·35e
Volendam 52·30n 5·04e
Voorburg 52·05n 4·23e
Voorschoten 52·07n 4·27e
Vriezenveen 52·25n 6·38e
Vught 51·40n 5·17e
Waalre 51·24n 5·26e
Waalwijk 51·42n 5·04e
Waddinxveen 52·03n 4·40e
Wageningen 51·58n 5·40e
Wassenaar 52·07n 4·24e
Weert 51·15n 5·43e
Weesp 52·17n 5·02e
Westerbork 52·51n 6·37e
Wijchen 51·48n 5·43e
Willemstad 51·42n 4·26e
Winschoten 53·08n 7·02e
Winterswijk 51·58n 6·44e
Woerden 52·05n 4·54e
Wolvega 52·53n 6·00e
Wormerveer 52·28n 4·46e
Zaandam 52·26n 4·49e
Zaltbommel 51·48n 5·15e
Zandvoort 52·22n 4·32e
Zeist 52·05n 5·15e
Zevenaar 51·56n 6·05e
Zijpe 51·38n 4·05e
Zutphen 52·08n 6·12e
Zwijndrecht 51·49n 4·39e
Zwolle 52·30n 6·05e

Physical features and points of interest

Ameland, *island* 53·25n 5·45e
Brouwershavensche Gat, *channel* 51·45n 3·50e
De Hoge Veluwe, Nationale Park 52·02n 5·55e
Frisian Islands (Friese Eilanden) 53·25n 5·45e
Goeree, *island* 51·50n 3·55e
Haringvliet, *channel* 51·46n 4·12e
Het Loo, Paleis, *palace* 52·14n 5·56e
IJssel, *river* 52·30n 6·00e
IJsselmeer (Zuiderzee), *sea* 52·45n 5·25e
Lek, *river* 52·00n 6·00e
Maas (Meuse), *river* 51·49n 5·01e
Markerwaard, *physical region* 52·33n 5·15e
Meuse, see Maas
Noordoost Polder, *physical region* 52·42n 5·45e
Europoort, *dock* 51·58n 4·00e
Noordzeekanaal (North Sea Canal) 52·27n 4·33e
North Sea 53·30n 4·00e
North Sea Canal see Noordzeekanaal
Oostelijk Flevoland, *physical region* 52·30n 5·45e
Ooster Schelde, *inlet* 51·30n 4·00e
Rhine, *river* 51·52n 6·02e
Rottumeroog, *island* 53·30n 6·30e
Schiermonnikoog, *island* 53·28n 6·15e
Soestdijk, Paleis, *palace* 52·12n 5·15e
Terschelling, *island* 53·24n 5·20e
Texel, *island* 53·05n 4·45e
Twenthekanal, *canal* 52·13n 6·50e
Vaalserberg, *mountain* 50·46n 6·01e
Veluwe, *physical region* 52·10n 5·55e
Vlieland, *island* 53·15n 5·00e
Waal, *river* 51·55n 4·30e
Waddenzee, *sound* 53·15n 5·15e
Wester Schelde, *estuary* 51·25n 3·45e
Zuidelijk Flevoland (Zuidelijke IJsselmeerpolders), *physical region* 52·22n 5·20e
Zuidelijke IJsselmeerpolders, see Zuidelijk Flevoland
Zuiderzee, see IJsselmeer

ization is concentrated in areas of industrial growth, especially around the harbours of the Nieuwe Waterweg (New Waterway) centred on Rotterdam or the Noordzeekanaal (North Sea Canal) centred on Amsterdam.

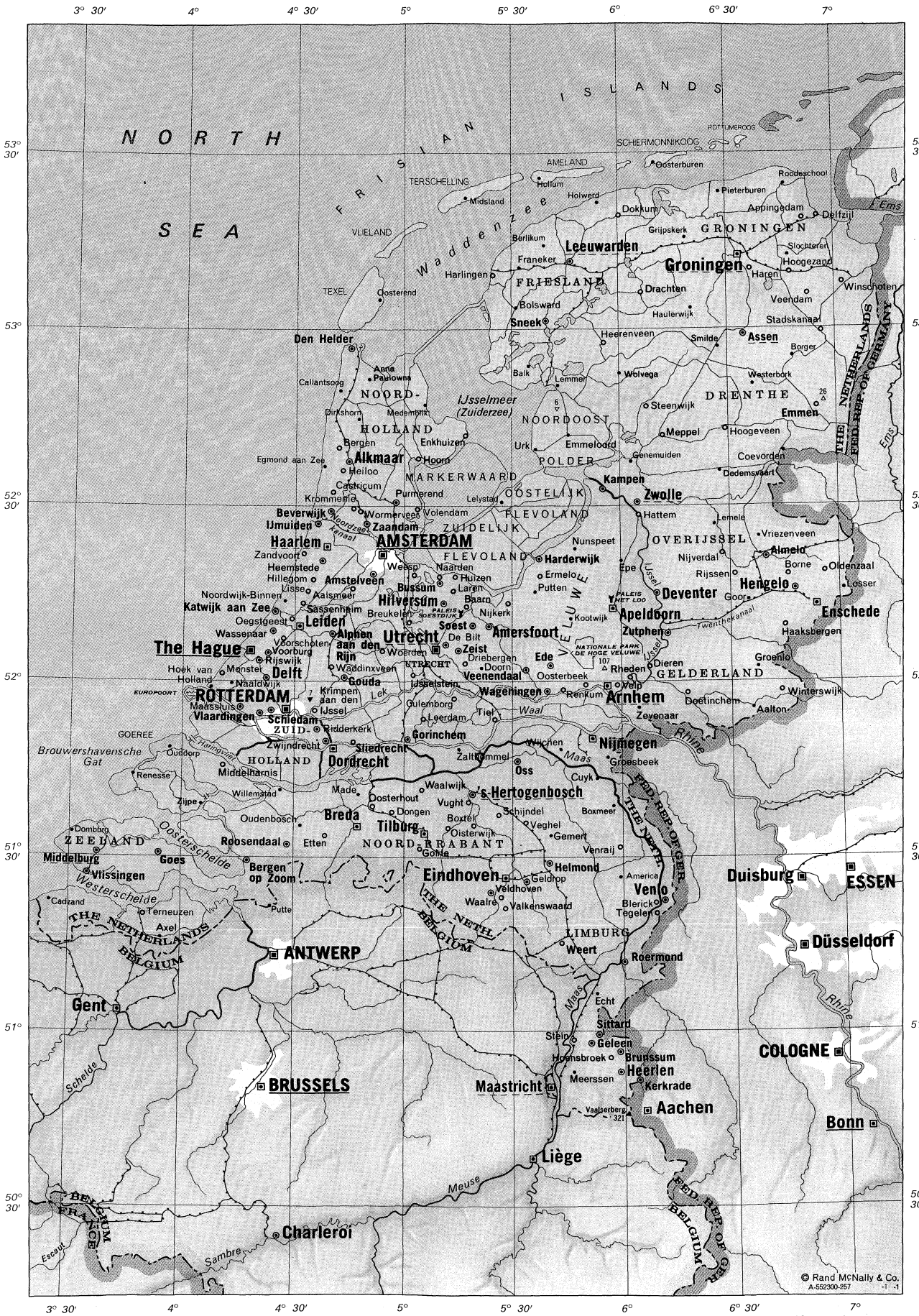
The bulk of the Dutch population is concentrated in the urban circle (known as the Randstad) comprising Rotterdam, The Hague, Leiden, Haarlem, Amsterdam, Hilversum, and Utrecht; in Noord-Brabant (in the cities of Breda, Tilburg, and Eindhoven); and in the Twente District around the city of Enschede. It is the policy to keep traditional towns and cities separated by strips of agricultural or recreational land. Even in the Randstad, authorities try to maintain a "green heart."

The people. *Ethnic groups and religions.* In origin, the Dutch people are a mixture of Frisians, Saxons, and Franks. Under the Romans, a German infiltration started that continued for centuries. The result was a predominantly Anglo-Saxon Frisia in the northeast, while the original Holland developed under Frankish influence.

Also of significance has been the attraction of countless refugees to the tolerant Dutch cities. Over the centuries, these have included southern Lowlanders, French Huguenots, Salzburgers, Swiss, and Portuguese Jews, who left their respective homelands as a result of religious or political pressures. In addition, many Germans and non-Iberian Jews have also tended to immigrate to The Netherlands for more purely economic reasons. In the 20th century, Dutch-Indonesians and nonwhite peoples from Ambon, in the East Indies, and Suriname, on the northeast coast of South America, added to the influx.

The heritage of Dutch Humanism

The Netherlands, a small state among big ones, has consistently preached freedom of thought. Dutch Humanism has always been a judicious mixture of piety and knowledge, epitomized in the development during the 14th and



THE NETHERLANDS

Size of symbol indicates relative size of town • ○ ◉ ▣ ■

Elevations in metres

15th centuries of a bourgeois civilization with a military and commercial spirit. These traditions, together with rigid Calvinism and a more tolerant Protestantism and Roman Catholicism, continue to characterize much of Dutch life. Political, educational, and territorial divisions are along religious lines. Not even World War II and the unity forged under German occupation were able to break up the pattern, and only late in the 20th century were there signs of change appearing.

Finally, even though they are parts of such a small country, individual regions have acquired their own distinctive personalities. The Zeelanders of the coast are often thought of as possessing a certain asceticism in their lifestyle developed through centuries of struggle against the sea—in contrast to the more easygoing, almost Burgundian French, way of life in the south.

Demographic trends. As a result of the postwar government policy mentioned above, emigrants exceeded immigrants by about 10,000 each year from 1947 to 1954. Thereafter, however, the economy and labour potential of western Europe attracted an increasing number of Spaniards, Yugoslavs, Turks, and Moroccans, so that the balance remained static. By the mid-1970s the growing number of arrivals from Suriname and the Netherlands Antilles resulted in an immigrant surplus.

Concentration of population in the two western provinces

Nearly half the population lives in the two western provinces of Zuid-Holland and Noord-Holland and in the central province of Utrecht. The four largest cities—Amsterdam, Rotterdam, The Hague ('s-Gravenhage), and Utrecht—are located in this area. In contrast, the three northern provinces—Groningen, Friesland, and Drenthe—and Zeeland province in the southwest are sparsely settled.

Figures on internal migrations suggest that the continuous emigration to the west from the east and north was reversed after 1970. Subsequent emigration was mainly from Zuid-Holland and Noord-Holland toward Utrecht and the less densely populated provinces, where the government has stimulated industrial growth—Groningen, Friesland, Drenthe, Gelderland, and Zeeland. In Limburg, the southernmost province, the closing of the coal mines between 1966 and 1974 contributed to emigration from the area.

The economy. Since World War II, The Netherlands has become a highly industrialized nation occupying a central position in the economic life of western Europe. Although agriculture accounts for a very small percentage of the national income, it remains a highly specialized contributor to Dutch exports. Because of the scarcity of mineral resources—with the important exception of natural gas—the country is dependent on large imports of basic materials.

Dutch trading partners

Trade relations are mainly with Europe and North America. The European Economic Community (Common Market) is, of course, the dominant factor in foreign trade. In 1958 (just before the Common Market was set up) only 41.5 percent of Dutch exports went to West Germany, Belgium, Luxembourg, France, and Italy, but in the late 20th century these nations accounted for about two-thirds.

Energy. With the increasing use of oil and natural gas, coal mining (concentratedin southern Limburg) was discontinued because of the rising cost of production. The Netherlands imports several million tons of coal annually to meet its needs.

The production of crude oil covers only a small part of Dutch requirements. The wells are located in southeastern Drenthe, near Schoonebeek, and in Zuid-Holland. Large amounts of crude oil are imported for refining in The Netherlands, and much of the refined petroleum is exported.

The discovery of natural gas during the 1950s had a tremendous influence on the development of the Dutch economy. The gas fields are near Slochteren, in Groningen, and beneath the Dutch sector of the North Sea. About half of the natural gas produced is exported to West Germany, Belgium, and France, helping to improve the balance of payments in the area in which The Netherlands has usually had its largest deficit. An aluminum smelter using natural gas as fuel was set up at Delfzijl, in Groningen. It became the core of an important industrialization project in a traditionally agricultural area; there is another aluminum smelter at Vlissingen, in Zeeland.

The use of coal for producing coke is no longer an economical proposition in The Netherlands, and coke-oven gas has been superseded by natural gas as fuel for the chemical industry of south Limburg.

Intense development of dairying and horticulture

Agriculture. The country's agricultural land is divided into grassland, arable farmland, and horticultural land, especially for tulips. Dutch cattle farming is highly developed: the milk yield per acre of grassland and the yield per cow are among the highest in the world. A good percentage of the total milk production is exported after being processed into such dairy products as butter, cheese, and condensed milk. Large-scale crop cultivation is hardly possible in The Netherlands, and most fodder as well as cereals for human consumption are imported from such countries as Canada and the United States.

Horticulture carried on under glass is of special importance. The export of hothouse tomatoes, cucumbers, and lettuce has greatly increased, and The Netherlands now contains almost half of the total European horticultural area under glass.

Fishing accounts for a small fraction of the national income. The delta works constructed along the coastline to protect against heavy spring tides limited the oyster culture and the shrimp and mussel fishing.

Industry. Dutch industrial development started relatively late, in 1870, and production rose even during the depression of the 1930s. Further development became a priority after World War II, when ascending population figures and growing farm-labour surpluses necessitated tens of thousands of additional jobs each year. A spectacular contribution to the solution of the employment problem was made by the chemical industry, whose annual output increased ten-fold between the late 1940s and the early 1970s—the greatest growth rate of any sector of Dutch industry.

The food, drink, and tobacco industries, based as they are on the highly productive agriculture sector, are the country's most important industries. The petroleum, chemical, and electronics industries are the other major manufacturers. Metal manufactucturing, formerly the most important industry, declined in the late 20th century in part because of increasing production costs. Dutch industry has been able to redevelop commercial aircraft since World War II, and automobile manufacturing has been expanded. In contrast, employment in textile manufacturing and shipbuilding has decreased. The government has encouraged new industrial development in the fields of microelectronics, aerospace, and biotechnology.

Banking

Finance. Commercial banking in The Netherlands is in the hands of a few large concerns. The Algemene Bank Nederland and the Amsterdam-Rotterdam Bank handle the majority of Dutch financial business between them, while the Coöperatieve Centrale Raiffeissen-Bank en Aangesloten Banken and the Coöperatieve Centrale Boerenleenbank specialize in agricultural credit. The many independent savings banks are centred on the state-owned National Savings Bank, which is operated by the Post Office. Since 1948 the central bank, the Nederlandsche Bank, has also been state owned. Its function, defined by the 1948 Banking Act and the 1956 Credit Control Act, is to supervise the entire system. An independent agency, it has authority to strengthen traditional monetary controls.

The Dutch economy nevertheless remains basically in private hands. Postwar industrial development has been consciously stimulated by government policy, and state subsidies have been granted to attract industry toward the relatively underdeveloped northern provinces, as well as Limburg and certain other pockets of economic stagnation. Despite these subsidies, the western provinces remain the centre of new industry, especially in the services sector.

Management of the economy. The state completely owns such corporations as Netherlands Railways and the Dutch State Mines and is a major participant in others: the Royal Netherlands Blast Furnaces and Steel Works, the Royal Dutch Airlines (KLM), De Nationale Investeringsbank (Herstelbank), and the Gas Union, which distributes natural gas. The state thus has a direct, continuous

influence on the management of some sectors of the economy and especially on investment policy. The government also employs a substantial percentage of the total labour force. Dutch employers are organized mainly in three separate but closely cooperating organizations: one Catholic, one Protestant, and one nondenominational. Labour has a similar tripartite organization, comprising Catholic, Protestant, and Socialist groups with some other independent fringe organizations. These employer organizations and labour unions are represented on the Joint Industrial Labour Council, established in 1945 as a collective bargaining vehicle, and on the Socio-Economic Council, which serves mainly as an advisory body to the government.

Transportation. Transportation is of special importance in The Netherlands, since the country functions as a gateway for goods traffic between western Europe and the rest of the world. Goods flow through Dutch harbours, continuing their passage by river vessel, train, truck, pipeline, and—a more recent development—by modern container units.

The maritime traffic of The Netherlands is about equal to the land traffic, and indeed the whole southern part of the North Sea may be likened to an immense traffic square, fed as it is by the Thames, Rhine, Maas, and Scheldt, with links into the hinterland of the Continent that make it the greatest commercial artery of the world after those of North America. Rotterdam has the country's best equipped modern harbour. Europoort, the region between Rotterdam and the sea, can easily be reached by the biggest oceangoing ships; it serves as an approach to Rotterdam harbour, itself handling more tonnage than any other harbour in the world. Rijnmond (the mouth of the Rhine) and IJmond (the mouth of the IJ) are also equipped to receive big ships. In oil processing, too, Rotterdam is one of the world's biggest centres. A trench about 1,500 feet wide and 75 feet deep extends eight miles into the North Sea to facilitate the entry of supertankers with 200,000-ton capacities. The number of rivercrafts is probably unsurpassed by any other country.

Royal Dutch Airlines (KLM) initiated a scheduled service between Amsterdam and London in 1920 and now has a worldwide network. One subsidiary operates an airline in the West Indies, while another runs a modest domestic network between Amsterdam and the outlying provinces. Schiphol, Amsterdam's airport, and Zestienhoven, the airport at Rotterdam, both are of international importance. The Dutch government conducts a school for civil aviation and pilot training.

As might be expected in so thickly settled and small a country, the road network is comprehensive.

The continued importance of waterways

Despite the competition offered by railroads, motor freight, containers, and pipelines, it appears that the important waterways will never be surpassed for traffic in freight. For this reason, much attention and financial support are devoted to the waterways in order to meet the growing international demand for efficient goods transport.

The Netherlands has about 3,000 miles of inland waterways linked with Belgian, French, and German systems. Besides such natural waterways as the rivers Maas, Rhine, Waal, and Lek, many artificial waterways—the Juliana Kanaal, the Amsterdam–Rijn Kanaal, and the Maas-Waal Kanaal—connect the major ports on the coast with the hinterland. The state-owned railroads are modest but well developed and in total length are about average for Europe. The combined effect of inland navigation and freight traffic on roads has prevented the expansion of goods traffic by rail. More than half of the rail network is electrified; elsewhere, steam has been replaced by diesel.

Administrative and social conditions. The Kingdom of The Netherlands is a constitutional monarchy. The monarchy is hereditary in both the male and female lines; the constitution was amended in 1980 to allow for succession of the eldest child, regardless of its sex.

Government. The constitution, which dates from 1814, maintains that the head of state is inviolable, but it also embodies the concept of ministerial responsibility; it further provides that no government can remain in power against the will of parliament. The States-General (Staten-Generaal), as parliament is officially known, consists of two chambers: a First Chamber, whose members are elected by the members of the councils of the provinces, and a directly elected Second Chamber. Both chambers share legislative power with the government (officially known as the crown, defined as the head of state acting in conjunction with the ministers) and control government policy, but, whereas the First Chamber has only the power of veto in the legislative process, the directly elected Second Chamber, controlling the government, plays a dominant role.

Formation of the government

Every four years, after elections for the Second Chamber have been held, the government resigns, and a process of bargaining starts between various parliamentary leaders who hope to form a government that will control a majority in parliament. It usually takes a few months before a *formateur,* as the architect of such a coalition is known, can accept a royal invitation to form a government; the head of state then formally appoints the ministers. In the event of political crises resulting in the fall of the government before the end of a four-year period, the same process of bargaining takes place. The monarch, however, has the right to dissolve one or both chambers, at which time new elections are held.

In local government, the most important institutions are the municipalities. Their jurisdiction extends over the entire country (except for some areas of recently reclaimed land). Each municipality is run by a directly elected council of between seven and 45 members, depending on the size of the population. This council is presided over by a *burgemeester* (mayor), who is appointed by the government and serves as chairman of the executive, members of which are elected by and from the council. Where the councils' own ordinances are applicable, the municipalities are autonomous, and in many instances national legislation or provincial ordinances provide for the cooperation of municipal authorities.

The country is also divided into provinces: Noord-Holland, Zuid-Holland, Zeeland, Utrecht, Noord-Brabant, Limburg, Gelderland, Overijssel, Drenthe, Groningen, and Friesland. Their administrative system has the same structure as the municipal one: directly elected councils (*staten*) who elect the members of the executive, except for the chairman, who is appointed by the government. The main function of the provinces consists of controlling the municipalities within their borders, as well as controlling the district water-control boards (*waterschappen*).

Political life. The Second Chamber, the provincial councils, and the municipal councils are elected according to a system of proportional representation. In general elections for the Second Chamber, it can take as little as 0.66 percent of the overall vote to get one of the seats in the chamber. As a result, a large number of parties and political movements are represented in parliament.

Objections to the political system

Because of the complicated party system and the time-consuming and complicated process by which a government comes into being, constitutional and party reforms are advocated by many people. Dissatisfaction with "the system" expresses itself by widespread utterances of frustration about official politics, often embodied in extra-parliamentary movements.

Justice. In The Netherlands, the ordinary administration of justice is entrusted exclusively to appointed judges; trial by jury is unknown. There are cantonal courts (*kantongerechten*), which exercise jurisdiction in a whole range of minor civil and criminal cases. More important cases are handled by one of the district courts (*rechtbanken*), which also can hear appeals over cantonal court decisions. Appeals against decisions from the district courts are heard by one of five courts of appeal (*gerechtshoven*).

The Supreme Court (Hoge Raad) ensures a uniform application of the law, but it cannot suspend a law as being contrary to the constitution. In the legislative process itself, the government and the parliament together pass judgment on the constitutionality of a bill under consideration. Laws that are contrary to an international agreement to which The Netherlands is a party can be abrogated by the courts.

Armed forces. The Dutch armed forces consist of an

army, a navy, and an air force. All male citizens become liable to military service at the age of 18. The normal period of obligatory service is between 14 and 17 months, depending on the rank and the branch of service. The Netherlands is a member of NATO.

Education. The Dutch educational structure consists of a public and a private system. Private schools, as long as they conform to legally determined standards of such things as curriculum, examinations, and qualifications of personnel, are financially subsidized up to the highest level, according to the same rules in force for the public system. With a few exceptions, private schools are run by religious organizations, either Protestant or Catholic. Higher education is provided by state universities at Leiden (established in 1575), Groningen (1614), Utrecht (1636), and Limburg (1975); other universities include that of the city of Amsterdam, which is the largest (1632), and two private universities, the Protestant Free University (Vrije Universiteit) in Amsterdam (1880), and the Catholic University of Nijmegen (1928). The Erasmus University of Rotterdam was founded in 1973 by amalgamation of the former School of Economics and Faculty of Medicine.

In addition, higher technical education at the university level takes place in special institutions in Delft, Eindhoven, and Enschede. A private institution for the teaching of economics at university level is in Tilburg (Catholic). All these are financed by the state, and, although different rules have been considered, the right of entrance is determined by final examinations in secondary schools.

Housing. Since World War II, housing policy has been one of the most controversial topics of political and social life. Because of war damage and a stagnant building trade, a considerable housing shortage had developed by the early 1950s. The rate of population growth, moreover, had accelerated, and rising living standards meant that people expected more space per person and generally higher quality in housing facilities.

The government's role in housing

Construction is heavily subsidized in The Netherlands. Subsidies are paid to municipalities and building associations, which are mainly concerned with the housing needs of lower income groups, or to private persons who conform to certain building standards. In urbanized areas, rent control is still maintained. Local authorities have power to set up a system of housing allocation, a power particularly exercised in metropolitan areas, where the housing shortage is most acute.

Social welfare. The Netherlands has an extensive system of social insurance. For instance, the state has direct supervision of old-age, widows', and orphans' pensions, to which all citizens are entitled. The enforcement of the Health Insurance Act, the Working Incapacity Insurance Act, and the Unemployed Benefit Act is entrusted to special associations set up by employers' and employees' organizations. The last two acts cover all employees, whereas participation under the Health Insurance Act is compulsory only for employees below a certain wage level, determined yearly according to the general level of wages and prices. Employees and self-employed persons also receive children's allowances, though rules and amounts differ for each group.

In 1968 a special act was passed to cover certain specialized or prolonged medical expenses for all citizens. The great majority of the population is covered by the Health Insurance Act, and most other persons are more or less covered by some form of private health insurance. A guaranteed minimum income is also maintained.

Cultural life. The cultural climate in The Netherlands is varied and lively. There are many state-supported museums, the most famous of which are the Rijksmuseum at Amsterdam, the Mauritshuis at The Hague, and the Stedelijk Museum, also in Amsterdam; the first is noted for its magnificent collection of works by the great 17th-century Dutch masters (especially Rembrandt van Rijn) and the last is known for its collection of contemporary paintings. Two museums, the Rijksmuseum Vincent van Gogh in Amsterdam and the Rijksmuseum Kröller-Müller in Otterlo, are renowned for their collections of paintings by Vincent van Gogh. Modern museums, with their progressive policies, attract large sections of the public. There are several hundred public collections of art and culture.

Connected with various cultural institutions are the Preservation Service for Monuments, which restores valuable buildings and protects archaeological treasures; the State Office for the Documentation of the History of Art; and the Central Laboratory for Art and Science.

The history of painting could scarcely be written without including, besides those of Rembrandt and van Gogh, the names of the great painters of Holland, such as Jan van Eyck, the founder of the Flemish school in the 15th century; the allegorical Hieronymus Bosch in the 16th; portraitists such as Frans Hals and landscapists such as Jan Vermeer, Aelbert Cuyp, and Jacob van Ruisdael in the 17th; the 17th-century still-life artists such as Willem Heda and Willem Kalff; and the geometrically inclined Piet Mondrian in the 20th. Interest in painting in The Netherlands is still widespread, and such artists as Kees van Dongen, Willem de Kooning, and Karel Appel are internationally famous and their works much in demand. Sculpture, on the other hand, has always been overshadowed by the more prominent pictorial developments. Thanks to a complicated system of public assistance and scholarships, young artists can indulge in their experiments with the certainty of a minimum income, while the government and municipalities regularly buy their works to adorn various administrative centres.

The problems of Dutch theatre and literature

Dutch theatre and literature have always been handicapped by the very tiny proportion of the human race that speaks Dutch. This limits the possibility of Dutch writing making a worldwide impact; except for the translation of some Dutch novels and poems, Dutch literature, though very much alive, has received little of the attention accorded the universal visual vocabulary of Dutch painting.

Theatre companies are all private foundations, though the state and the municipalities provide financial assistance. Players' salaries are legally determined, and performances of Dutch plays are encouraged, though the administration is not concerned with artistic quality.

Young Dutch filmmakers, less handicapped than their writing colleagues, have chalked up many successes abroad, notably with documentaries about Dutch painters and with feature films.

Musical life is intense. Although private orchestras in The Netherlands are supported by the government, the latter makes no attempt to interfere in the operation of these organizations or the choice of repertory. The Amsterdam Concertgebouw orchestra is world-famous, and Dutch composers are encouraged. Other organizations, such as National Ballet, the Netherlands Dance Theatre, several choirs, and music schools, are also subsidized.

The first European newspapers were printed in Amsterdam in 1620, and the Dutch press has a reputation for high-quality reporting; yet, despite ample advertising, it is plagued by financial depression. This situation stems from rising costs of newsprint and printing and the competition of radio and television in both news reporting and advertising.

The Dutch broadcasting system is complicated and may be considered unique, but above all it is democratic. Eight organizations, all under private initiative and with ideological aims, account for the majority of radio and television broadcasting. These Catholic, Protestant, Christian, Socialist, and neutral stations all operate within the framework of the Netherlands Broadcasting Foundation (NOS), which is responsible for the actual transmission and for the joint programs filling the rest of the time. Church societies, political parties, and small representative groups are granted some reserved broadcasting time.

The government itself has no influence on the programs, and even permission to advertise is granted to a separate foundation, with a limited allocation of time. All broadcasting is financed by a licensing fee and by the yield from television and radio advertisements. These sums are paid into the Ministry of Cultural Affairs, Recreation and Social Welfare, which in turn pays the cost of the transmitters and allots the broadcasting systems a fixed amount per hour for radio and television programs. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.* (J.M.Lu./J.P.A.G.)

HISTORY

The early history of the Low Countries, including the area now occupied by the Kingdom of The Netherlands, is treated above in the general section on the Low Countries. Treated in this section is the history of the modern state of The Netherlands, beginning with its establishment in 1579.

The Union of Utrecht. On January 23, 1579, an agreement was concluded at Utrecht forming a "closer union" within the larger union of the Low Countries led by the States-General sitting in Brussels. Included in the Union were the provinces and cities committed to carrying on resistance to Spanish rule: Holland, Zeeland, Utrecht, Gelderland (Guelders), and Zutphen (a part of Overijssel) as the first signatories, followed in the next year by the whole of Overijssel, most of Friesland, and Groningen, all in the north; and in the south by the cities of Antwerp and Breda in Brabant, and Ghent, Bruges, and Ypres in Flanders. Designed to establish a league for conduct of the war of independence and ultimately to strengthen the central government in Brussels, the "Union of Utrecht" became in fact the foundation of a separate state and a distinct nation in the northern Netherlands. The new state was named the United Provinces of the Netherlands, or, more briefly, the "Dutch Republic," and was known in the international community as the "States-General." The people of the northern Netherlands began to be distinguished from the inhabitants of the south (to whom the name of Flemings continued to cling) by the appellation Hollanders (French Hollandais, Italian Olandese, German Holländer, etc.), after their principal province. The English, however, came to apply exclusively to the Hollanders the name of Dutch, which previously they had applied to all German speakers (from German Deutsch, Dutch Duits). The name Netherlanders, which remained in use in the Low Countries for both the inhabitants of the United Provinces specifically and for all those, north or south, who spoke Dutch (Nederlands), passed out of currency in most foreign countries or came to be restricted to the northerners. The transformation had a price: the erosion of the bond of historical identity between northerners and southerners, Dutchmen and Belgians as they would be called in the 19th and 20th centuries.

The treaty that formed the basis of the new northern union established a military league to resist the Spaniards but on a "perpetual" basis, and it provided for closer political arrangements among the provinces than those of "allies" in the ordinary sense. The provinces united "for all time as if they were a single province"; each remained sovereign in its internal affairs, but all acted as a body in foreign policy. Decisions on war and peace and on taxation could be made only unanimously. The Union did not throw off the formal sovereignty of the king of Spain, but it confirmed the effective powers of the provincial stadholders (formally the "lieutenants," or governors, of the king) as their political leaders (there was no "stadholder of the United Provinces," as foreigners often assumed). The Union moved away from the religious settlement embodied in the Pacification of Ghent of two years before (see above) and toward a predominance of the Calvinists and their monopoly of public practice of religion in the key provinces of Holland and Zeeland.

The immediate political significance of the Union was that it dovetailed with the Union of Artois, concluded earlier in the month, which began the reconciliation of the southerners with Philip II. The two "unions," parallel but opposite, thus undermined William's policy of collaboration between Catholics and Calvinists throughout the Low Countries in resistance to the Spanish domination, which required mutual toleration among the religions. But it took some time before the "general union," with its base in the States-General at Brussels, fell apart irrevocably. For another half decade the Prince struggled to keep intact the broader union and at the same time to assure its military and political support from abroad. Although Archduke Matthias of Habsburg, named governor general by the States-General in 1577 after the deposition of Don Juan, remained the formal head of state until 1581, the Prince continued to exercise his leadership. That the Prince was the head and heart of the rebellion was recognized by Philip II in 1580 when he put him under the ban of outlawry. William's *Apology* in defense of his conduct was followed in 1581 by the Act of Abjuration (literally, "Act of Abandonment," Akte van Verlating) by which the States-General declared that Philip had forfeited his sovereignty over the provinces by his persistent tyranny. This was a declaration of independence for the whole of the Low Countries, but the military and political events of the next decade limited its permanent effect to the northern provinces under the "closer union" of Utrecht.

Declaration of independence

Foreign intervention. Yet independence did not become William's objective even after the proclamation of the Act of Abjuration. Archduke Matthias returned home in 1581 after Orange turned to the Duc d'Anjou, who agreed to take over the "lordship" of the Low Countries in 1580. The Prince hoped for assistance from the Duc's brother, King Henry III of France, and considered the "lordship" of Anjou as only a kind of limited, constitutional "sovereignty" like that which the rebels had hoped to impose on Philip II at the beginning of their rising. Anjou, however, saw the lordship as a means to total dominion over the Netherlands. Irritated by restraints upon his authority, he even attempted the seizure of power by military force, resulting in the so-called French Fury of January 17, 1583, when his troops tried to capture Antwerp. The coup misfired, but Orange managed to keep Anjou (who returned to France) in his post despite the outraged feelings of the Netherlanders. Holland and Zeeland were on the verge of offering the title of count to William when he was assassinated on July 10, 1584, at Delft, by Balthasar Gérard, a fanatical young Catholic from Franche-Comté, spurred by the promises of the ban of Philip II. William's death did not end the rebellion, as Philip had hoped, but it did result in the almost unnoticed disappearance of the central government in Brussels. The States-General, which now met at The Hague in Holland, represented only the provinces in the Union of Utrecht.

With the Spaniards steadily overrunning Flanders and Brabant, the Dutch in their plight did not immediately abandon William's policy of seeking foreign assistance. But after Henry III of France and Elizabeth of England both refused sovereignty over the country, the States-General in 1586 named as governor general Robert Dudley, earl of Leicester, whom Elizabeth had sent to command Dutch and English auxiliary forces against the Spaniards after the fall of Antwerp. Leicester, like Anjou before him, endeavoured to make himself absolute master of the country, relying on the support of popular Calvinism and of the outlying provinces that were jealous of Holland to create a strong centralized government under his authority. Holland thwarted Leicester's efforts, which culminated in an attempted invasion of Holland from Utrecht in 1587. With Leicester's departure, the United Provinces put aside all efforts to obtain a foreign protectorate and stood forth as an independent state.

Leicester as governor general

The formation of a new government. Although derived from historical institutions, the government of the United Provinces was in practice largely a new set of institutions, not created, but confirmed, by the Union of Utrecht. Their primary force lay in the provinces, seven in number (Holland, Zeeland, Utrecht, Gelderland, Overijssel, Friesland, and Groningen), which were ruled by assemblies of provincial States representing the towns and the landed nobility. Although the stadholders (who after a few years came to be drawn exclusively from the House of Orange) were elected by the States of the provinces, they at the same time possessed important prerogatives in the selection of members of the town governments from which the provincial assemblies ultimately derived their authority, and they were the acknowledged military leaders of the republic. Central government passed from the Council of State to the States-General, which was more explicitly subordinated to provincial authority. Although it conducted the military and diplomatic work of the republic, the States-General failed to obtain effective rights of direct taxation (except for import and export duties assigned to the admiralties), and its major decisions were taken under the rule of unanimity. In practice the province of Hol-

Dominance of Holland

land, by far the wealthiest province in the Union and the contributor of more than half of the revenues of the central government, became the preponderant political force in the country, along with the stadholders of the House of Orange. The relationship between Holland and Orange governed the republic's politics for the two centuries of its existence. As collaborators, Holland and the princes of Orange could make the clumsy governmental system work with surprising effectiveness; as rivals, they imperilled its potency as a state, at least until one or the other emerged a temporary victor: but neither force was able to rule permanently without the other.

The decades immediately after 1587 were marked by close collaboration between Johan van Oldenbarnevelt, "advocate" of Holland (the legal and executive secretary of the provincial States), and Maurice of Nassau, William the Silent's second son (the first, Philip William, became prince of Orange and remained loyal to Spain), who was named stadholder of Holland and Zeeland and became the commander of the republic's armies. The result was a series of military triumphs over the Spanish forces under Parma. Maurice recaptured the Dutch territories north of the great rivers and extended them southward into much of Brabant and enough of Flanders to cut off Antwerp from the sea. These victories are recorded in the historical memory of the Dutch as "the closing of the garden," the territory that became the republic of the United Provinces and then (with a few additions) the modern Kingdom of The Netherlands. These victories were accompanied by the emergence of the States-General diplomatically recognized by England and France as an independent state.

Ascendancy of the Dutch economy. The military prowess of the fledgling republic rested upon the wealth of Holland—which managed in wartime to maintain and extend its trade to all of Europe and, after the turn of the century, even to the Far East. Amsterdam replaced Antwerp, the great port on the Scheldt River, as the principal warehouse and trading centre for all Europe, even while Holland maintained the leadership in shipping it had already garnered during the 16th century. The foundation of Dutch economic prosperity lay in the fishing and shipping industries. Even during the period of Antwerp's ascendancy, ships from Holland and Zeeland had carried a large portion of the goods that passed through the Scheldt, and now that Amsterdam had taken over from Antwerp, Dutch shipping only expanded its predominance. Dutch fishermen had harvested the North Sea for centuries, and the salted cargoes were sold widely through western and central Europe. Dutch trade benefitted, as had that of Flanders, from the location of the country at the nexuses of the great north–south and east–west trade routes of Europe. To these were added the route to the East Indies early in the 17th century. Amsterdam and the lesser ports of Holland and Zeeland became the principal European suppliers of grain and naval stores from the Baltic, to which they shipped manufactured goods and wines from the south. Germany's principal exports were now shipped down the Rhine, as Dutch ports replaced the Hanseatic towns of northern Germany. The bulk of French exports were carried in Dutch ships, and even Spain and Portugal depended on the Dutch for grain and naval stores (thereby enabling the Dutch to finance their war of independence).

Amsterdam supplants Antwerp

Banking

During the 17th century the Dutch assumed a major role in supplying grain and other northern commodities to the countries of the Mediterranean and also became the principal importer of spices and other luxury goods from the East. Even England relied to a great extent upon Dutch shipping. The Dutch advantages lay not only in their situation but also in the efficient design of their bulky flyboats (*fluiten*), manned by small crews at less cost than any of their competitors. Modern banking institutions developed to meet the needs of the vastly expanding trade. Amsterdam's "exchange bank" was instituted in 1609 to provide monetary exchange at established rates, but it soon became a deposit bank for the safe settling of accounts. Unlike the Bank of England, established almost a century later, it neither managed the national currency nor acted as a lending institution (except to the government in emergencies). Private bankers met the need for credit, as well as acting as brokers in financial transactions. The need for commercial exports, as well as a growing population at home, spurred industry in many towns. Although the shipbuilders on the Zaan, northwest of Amsterdam, and the sugar refiners in particular developed large-scale operations, sometimes including machinery, Dutch industry generally remained at the level of handicraft production. Dutch agriculture also responded to the new needs. Grain production was encouraged in the inland provinces, while the polderland of the delta islands and along the coast was drained and diked, providing pastureland for a flourishing dairy industry.

The Twelve Years' Truce. The Twelve Years' Truce that began in 1609 arose out of political controversies that were to dominate the republic for the next two centuries. The collaboration between the House of Orange and the leaders of the province of Holland, which had thwarted Spain in its reconquest of the Netherlands north of the great rivers, was replaced by an intermittent, but often fierce, rivalry between them, in which the other tensions of Dutch political life were reflected and incorporated: the jealousy among the lesser provinces of a Holland that they considered too wealthy, too mighty, and too arrogant but which they knew they needed for their own defense; the misunderstanding between maritime and landward provinces; the annoyance of landed noblemen that they were dependent upon the goodwill of burghers in Holland (they preferred the prince of Orange, whom they saw as one of themselves); the resentment of the popular classes, men of small property and of none, toward the town regents, against whom they looked to the princes of Orange to protect them; and the antipathy of the Reformed clergy toward the regents, who obstructed their desire to make the state serve the church. The debate over whether or not to conclude a peace with Spain mingled these various interests with that of the House of Orange, partly because Maurice opposed peace, partly because it involved making some compromise with Spain, and partly because it would mean a reduction of his influence in the state; but the province of Holland in particular, under Oldenbarnevelt's leadership, felt that the independence and security of the United Provinces had been sufficiently assured to permit a reduction of the immense expenditures for the war. When Spain reduced its immediate proposal to a truce rather than permanent peace, agreed to treat the United Provinces as independent and sovereign, which was just short of outright recognition, and put aside efforts to win guarantees for Dutch Catholics, the pressure for conclusion of a truce could not be withstood.

But the Twelve Years' Truce did not end internal controversy within the republic; if anything, it only sharpened Maurice of Nassau's opposition to Holland and Oldenbarnevelt. The staunch Calvinists endeavoured to hold the Reformed Church to the strict orthodoxy expounded by Gomarus, a Leiden professor of theology, against the broader, less rigorous tenets upheld by his colleague Arminius. The Gomarists demanded the government uphold their principles because the Reformed Church was the only true church, but they reserved for themselves the right to declare what the correct doctrines were; and they vigorously asserted that other religious groups, Catholic, Protestant, and Jewish alike, should be suppressed or at least penalized and restricted. On the other hand, the Arminians had the support of the leaders of Holland and a majority of its towns, who felt that what was in effect the state church had to be under the authority of the government. Both out of principle and out of a desire not to hamper trade with men of all religions, they favoured a broadly inclusive Reformed Church and toleration for those outside its ranks.

Gomarists versus Arminians

The efforts of Gomarists to seize churches for their own use in defiance of town authorities led to incipient civil war. Maurice broke openly with the dominant party in Holland when it attempted to set up little provincial armies in Holland and Utrecht. In 1618, he acted under the authority of the States-General, in which the majority of provinces favoured the Gomarists (now called the Contra-Remonstrants because they had opposed an Arminian petition) over the Remonstrants (Arminians) to

crush the resistance of Oldenbarnevelt's party. Oldenbarnevelt, together with two of his chief supporters in Holland (including the great jurist Grotius) and an ally in Utrecht, was arrested and tried for treason by a special court instituted by the States-General. The defendants affirmed that they were subject only to the authority of the sovereign province that they served. The sentence, which to foes of the House of Orange over the centuries became an act of judicial murder, sent Oldenbarnevelt, then 71 years old, with almost four decades of service as Holland's leader, to his death by beheading in May 1619. Grotius and another defendant (the third had committed suicide) were sentenced to life imprisonment, although Grotius escaped, sensationally, a few years later.

Execution of Oldenbarnevelt

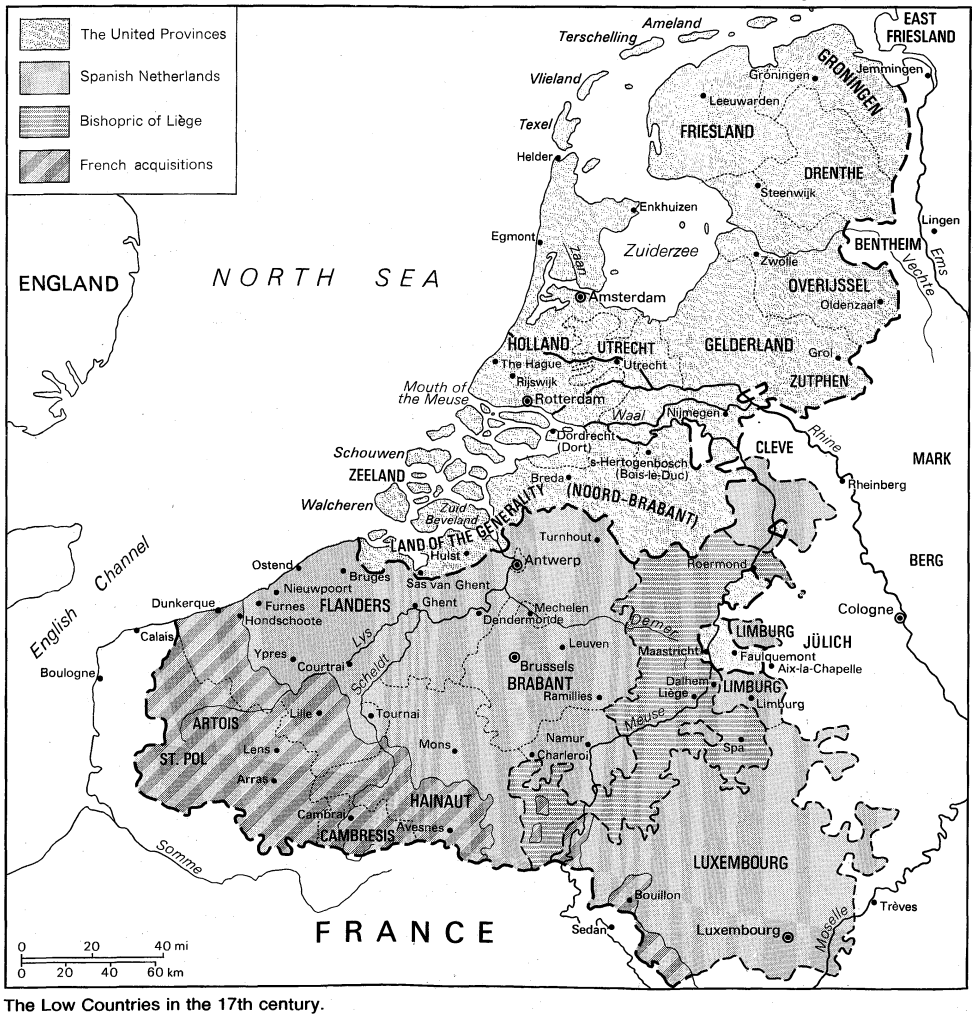
During these fateful months, the Reformed Church held a national synod at Dordrecht. Dominated by the Contra-Remonstrants, the synod expelled the Remonstrants, reaffirmed the doctrines of the church along Gomarian lines, and ordered the preparation of a new translation of the Bible (the famous States Bible that consolidated the Dutch language much as the contemporary King James Version consolidated English). The triumph of Maurice and the Contra-Remonstrants meant that war with Spain would be a virtual certainty upon the expiration of the Twelve Years' Truce in 1621—all the more because the Spanish authorities in the southern Netherlands insisted upon including rights for Dutch Catholics in a permanent treaty and even sought an acknowledgment by the States-General of the nominal overlordship of the king of Spain. Maurice did not use his new uncontested power to reform the complicated incoherence of the Dutch constitution; the structure of government and the distribution of formal power remained the same. Maurice was not a politically minded ruler and was satisfied as long as he had his way in military matters. The United Provinces continued essentially republican in character.

War with Spain, 1621–48. The war resumed in 1621 under Maurice's leadership. But his touch of victory was gone, and the republic appeared in danger when the great fortress of Breda on the southern frontier fell to the Spaniards in 1625. Only a few weeks before, Maurice had died. The danger was all the greater because the Austrian Habsburgs, in alliance with their Spanish cousins, were waging a successful struggle against their Protestant foes in Germany in the first stages of the Thirty Years' War. But Maurice's half-brother, Frederick Henry, who succeeded him as prince of Orange, stadholder, and commander in chief, resumed the course of victory. He completed the recapture of the towns recently gained by the Spaniards and extended the territory under the States-General to the key fortress of Maastricht on the Maas (Meuse) well to the south. At the same time, the Dutch navy won a series of victories over the Spaniards, including Piet Heyn's celebrated capture of their silver fleet off the coast of Cuba (1628) and the destruction of a Spanish fleet in the Downs, off the English coast, by Maarten Tromp in 1639.

Victories under Frederick Henry

Frederick Henry turned out to be a more subtle and purposeful politician than Maurice. On the one hand, he ended the suppression of the Remonstrants, with whose religious views he sympathized, without exasperating the Contra-Remonstrants beyond repair. On the other hand, he established a firm grip over the policies of the republic, notably by establishing a close alliance with France aimed at the joint conquest of the Spanish Netherlands. Frederick Henry's political predominance within the republic was based upon his control of the lesser provinces, which had a majority in the States-General and which could outweigh the influence of Holland. Gradually Holland turned against

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Medieval and Modern*, 9th ed. (1962); George Philip & Son Ltd., London



The Low Countries in the 17th century.

him, especially after he arranged the marriage of his young son William to Princess Mary Stuart, daughter of Charles I of England, on the eve of the English Civil War (1642–51). This fateful dynastic bond tied the interests of the House of Orange to the royal families of England, first to the Stuarts and later to the Hanoverians. The position of the House of Orange, however, was elevated by the connection: the French monarchy granted Frederick Henry the honorary address of "His Highness," normally restricted to royalty; and the debate over the function of the princes of Orange in Dutch politics began to be conducted as a controversy over monarchy. A quasi-royal court rose up around Frederick Henry, and this in turn only clarified and strengthened the republicanism of his opponents, especially in Holland, who feared that the political leadership of the princes of Orange would be turned into an explicit monarchy.

During the 1640s, however, Frederick Henry lost his physical and intellectual powers and was unable to prevent Holland from reasserting its predominance over the republic's policies. The States-General entered into peace negotiations with Spain at Münster in Westphalia. Frederick Henry died in 1647 before the conclusion of the talks, but his son, William II, could not prevent the signing and ratification of the treaty in January 1648. Spain now formally acknowledged the independence of the Dutch, and indeed even urged its friendship upon the United Provinces, warning of the threat to both the Dutch and the Spanish from the rising power of France.

Peace of Westphalia

Prince William was not ready to accept a permanent peace, and he negotiated secretly with the French for a resumption of the war, not only against Spain but also against republican England, which had executed his father-in-law, King Charles I, in January 1649. Needing a powerful army to meet the anticipated war, William bitterly fought the efforts of Holland to reduce the standing army and thereby to permit more rapid payment of the huge debt accumulated over the eighty years' struggle for independence. Efforts at compromise broke down during the spring of 1650, as the Hollanders and William each sought to compel the other to concede political inferiority. William decided to make use of his preponderance in the States-General, and he led a delegation from that body to the towns of Holland to seek a change of their vote in the States of Holland; such a delegation was a direct violation of what Holland saw as its provincial sovereignty. Rebuffed by a number of town governments, most importantly by those of Amsterdam and Dordrecht, William decided to cut through the resistance by force. At The Hague, on July 30, 1650, he arrested six of the States' deputies from the recalcitrant towns and sent them to the castle of Loevestein (where Grotius had been imprisoned) on charges of having resisted lawful orders of the States-General. At that same time he sent an army to seize Amsterdam, but it was thwarted by delays on its march and by the determined resistance of the municipal authorities, supported by the common people. Amsterdam, however, faced a siege that might gravely imperil its trade, while the besiegers themselves ran the danger of being drowned should Amsterdam open the dikes. A compromise was soon worked out whereby William's opponents were released but were required to withdraw from government. William had cleared the way for his policies but at the price of arousing deep fears among the Dutch people—most of all in the powerful province of Holland—of military dictatorship, monarchical rule, and renewed involvement of the nation in war. But before he could carry out his plans, William II died of smallpox in early November. A posthumous son, William III, was born a week later.

Arrest of the deputies

The first stadholderless period. Fate thus intervened to give Holland's leaders, now intensely distrustful of Orangist influence, a chance to take over the country from the leaderless party of their antagonists. They governed the country for a little more than two decades, during what is known as the "first stadholderless period" (1650–72) because the five leading provinces did not appoint a successor to William II. (William II's cousin, William Frederick, of the junior branch of Orange-Nassau, continued to govern Friesland as well as Groningen, which also elected him as stadholder.) During the early months of 1651, a "Great Assembly" of the States-General, with expanded delegations from all the provinces, met at The Hague to consider the new situation. Holland was satisfied to consolidate the leadership it had so unexpectedly regained and conciliated the lesser provinces by leaving undisturbed the religious settlement of 1619 and by granting amnesty to those who had supported William II in 1650. But Holland's fears of the increased powers of the central government had been so stiffened that it depended upon its own preponderance, rather than upon constitutional reforms, to achieve effective government.

Yet efficiency of rule, so difficult to obtain when the powers to make and apply policy were so widely scattered, became all the more necessary when the republic became embroiled in war with the English Commonwealth in 1652. That conflict arose out of a medley of causes: first, the English republicans, after their successes against the royalists, took up the cause of defending English commercial interests against the Dutch and passed the Navigation Act of 1651 forbidding Dutch shippers from acting as middlemen in English trade both in Europe and overseas; second, the English sought to bring the Dutch into a political union directed primarily against the Stuarts and their Orange cousins. But the Dutch, whatever resentment the Hollanders bore against the Orange dynasty, were unwilling either to court civil war or to abandon their dearly won independence in a union that would make them junior partners to the English. An accidental clash between the Dutch and English fleets led to full-scale war in which a greatly improved English navy won the upper hand. By 1654 the Dutch were compelled to accept peace on English terms, including a secret promise by Holland ("Act of Seclusion") to exclude forever the prince of Orange from the stadholderate or the supreme command.

English navigation acts

The decision to accept a humiliating peace as the only way to terminate a disastrous war had been taken at the insistence of the young Johan de Witt, who had taken office in 1653 as councillor pensionary of Holland (the same office once held by Oldenbarnevelt). With the return of peace, de Witt became the brilliant leader of the republic's foreign and domestic policy. He rebuilt the Dutch navy, reduced indebtedness, improved the financial condition of both the States-General and the States of Holland, and restored the republic's prestige in Europe. Carefully averting any renewal of strife with England, he was able not only to compel France to back down in a naval dispute but also to send a powerful Dutch fleet to save Denmark from Swedish conquest in the first Northern War (1657–60).

When the exiled English king Charles II was restored to his throne in 1660, de Witt continued his policy of staying on good terms with England no matter who ruled there; this policy, however, foundered on the same two issues—commercial rivalry and the status of the House of Orange—that had brought about the war of 1652–54. Charles not only accepted the renewal of the Navigation Act of 1651 but intensified the rivalry with the Dutch by demanding forcefully that they acknowledge his sovereignty over the adjacent seas, pay tribute for the right to fish in the North Sea, and open the Dutch East Indies to English traders. When naval warfare resumed in 1664 off Africa, followed by war in Europe the next year, Charles took up the cause of the young Prince of Orange. By persuading the Orangists that his price for peace was restoration of William III to the offices of his forefathers, the English monarch built up a friendly party in the United Provinces that urged acceptance of his terms and even fostered a conspiracy to overthrow the government of de Witt and his friends. But de Witt managed to meet the new threat. An Orangist plot in Holland was uncovered and put down in 1666.

When Charles had demanded too high a price for Dutch friendship in 1660–62, de Witt had negotiated an alliance with the French, who feared that the restoration of the Prince of Orange would create a hostile Anglo-Dutch coalition. Furthermore, the fighting at sea increasingly went to the newly rebuilt Dutch navy. In 1667 the Dutch fleet sailed up the Thames and the Medway to Chatham, destroying the English shipyards and burning the fleet at its

moorings. In that same year, however, the French, under Louis XIV, who had only belatedly sent naval and land forces to aid the Dutch, began an invasion of the Spanish Netherlands in the War of Devolution. As French conquest of the southern Low Countries constituted a threat to both the Dutch Republic and Britain, the belligerents came to terms in the stand-off Peace of Breda (July 31, 1667), followed in January by an Anglo-Dutch alliance compelling France to make peace with Spain.

The Triple Alliance

This Triple Alliance (so-called because Sweden became a third partner) proved to be de Witt's undoing, although he had no effective diplomatic strategy to put in its place. Louis XIV, balked in his aim of conquest, considered that the Dutch had betrayed their alliance and turned to Charles II with proposals for a joint war against the United Provinces. Charles, bitterly resentful at his humiliating defeat, accepted the French offer of a richly subsidized alliance. Even as the threat from France emerged more clearly, the Orangists imagined that the Dutch could still win over Charles by the restoration of William III, but they were able to obtain only the Prince's appointment as commander in chief early in 1672. Charles joined the French in open war in the spring of 1672, counting upon William to accept rule of a rump Dutch Republic after France and Britain had taken away important territories for themselves. But William, who was given full power, including the stadholdership, during a storm of riots and near rebellion that swept the country in June and July after the French invasion penetrated to its heart, took over the leadership of the Dutch defense from de Witt, who was lynched by a mob in The Hague in August. With his support, the States-General rejected the Anglo-French terms.

William III. The tide of war now turned against the aggressors. The Dutch navy under Admiral de Ruyter repeatedly defeated the allied fleets off the coast of the republic, while the Dutch armies held on behind the flooded polders of the "water line." When other powers—Spain, at first as an auxiliary, then as a full participant, the Emperor, and Brandenburg—joined the Dutch side, the French armies withdrew from the republic. During six years of bitter war, William III was able to bring about the withdrawal of England (1674) and the defeat of all French war aims against the Dutch; yet his Grand Alliance was unable to bring Louis XIV to his knees, although Spain paid the price of a peace negotiated at Nijmegen in 1678. But during these years in which his political control of the republic, while strong, was not absolute, William was no more interested in constitutional reform than de Witt, his predecessor in the leadership of the country, had been. He was satisfied to expel adversaries from office and dominate the decisions taken by men who represented the same groups and the same social principles as those whom they replaced; but Holland, whose wealth ultimately was the basis for all Dutch power, political and military, slipped from under his thumb and asserted its autonomy of judgment and decision. The transformation of the republic, which had been from its origins an aristocracy dominated by mercantile wealth, into an oligarchy of inherited power, continued unimpeded by William: he had used the violence of the urban citizenry during the crisis of 1672 to unseat his opponents, but he was no more sympathetic than they had been to the vague democratic aspirations that were expressed here and there.

During the decade after the conclusion of the Peace of Nijmegen, the tension between William and Holland (particularly Amsterdam) worsened because the Prince was fixed upon a policy of renewed resistance to Louis XIV, while the Hollanders preferred peace at any reasonable price. But the upsurge of the threat from France in the late 1680s—the French incursions into western Germany and the threat of French domination of England under James II, a stalwart Catholic and a pensioner of Louis XIV—brought William and Holland into agreement upon the need to support the Prince's expedition to England in 1688, which resulted in his acceptance of the English throne, jointly with his wife Mary Stuart, early the next year. William, as king-stadholder, had to give primacy to English interests because England was the more powerful partner in the alliance. He therefore approved the arrangement whereby England concentrated her efforts against France on the sea, while the Dutch did so on land; the result was neglect of the Dutch navy. Ironically, the final triumph of the English over the Dutch in their commercial rivalry was a consequence of their alliance, not their enmity.

William accepts the English throne

The war begun in 1689 ended with a stalemate peace in 1697, followed by two treaties between the maritime powers and France for partition of the Spanish monarchy. In 1700, however, Louis XIV accepted the bequest of the Spanish throne for his grandson, Philippe d'Anjou (Philip V of Spain), and war was resumed the next year.

William died, childless, in 1702. When Holland again took the initiative for government without a stadholder, it was followed by the other provinces with much greater alacrity than had been true in 1650–51. Resentment had built up against the harshness with which the Prince had governed the country for three decades, and the absence of an adult heir meant that there was no effective opposition to the new course. Leadership of the Dutch state for the next 45 years came from the councillor pensionaries of Holland, who were often able men but either unwilling or unable to do more than conduct current business without attempting the delicate and explosive task of restructuring the government. On the contrary, constitutional rigidity became the credo not only of Dutch republicans but also of the Orangist party, with the only point in contention between them whether or not the Prince of Orange-Nassau, who was stadholder of Friesland, should be elected to the same office in the other provinces. William IV, who followed his father in Friesland in 1711, was chosen stadholder in Groningen in 1718 and in Gelderland (and the district of Drenthe) in 1722. Even without a stadholder in the principal provinces, Dutch subordination to English interests remained intact during the War of the Spanish Succession (1701–13) and the succeeding years of peace.

Dutch civilization in the Golden Age (1609–1713). The century from the conclusion of the Twelve Years' Truce in 1609 until either the death of Prince William III in 1702 or the conclusion of the Peace of Utrecht in 1713 is known in Dutch history as the "Golden Age." It was a unique era of political, economic, and cultural greatness during which the little nation on the North Sea ranked among the most powerful and influential in Europe and the world.

The economy. It was a grandeur that rested upon the economic expansion that continued with scarcely an interruption until 1648, at the end of the Thirty Years' War. The half century that followed was marked by consolidation rather than continued expansion, under the impact of the revived competition from the other nations, notably England and France, whose policies of mercantilism were in large degree directed against the near monopoly of the Dutch over the trade and shipping of Europe. Although the Dutch tenaciously resisted the new competition, the long-distance trading system of Europe was transformed from one largely conducted through the Netherlands, with the Dutch as universal buyer-seller and shipper, to one of multiple routes and fierce competitiveness. Nonetheless, the wealth earned during a long century of prosperity made the United Provinces a land of great riches, with more capital by far than could find outlet in investment. Yet the economic burden of repeated wars caused the Dutch to become one of the most heavily taxed people in Europe. Because the role of agriculture was secondary, taxes were to some extent imposed on the transit trade in and out of the country. But as mercantile competition became stiffer, the rate of such taxation could not be safely increased, and the burden therefore fell increasingly on the consumer. Excise and other indirect taxes made the Dutch cost of living one of the highest in Europe.

Increasing competition abroad

Dutch prosperity was built not only upon the "mother trades"—to the Baltic and to France and the Iberian lands—but also upon the overseas trades with Africa, Asia, and America. The attempt of the Spanish monarchs (who also ruled Portugal and its possessions from 1580 to 1640) to exclude Dutch merchants and shippers from the lucrative colonial commerce with the Far East led the Dutch to trade directly with the East Indies. Individual companies

The Dutch East India Company

were organized for each venture, but the companies were united by command of the States-General in 1602 in order to reduce the costs and increase the security of such perilous and complex undertakings; the resulting United East India Company established bases throughout the Indian Ocean, notably Ceylon, mainland India, and the Indonesian archipelago. The Dutch East India Company, like its rival English counterpart, was a trading company granted quasi-sovereign powers in the lands under its dominion. Although the East India fleets that returned annually with cargoes of spices and other valuables provided huge profits for the shareholders, the East India trade of the 17th and 18th centuries never provided more than a modest fraction of Dutch earnings from European trade. The West India Company, established in 1621, was built upon shakier economic foundations; trade in commodities was less important than the trade in slaves, in which the Dutch were pre-eminent in the 17th century, and privateering, which operated primarily out of Zeeland ports and preyed upon Spanish (and other) shipping. The West India Company had to be reorganized several times during its precarious existence, while the East India Company survived until the end of the 18th century.

Society. The social structure that evolved with the economic transformation of Dutch life was complex and marked by the predominance of the business classes that later centuries called the bourgeoisie, although with some significant differences. It was an unabashedly "aristocratic" country, socially and politically. The social "betters" of Dutch aristocracy were only to a limited extent landed noblemen, most of whom lived in the economically less advanced inland provinces. Most of the Dutch elite were wealthy townsmen whose fortunes were made as merchants and financiers, but they frequently shifted their activities to government, becoming what the Dutch called regents, members of the ruling bodies of town and province, and drawing most of their incomes from these posts and from investments in government bonds and real estate. The common people comprised both a numerous class of artisans and small businessmen whose prosperity provided the base for the generally high Dutch standard of living and a very large class of sailors, shipbuilders, fishermen, and other workers. Dutch workers were in general well paid, but they were also burdened by unusually high taxes. The farmers, producing chiefly cash crops, prospered in a country that needed large amounts of food and raw materials for its urban (and seagoing) population. The quality of life was marked by less disparity between classes than prevailed elsewhere, although the difference between a great merchant's home on the Herengracht in Amsterdam and a dockworker's hovel was too obvious to be remarked. What was striking was the comparative simplicity even of the wealthy classes and the sense of status and dignity among the ordinary people, although the exuberance that had earlier marked the society was toned down or even eliminated by the strict Calvinist morality preached and to some extent enforced by the official church. There was, too, a good deal of mingling between the burgher regents who possessed great wealth and political power and the landed gentry and lesser nobility who formed the traditional elite.

The regents

Religion. One of the characteristic aspects of modern Dutch society began to evolve in this period—the vertical separation of society into "pillars" (*zuilen*) identified with the different Dutch religions. Calvinist Protestantism became the officially recognized religion of the country, politically favoured and economically supported by government. But the Reformed preachers were thwarted in their efforts to oppress or drive out other religions, to which a far-reaching toleration was extended. Mass conversion to Calvinism had been confined mainly to the earlier decades of the Eighty Years' War, when Catholics still frequently bore the burden of their preference for the rule of the Catholic monarchs in the southern Netherlands. Sizable islands of Catholicism remained in most of the United Provinces except Zeeland, while Gelderland and the northern parts of Brabant and Flanders conquered by the States-General were overwhelmingly Catholic, as they remain today. Although public practice of Catholicism was forbidden, interference with private worship was rare, even if Catholics sometimes bought their security with bribes to local Protestant authorities. Catholics lost the traditional form of church government by bishops, whose place was taken by a papal vicar directly dependent upon Rome and supervising what was in effect a mission; the political authorities were generally tolerant of secular priests but not of Jesuits, who were vigorous proselytizers and linked to Spanish interests. Protestants included, along with the predominant Calvinists of the Reformed Church, both Lutherans in small numbers and Mennonites (Anabaptists), who were politically passive but often prospered in business. In addition, the Remonstrants, who were driven out of the Reformed Church after the Synod of Dort (Dordrecht; 1618–19), continued as a small sect with considerable influence among the regents. There were also other sects emphasizing mystical experiences or rationalist theologies, notably the Collegiants among the latter. Jews settled in the Netherlands to escape persecution; the Sefardic Jews from Spain and Portugal were more influential in economic, social, and intellectual life, while the Ashkenazim from eastern Europe formed a stratum of impoverished workers, especially in Amsterdam. Despite unusually open contacts with the Christian society about them, Dutch Jews continued to live in their own communities under their own laws and rabbinic leadership. Successful though some Jews were in business, they were by no means the central force in the rise and expansion of Dutch capitalism. Indeed, no clear pattern can be detected of religious affiliation affecting the growth of the Dutch business community; if anything, it was the official Dutch Reformed Church that fulminated most angrily against capitalist attitudes and practices, while the merely tolerated faiths often saw their adherents, to whom economic but not political careers were open, prospering and even amassing fortunes.

Religious "pillars"

The Jews

Culture. The economic prosperity of the Dutch Republic in this "golden century" was matched by an extraordinary flowering of cultural achievement, which drew from the nation's prosperity not only the direct resources of financial nourishment but also a driving and sustaining sense of purpose and vigour. This was reflected in the first instance by a notable series of historical works: from the contemporary chronicles of the revolt by Pieter Bor and Emanuel van Meteren; through the highly polished account by Pieter Corneliszoon Hooft, a masterpiece of narration and judgment in the spirit of Tacitus; to the heavily factual chronicle of Lieuwe van Aitzema, with its interspersed commentary of skeptical wisdom, Abraham de Wicquefort's history of the republic (principally under the first stadholderless administration), and the histories and biographies by Geeraert Brandt. These were works in which a proud new nation took account of its birthpangs and its growth to greatness. Only in the latter part of the century did Dutch historians begin to express a sense that political grandeur may be transient.

Political theorists shared the same concerns, although the effort to fit new experience and ideas into the traditional categories derived from Aristotle and Roman law created an air of unreality about their work, perhaps even more than was true of political thinkers elsewhere in Europe. Theorists such as the Gouda official Vrancken in the days of the foundation of the republic and Hugo Grotius in the early 17th century portrayed the republic as essentially unchanged since the early Middle Ages or even since antiquity—a country where sovereignty resided in provincial and town assemblies, which had partly lost their control to counts and kings before regaining it in the revolt against Philip II. The next surge of political debate came after mid-century, when for a little more than two decades the country was governed without a prince of Orange as stadholder.

Political theory

The controversy over whether the young prince William had any right by birth to the offices of his forefathers probed the fundamental character of the republic, for even a quasi-hereditary stadholdership created an incipient monarchy within the traditional structure of aristocratic republicanism. The debate involved the issue not so much of centralization versus provincialism as where the leadership of the republic properly lay, whether in

the House of Orange or in the province of Holland and notably its greatest city, Amsterdam. Only the celebrated philosopher Benedict de Spinoza, an outsider by origin and character, elevated these political questions to the level of universality.

Two of the greatest philosophers of the 17th century resided in the Dutch Republic, but both were to some extent outsiders. The Frenchman René Descartes found there a freedom from intellectual inquisitions and personal involvements and lived in the Netherlands for two decades while engaged in the studies that helped transform modern thought. The second was Spinoza, a Jew by birth and upbringing but a key figure in the history of all modern Western philosophy. He was more than just Descartes's greatest exponent, for he was an independent thinker of the first order. Scientific activity in the United Provinces reached a high level. The physicist Christiaan Huygens approached Newton himself in power of mind and importance of scientific contribution. The engineer and mathematician Simon Stevin and the microscopists Antonie van Leeuwenhoek and Jan Swammerdam rank in the front of their fields.

Dutch literature, which knew great creativity during the "golden age," remained the possession of the relatively small number of those who spoke and read Dutch. Figures like the historian P.C. Hooft or the poets Constantijn Huygens and Joost van den Vondel, who was also a distinguished playwright, wrote with a power and a purity worthy of the best that France and England produced at that time. Music was hampered by the Calvinists' antipathy to what they saw as frivolity. Organ music was barred from services in Reformed churches, although town authorities frequently continued its performance at other times. The great organist-composer J.P. Sweelinck was more influential in encouraging the creative wave in Germany than among his own countrymen. The art whose achievements rank at the very top was painting, which rested upon the broad patronage of a prosperous population. Group portraits of regents and other influential citizens adorned town halls and charitable establishments, while still lifes and anecdotal paintings of popular life hung in profusion in private homes. Some of the greatest work, from the brushes of such painters as Hals, Steen, and Vermeer, were painted for these markets, but the greatest of Dutch painters, Rembrandt van Rijn, broke through the boundaries of the group portrait to create works with his own extraordinary mood and inward meaning. The landscape painters, notably Jacob van Ruisdael, captured the distinctive Dutch flat land, broad skies with massed clouds, and muted light. Architecture remained at a lesser level, merging with some success the native traditions of brick buildings and gable roofs and fashionable Renaissance styles. Sculpture remained a largely foreign art.

Painting

The 18th century. *Economic and political stagnation.* Once the Dutch fleet had declined, Dutch mercantile interests became heavily dependent upon English goodwill; yet the rulers of the country were more concerned to reduce the monumental debt that weighed heavily upon the country. During the 18th century Dutch trade and shipping were able to maintain the level of activity reached at the end of the 17th century, but they did not match the dramatic expansion of French and especially English competitors. The Dutch near monopoly was now only a memory. Holland remained rich in accumulated capital, although much of it could find no outlet for investment in business. Some went into the purchase of landed estates, but a great deal was used to buy bonds of foreign governments; the bankers of Amsterdam were among the most important in Europe, rivalling those of London and Geneva. Dutch culture failed to hold its eminence; a medical scientist like Hermann Boerhaave or a jurist like Cornelis van Bynkershoek was highly respected, but they were not the shapers and shakers of European thought. Dutch artists were no longer of the first order, and literature largely followed English or French models without matching their achievements. The quality of life changed: instead of the seething activity of the 17th century, the 18th century was one of calm and easeful pleasantness, at least for men of property. The middling classes in town and countryside also knew continuing prosperity; conditions for the labouring classes continued to be hard, although foreign visitors thought the workers lived better than elsewhere. The poor as such were not labourers but a residual class of unemployed who subsisted on the charity of town governments and private foundations. Religious life was more relaxed, particularly among Protestants. Catholics, still without political rights but facing milder restrictions, fell into a quarrel between adherents of Jansenism, which denied free will, and supporters of Rome; the former split off to form the Old Catholic Church, a small denomination that still exists. The educated classes widely accepted the principles and attitudes of the Enlightenment although without the sharp hostility to religion that characterized the French Philosophes.

During the second stadholderless period of Dutch government (1702–47), the republican system became an immobile oligarchy. The "liberty" defended by the regents as soundly republican was in practice the rule of hereditary patricians, responsible to neither the citizenry below nor a stadholder above. Although William IV yearned for restoration to the offices held by the princes of Orange before him in the provinces to the south, he accepted, with no less admiration and commitment than the regents, the perfection and immutability of the Dutch constitutional system, with the single difference that he envisioned it including the stadholderate for all the provinces. It was not until the War of the Austrian Succession (1740–48) that the power of the regents began to crumble. As in 1672, disaster on the battlefield proved the Achilles' heel of a regime that had not built up a broad popular political base. The regents had not been able to overcome the traditional commitment of the people to the House of Orange as their natural leader and saviour. French and Prussian armies swarmed over the Austrian (southern) Netherlands and were poised for invasion of the United Provinces, which were linked by alliance with England, although they had remained formally neutral. When the French forces crossed into Dutch territory, rioting reminiscent of 1672, although less widespread and violent, led to the fall of the second purely republican government and the election of William IV as hereditary stadholder of all the provinces. Otherwise there was little change; some regents were compelled to step down from their posts, and leadership in the hands of the Prince of Orange was uncontested. William rebuffed the efforts of burghers in Amsterdam and other towns who had supported his restoration in order to achieve democratic reforms, in which participation in government would be extended to men of modest property who had been completely disfranchised (although not to wageworkers or to paupers).

War of the Austrian Succession

The Patriotic movement. During the next decades, in the face of the rigid conservatism of the princes of Orange (William V succeeded his father in 1751 and assumed personal government in 1759) and under the influence of the French Enlightenment, an essentially new political force began to take shape. Known as the Patriot movement after an old party term used by both republicans and Orangists, it applied fundamental criticism to the established government. Although the Patriot movement was representative of the new democratic and Enlightenment ideals, it had strong roots in native Dutch traditions. From the beginning, the United Provinces had rejected specifically democratic institutions in favour of frankly aristocratic government (in the Aristotelian sense), but the notion that the regents had a duty to serve not their own private interests but those of the country and the people had persisted in theory and in mood. When the aristocracy ceased to recruit new members from below and thus became an enclosed caste, the discrepancy between its claim of service to the general welfare and the reality of its practice became evident. The Patriot movement took in a wide range of supporters: discontented noblemen like the Gelderland baron Joan van der Capellen tot den Pol; wealthy bankers and businessmen without a voice in government; artisans and shopkeepers, traditionally Orangist in sympathy, who were dismayed to find their claims to an effective role in the politics of their towns rebuffed by the princes; and intellectuals committed to the new

Traditionalism of the Patriots

Enlightenment rejection of arbitrary power. The Patriots included in their ranks many Protestant dissenters and Catholics, but the Jews continued to look to the prince of Orange as their protector. Some regents, holding firm to the republicanism of their ancestors and resenting the return of the stadholderate, found a new base for their ideas in the Patriot movement. Most regents, however, saw more peril in the new movement for broader popular government than in the stolid conservatism of the princes of Orange; a reconciliation between the camps of the patrician republicans and the Orangists began to take shape under the impact of a common threat from below.

Again the events of war imperilled the established regime. Although the diplomacy of William V was firmly based upon the alliance with England, London became exasperated with the Dutch during the United States War of Independence (1775–83), when they attempted to continue and expand their profitable trade with the new American nation as well as with France. Dutch flirtations with the Russian-sponsored League of Armed Neutrality, resisting British searches of neutral vessels, and indications of Dutch negotiations for an alliance with the Americans only worsened relations. Finally open hostilities erupted in the fourth Anglo-Dutch War (1780–84). The Dutch navy, sorely neglected for more than a half century, was utterly unprepared to battle the powerful British fleet, and the Dutch fleet's attempts to convoy their merchantmen brought only disaster.

The onus of defeat fell upon the Stadholder. He was unable to stand firm against the increased agitation of the Patriots, who forced their way into governments of town after town in Holland and other provinces. Holland began organizing its own army, distinct from that under the Prince's command, and civil war seemed in the offing. William V fled to Gelderland with his wife, Wilhelmina, the sister of the Prussian king Frederick II. Holland declared him deposed.

It was the strong-willed Wilhelmina, rather than her hesitant and rather docile husband, who took the lead in the restoration of the stadholderate. Dutch politics had now become a concern of the great powers. France sided with the Patriots, not out of sympathy with their principles but because they opposed the Stadholder, who had fallen back into dependence upon English and Prussian support. As long as Frederick II ruled in Prussia, Wilhelmina's pleas for armed intervention fell on deaf ears, but when the throne passed to his nephew Frederick William II in 1786, the way opened for action. The Patriots counted upon the support of the French, but the government at Versailles, then entering the final financial and political crisis of the monarchy that erupted in the Revolution of 1789, could give no more than verbal encouragement. Wilhelmina, working closely with the English ambassador, arranged to create a crisis by seeking to return to Holland; her detention at the provincial border was taken by Prussia as justification to send an army into the United Provinces. The Prussians quickly swept away the makeshift militias of Holland and Utrecht and restored the Stadholder to his offices. A period of repression of Patriots followed; many went into exile, first in the Austrian Netherlands and then in France. The outbreak of the French Revolution in 1789 gave new hope to the exiles and their friends at home. They looked now for more effective French assistance and at the same time found in the French Revolutionary experience practical ideas for the reorganization of the government at home, notably the principle of a single, indivisible republic. The Patriots' hopes rose when French Revolutionary armies swept over the Austrian Netherlands (which had had a brief interlude of independence in 1789–90) in 1792, but the French forces retreated the next year. It was not until 1794 that they returned to Belgium (as it now became customary to call the southern Netherlands), driving up to and then across the frontier of the United Provinces. The moment for which the Dutch Patriots had long been waiting was at hand: French power would more than overweigh the English and Prussian strength upon which the stadholder relied (Prussia made a separate peace with France in 1795), and a democratic revolution, thwarted in 1787, would be possible. The freezing of the great rivers during the winter permitted the French forces to cross into the Dutch heartland, but even before they arrived the Patriots seized the reins of state from helpless William V, who abandoned office and fled to England.

Impact of the French Revolution

The period of French rule, 1795–1813. The old republic was replaced by the Batavian Republic, and the political modernization of the Netherlands began—a process that would take more than half a century and pass through many vicissitudes; yet it was one marked by an extraordinary lack of violence. For all its flaws and inconsistencies, the old regime of the United Provinces had enjoyed many of the institutions and practices that other countries had to create in the fire of revolution: the sovereignty of parliamentary assemblies; wide-ranging political and religious toleration; equality of all citizens before the law; and an unusually broad distribution of the benefits of economic prosperity, however far the social system was from equality. Even the sense of nationhood had put down deep roots, although the awareness of differences of religion remained powerful. In a word, the Dutch had already achieved a large measure of the "liberty, equality and fraternity" that had become the slogan of the French Revolutionaries. The task that confronted the Batavian and the successor regimes was to adapt old institutions and create new ones that could meet the needs of a new era. But the Dutch statesmen had to operate within the confines of a small power shorn of most of its military and naval strength and yet more dependent than most other countries upon its trading and shipping.

The Batavian Republic, 1795–1806. The Batavian Republic lasted 11 years, during which it proclaimed the sovereignty of the people but was in fact a protectorate of France. The organization of government had to be approved not only by the Dutch people but also by whatever government happened to be in control in France. The constitutions therefore reflected not only Dutch conditions and ideas but also the arrangements in effect in Paris; nonetheless, they did create a political system of the new type, a new regime, in the Netherlands. After much debate, the ancient historic provinces—so unequal in wealth, population, and influence—were replaced by a unitary republic divided into departments and electoral constituencies that were roughly equal in population, if not in wealth. The representatives elected to the National Assembly (which replaced the historic States-General) were not delegates of provincial assemblies by whose decisions they were bound but deputies with full independence of judgment. The ancient system of government, with its medley of assemblies and boards with imperfectly differentiated functions, was replaced by a modern system of separate and explicitly defined legislative, executive, and judicial branches; functionally organized ministries directed the work of foreign affairs, internal affairs, war, and navy. The full legal equality of all citizens in all parts of the country was proclaimed; the residents of North Brabant, Zeeland-Flanders, Limburg, and Drenthe gained the same rights as all other citizens of the republic, just as their districts, once excluded from the States-General, now participated in the national government equally with all others. The Reformed Church lost its standing as the sole official, protected church, supported out of state revenues; equal status was accorded to all religious denominations, including Catholics and Jews. Yet full separation of church and state was not proclaimed, and their relationship was to continue as one of the central factors in Dutch politics for more than a century. The historic privileges of class and locality were abolished; the liberty of each and all under the law and before the courts replaced the diverse "liberties" of town and province, noble and regent. Where, before, town governments had co-opted their members, deputies to the National Assembly were now elected; but the franchise was limited to property holders, and these did not choose their representatives directly but through electors named by primary assemblies. Most of these institutional changes were permanent, though the republican form of government was replaced by a kingdom in 1806 and never reestablished.

Reorganization of the provinces

Abolition of ancient privilege

While these momentous changes were being debated and adopted, the ordinary work of state and nation had to con-

tinue amid conditions of almost unprecedented difficulty. England reacted to the French occupation of the Netherlands and the flight and overthrow of the Stadholder by a declaration of war and a blockade. Dutch overseas trade and fishing, the country's most essential occupations, were brought to a near standstill, while most of the Dutch colonies were seized by the English on behalf of William V. The French, however, remained relentless in their own exploitation of the occupied "fraternal republic." The Dutch government, which took over the whole accumulated burden of national and provincial indebtedness, had also to bear the costs of the French occupying forces and to pay immense sums in tribute to the Paris government; indeed, the forced circulation of vastly inflated French assignats (paper currency) at face value was a scarcely disguised and very effective form of French taxation directly upon the Dutch people. Nor did the successive French governments—republican, consular, or imperial—grant the Dutch any greater freedom of trade with France or other countries under its control in compensation for the lost overseas business.

Shift toward agriculture

As trade declined and industry languished, the Netherlands began to change in economic character toward a primarily agricultural country. The venturesome spirit for which Dutch businessmen had been so famed a century or two before seemed to be lost, replaced by what the Dutch themselves called a "jansalie" (stick-in-the-mud) attitude; once-bustling cities dwindled to mere market towns; even Amsterdam lost much of its population. As a result, it became difficult to consolidate the new government. A multiple executive modelled on the French Directory and lacking a firm base in established political institutions and practices reflected the intrigues of individuals rather than the programs of clearly delineated parties. The victors quarrelled among themselves and looked to Paris to decide among them, or at least passively accepted its dictum, given by coups d'etat organized or approved by the French army command. In 1805, Napoleon gave quasi-dictatorial powers to R.J. Schimmelpenninck. Schimmelpenninck, called Councillor Pensionary after the fashion of the old provincial leaders, was actually an uncrowned and nearly absolute monarch (although, ultimately, power continued in Napoleon's hands); he nonetheless carried into practice many of the reforms that had been proposed but not adopted. But Napoleon decided the next year to incorporate the Dutch state directly into his "Grand Empire" of vassal states.

The Kingdom of Holland and the French Empire, 1806–13. Renamed the Kingdom of Holland, the Netherlands received as its monarch Napoleon's younger brother Louis. The four years of his kingship were one of the strangest episodes in Dutch history. Louis Bonaparte was a stranger in the land, yet he took its interests to heart, evading his brother's commands and winning the respect, if not quite the affection, of his subjects. The reconciliation of former Orangists, republicans, and Patriots began under his rule, for, in the face of the apparent permanence of the Napoleonic empire, they entered his government and worked together. Nonetheless, the brute fact remained that, for Napoleon, Holland was the kingpin of the "continental system," which he hoped would bring England to its knees by cutting off its continental exports. French officials enforced the vigorous suppression of the smuggling of English and colonial goods to the Continent through Holland that had sprung up over the past decade with London's connivance. King Louis's resistance to his brother's efforts and his refusal to put French interests ahead of Dutch led to the Emperor's decision to oust his brother from his throne in 1810 and to incorporate Holland into the French Empire.

Kingship of Louis Bonaparte

Little changed, however; the same officials—some Dutch, some French—continued to do the work of government in the country, which remained outside the French tariff system. As long as the Napoleonic empire seemed firmly based and permanent, Dutchmen served the new sovereign as they had King Louis, all the more readily because the exiled Prince of Orange gave permission for such collaboration. Dutch soldiers continued to fight in Napoleon's campaigns, losing heavily in the Russian invasion of 1812. But as it became increasingly obvious, after the failure of the Russian and Spanish campaigns, that the Napoleonic empire was collapsing, influential Dutchmen began to prepare for the creation of a new and independent regime; it was taken for granted that its head would be the Prince of Orange—the son of William V, who had died in 1806—and that it was desirable that it be established by the Dutch people rather than imposed by the eventual allied victors. The movement for restoration was led by a remarkable figure, Gijsbert Karel van Hogendorp, a man of firm political principle who had refused to serve any of the governments that ruled in Holland after 1795, yet accepted the necessity for a re-established prince of Orange to govern the country as a limited constitutional sovereign.

Hogendorp's restoration

During the autumn of 1813, Hogendorp secretly planned a takeover of government from the French, which became possible without bloodshed during November as French troops withdrew to their homeland. On November 30, the hereditary stadholder, at the invitation of Hogendorp's provisional authority, returned from England to proclaim his reign as hereditary prince. In 1814, he granted a charter establishing a constitutional monarchy, with restricted powers for a Parliament elected by a narrow property suffrage. At the insistence of the victorious powers meeting at the Congress of Vienna, he took the title of king of The Netherlands and was given sovereignty over the southern Netherlands, which included both Belgium and Luxembourg. During the campaign against Napoleon after his return from Elba in 1815, Dutch troops played a major role in his defeat at Waterloo.

The Kingdom of The Netherlands (1814–1918). *King William I.* The reign of King William I, as the restored Prince of Orange was now called, was one of the most critical periods in the history of The Netherlands. During this quarter-century the adaptation of the country to the conditions and requirements of modernity moved in a complex and even contradictory way, guided by a monarch who in his economic policy was far more forward-looking than most of his countrymen but who in politics resisted the expansion of Parliament and the introduction of liberal principles. He was a 19th-century version of the "enlightened despot," a man intent upon power, although not so much for its own sake as in order to serve the welfare of his country as he saw it. The role of the States-General—which continued to represent a general electorate of taxpaying citizenry—was strictly limited to the enactment of laws proposed by the government and to approval of a long-term budget; it was in no sense the representative of a sovereign people. The ministers of state were the agents of the king and responsible to him, not to the States-General. Yet the basic structure of modern government had been created in The Netherlands; constitutional debate would be concerned with redistributing powers and responsibilities among existing institutions.

William I was at his best in confronting the problem of reviving the economic life of the country after the shattering impact of the long French occupation. He put the support of both the government and his own private fortune behind encouragement of commerce and, to a lesser extent, of industry. He sponsored the formation of The Netherlands Trading Society, a nominally private firm that undertook the important but costly and risky enterprise of reorganizing Dutch long-distance trade and shipping, particularly to the Netherlands East Indies, which were returned to Dutch sovereignty by England as part of the peace settlement. With the reopening of trade between the European continent and the wider world, the advantages of the Dutch position at the mouth of the great rivers favoured the revival of the traditional branches of Dutch enterprise; but competition from the ports of other countries, notably from Hamburg and Bremen, as well as from Britain, remained strong. Only in the Netherlands East Indies did the Dutch have a clear advantage over their rivals, despite the abandonment of the monopoly formerly maintained under the East India Company.

Encouragement of trade

The most difficult problem faced by the new regime in The Netherlands was the relations between Holland (which now became the everyday name for all the northern Netherlands, in Dutch as well as foreign usage)

and Belgium. The King was passionately devoted to the preservation of a single state encompassing all the Low Countries, a unity lost in the revolt against Spain more than two centuries before; but the sense of common nationhood, cultural and political, was quite weak among the people. The Belgians resented assuming a share of the burden of debt inherited by Holland; they were oriented toward industry, Hollanders toward trade. French was the language of the leading classes in the south, and the use of Dutch as the official language was bitterly opposed even by Flemings, who resented the Dutch version of the common Dutch-Flemish language. Most Flemings, as devout Catholics, were hostile to the predominantly Protestant Hollanders. William's efforts to assume the control that Napoleon had possessed over the Belgian Catholic Church met fierce resistance. At the same time the authoritarian character of William's government, particularly the sharp censorship of the press in Belgium, aroused the antipathy of liberals to the regime. The result was the outbreak of the Belgian Revolution of 1830 and the proclamation of Belgian independence. William, supported by a majority of Dutchmen, who were angered by what they saw as Belgian ingratitude, was able to defeat the hastily organized Belgian army; but the European powers intervened to secure Belgian independence, although it was not until 1839 that a final settlement was reached and the last Dutch troops withdrew from Belgian soil. William, deeply despondent, abdicated the next year, leaving to his son, King William II, the task of coming to terms with the new situation.

The Belgian Revolution of 1830

William II and William III. The new king was not a man of clear ideas or strong will, but he was able to do what his father dared not even envisage—to oversee the transformation of The Netherlands into a parliamentary, liberal state. When the crisis of the 1848 revolutions broke, first in France and then in central Europe, William II turned to the leading liberal thinker, J.R. Thorbecke, to guide the change. A new constitution was written, largely modelled on the British (and Belgian) pattern, which gave effective supremacy to the States-General and made the monarch a servant and not the master of government. The King died the next year, and the work of transformation continued under his son, William III (1849–90), who named Thorbecke prime minister. The constitutional monarchy was consolidated, even though Thorbecke stepped down in 1853 because of Protestant rioting against the re-establishment of a Roman Catholic archbishop at Utrecht. Gradually, over the next century, the scope of Dutch democracy was extended to include ever broader sections of the Dutch population in the franchise; universal male suffrage was achieved during World War I, and suffrage was extended to women after World War II. During this period, political parties of the modern type took shape, organized along religious and ideological lines; to the original Liberal, Protestant, and Catholic parties were eventually added Socialist, Conservative-Protestant, Communist, and minor parties. As no single party was able to emerge with a majority, coalition politics became inevitable. The central issue of political controversy became the "school conflict" (*schoolstrijd*), which pitted the liberal (and later socialist) advocates of public schools against the combined Protestant and Catholic parties, which demanded that the state support private ("special") schools equally with the public schools. For several decades, the liberals remained generally in control and made few concessions on the school issue. But when the Protestant leader Abraham Kuyper formed a coalition with the Catholics in 1888, the religious parties were able to gain power and to favour the special over the public schools. Their policy was assailed by the secular parties, the traditional liberals, the progressives, and the socialists. The liberals, however, were at odds with the other anti-clerical parties on other issues, notably economic policies and the extension of the suffrage. The liberals tended to be the most conservative party on economic issues and favoured a restricted electorate; the progressives were vigorously democratic in outlook, as were the socialists, who also favoured universal suffrage, protection of the right to strike, labour legislation, and other welfare measures. The other major issue of the latter half of the 19th century was the role of the Dutch

Extension of the franchise

The Dutch in the Indies

East Indies. The income received by the Dutch treasury from Indonesian taxes helped balance the national budget; yet the revelations of harsh conditions in the distant archipelago made it impossible to maintain the "culture system," which had been introduced to force the production of certain crops for export, while the long Achin war drained the treasury.

Queen Wilhelmina and World War I. During the first half of the reign of Queen Wilhelmina (1890–1948), the political situation remained fundamentally unchanged. The major parties came to recognize that the school struggle interfered with the solution of other problems. An agreement in principle was reached on the eve of World War I, by which the secular parties accepted state support for religious schools on a basis of equal funds in exchange for enactment of universal male suffrage. When war broke out in 1914, Holland, which had declared its neutrality, put aside the proposed reforms in order to concentrate on the immediate problem of maintaining the country's livelihood in the face of blockades. The "Pacification," as the compromise was called, was adopted in 1917 and put into effect after the return of peace. The war years saw almost all political controversies set aside, while the government took unprecedented action in maintaining trade and guiding economic life. Although spared the horrors of combat, the Dutch had to maintain a large standing army, and mutinies broke out among the soldiers in 1918.

The century from the restoration of Dutch independence in 1813 until World War I saw fundamental transformations of Dutch life. The economic base was modernized; the role of agriculture fell off, with most Dutch farmers producing dairy and meat products for the market; trade and shipping were revived in the face of fiercely competitive conditions. But most important was the rise of industry—first textiles in the eastern provinces, then coal in the southeast in Limburg, and finally modern manufactures, notably the great Philips electrical products factories at Eindhoven. Rotterdam became one of the world's busiest ports and the centre of chemical and other industries. These changes were paralleled in society by the gradual extinction of pauperism, the domination of the middle-class business and professional men, and the improvement of the conditions of working people and farmers, especially after the mid-19th century.

Economic transformation

Although religious freedom was generally as great as anywhere in Europe, Protestant conservatives faced major difficulties, especially during the first half of the century, when they protested the modernizing ideas of the Reformed (Hervormde) Church; their efforts to create independent religious communities met with sharp resistance from the government, which did not interfere in non-Protestant religions. Some of the Gereformeerden (the older name for "Reformed" used by the conservatives) emigrated, many of them to the United States; but, by the latter half of the century, persecution ceased.

The cultural life of Holland remained very largely confined within national boundaries; Dutch thinkers, writers, and artists responded strongly to influences from Germany, France, and England but themselves had little impact abroad. Dutch scientists maintained a respected position for their country; Hugo de Vries was one of the principal founders of the science of genetics, while physicist Hendrik Antoon Lorentz contributed greatly to Einstein's theories of relativity. Dutch artists were generally imitative, although the "Hague school" of Impressionists displayed great gifts; only Vincent van Gogh, who spent most of his active life in France, achieved world reputation. Dutch literature ran parallel to main currents abroad; the Réveil early in the century was a movement of intensely religious romanticism with strongly conservative ideas, while Eduard Douwes Dekker (pseudonym Multatuli) in mid-century expressed the moods of social criticism with great power; the movement of "Men of the 'Eighties" (Tachtigers) brought to the fore a stress upon aesthetic values and spirituality; and early in the 20th century a literature of social protest re-emerged.

19th-century literature

The Netherlands since 1918. The movement of The Netherlands into modernity was accelerated after 1918. Although the country became a member of the League

of Nations, it reaffirmed its neutrality, which seemed to have obtained the respect of the powers and which was symbolized by the presence of the International Court of Justice at The Hague. There was considerable harshness in relations with Belgium, which not only abandoned its neutrality for a close alliance with France but demanded territorial cessions from Holland. The Dutch government, although humiliated by a demand that it present its case before the peace conference at Versailles, successfully resisted any amputation of its territory. The Dutch, for their part, refrained from giving any official support to the Flemish nationalist movement in Belgium, although a "Great Netherlands" movement, principally among intellectuals, emphasized the underlying unity of the Dutch and Flemings. Domestic politics continued in the same course, with the Protestant political parties continuing to provide leadership for generally conservative policies, especially after the onset of the world depression in the 1930s.

World War II. At the outbreak of World War II in 1939, the Dutch sedulously maintained their neutrality, although their sympathies lay overwhelmingly with the Western powers. Nonetheless, when Nazi Germany undertook the campaign against France in the spring of 1940, its forces struck not only against Belgium in order to outflank the French defenses but also against Holland. The Dutch land armies were overwhelmed in less than a week, and the government, accompanied by Queen Wilhelmina and the royal family, withdrew to England, where they formed a government in exile.

The work of public administration under German occupation was continued by Dutch organs of state, which attempted to buffer German political repression, deportation of Jews, and forced employment of Dutch labour in Germany. A resistance movement sprang up, which, with the exception of a few Dutch Nazi collaborators, spanned all groups from the conservatives to the Communists. The Germans retaliated by executing Dutch hostages for such measures of resistance as the strike of Amsterdam dockworkers against the seizure and deportation of Dutch Jews to extermination camps in Germany. Some Jews were able to "dive under" (go into hiding) with the assistance of friends, but the large majority were taken away to their deaths. In the final phases of the war, particularly after the failure of Allied airborne attempts to capture bridgeheads across the great rivers at Nijmegen and Arnhem, the Dutch suffered from severe food shortages, and during the last months before liberation (May 1945) they were on the verge of famine.

The resistance movement

The postwar period. After the war many aspects of Dutch life changed dramatically. Wilhelmina and her government returned from exile to reestablish a government more strongly democratic than ever, with both universal suffrage and proportional representation in elections. Anticipating the characteristic difficulties of postwar reconstruction, the government, industry, and labour agreed upon a plan for industrial and commercial expansion, with avoidance of the rapid expansion of prices or wages that would bring a threat of inflation. The plan worked effectively for more than two decades, so that the Dutch were able to avoid drastic inflation until the introduction of a new taxation system in the late 1960s.

Dutch industrialization moved forward with speed and depth, expanding to include the large-scale production of steel, electronics, and petrochemicals. Holland, putting aside the policy of neutrality as a failure, entered vigorously into the postwar Western alliances, including the North Atlantic Treaty Organization (NATO), and the various organizations of European unity (the Common Market); but its influence was limited, even though it joined with Belgium and Luxembourg in a closer union ("Benelux"). Indonesia, where Dutch authority was reestablished after wartime Japanese occupation, soon became the scene of a nationalist revolution. After some hesitation and bitterness, the Dutch granted it full independence. The Netherlands Antilles remained part of the Dutch kingdom, although no longer under the authority of the government at The Hague; Surinam became independent in 1975 and was renamed the Republic of Suriname in 1978.

Abandonment of neutrality

(H.Ro.)

Dutch political alignments since the mid-20th century have undergone few shifts. The first postwar governments were dominated by an alliance of the Labour and Catholic parties, which continued until the Labour Party went into opposition in 1958. Thereafter, with the exception of 1973–77, when the country had a left-led government, and 1981–82, when it was ruled by a centre-left coalition that collapsed prematurely in a dispute over a job creation plan, governments were formed by centre-right coalitions. After the early 1980s the government was faced not only with recurrent economic problems but also with the emotion-charged issue of siting U.S. nuclear cruise missiles (as part of the NATO defense strategy) in the country. It finally reached the decision in 1985, against widespread popular opposition, that 48 missiles would be sited by 1988.

During the 1960s, the generally peaceful mood of Dutch public life was broken by rioting of youth and labour groups, especially in Amsterdam. The most difficult crisis affected the royal family. The marriage (1966) of Princess Beatrix, the heiress to Queen Juliana (who had succeeded Wilhelmina on her abdication in 1948), to a German diplomat aroused acrimonious debate. The marriage of Princess Irene to a Spanish Carlist prince had already come as a shock even to the Catholics, but it was less difficult politically because she lost her right of succession. Juliana's husband and consort, Prince Bernhard, was involved in a bribery scandal and withdrew from his military offices. Juliana abdicated in 1980 and was succeeded as queen by Beatrix.

Dutch politics, like Dutch society in general, continued to practice what was called pillarization (*verzuiling*)—the organization not only of political parties but of labour unions, businessmen's organizations, social and sport clubs, and many professional groups on the basis of membership in a religious-ideological "pillar," either Catholic, Protestant, or Humanist (the latter including Liberals and Socialists). Pillarization had received official confirmation in the Pacification of 1917 and removed most of the tinder from Dutch politics; but it also kept ordinary Dutchmen separated from each other, by religion, to a greater degree than in most other Western countries. Yet, because the leaders of the pillar organizations worked well with each other and the right of each pillar to exist and function was unquestioned, public life generally ran smoothly. A quarter-century after the end of World War II, there were signs of disquiet over the system, however, and efforts were being made to form new political parties, such as the Christian Democratic Appeal, on a comprehensive basis.

Continued pillarization

For later developments in the political history of The Netherlands, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Nine, Division II, especially Section 923; and Division VI, especially Sections 961 and 963.

(H.Ro./Ed.)

Belgium

One of the smallest and most densely populated nations of Europe, the Kingdom of Belgium has been, since its birth in 1830 and 1831, a hereditary, representative, and constitutional monarchy. Its area is 11,783 square miles (30,519 square kilometres). The country is bounded on the northwest by the North Sea, on the north and northeast by The Netherlands, on the east by West Germany, on the southeast by the Grand Duchy of Luxembourg, and on the west and southwest by France.

Culturally, Belgium is a heterogeneous nation, a bringing together without union of the ancient Latin and Germanic heritages of western Europe. Aside from a small German-speaking population in the east, the nation is divided geographically between the Walloons, a French-speaking people in the four southwestern provinces (Hainaut, Namur, Liège, and Luxembourg), and the Flemings, a Dutch-speaking people in the four northern and northeastern provinces (West-Vlaanderen, Oost-Vlaanderen [West and East Flanders], Antwerp, and Limburg). The central province, Brabant, is itself split, and in it lies the officially bilingual capital, Brussels.

Cultural heterogeneity

Internal disharmonies aside, Belgium and the political entities that preceded it have been rich with historical and cultural associations, from the Gothic grandeur of its medieval university cities and its small castle-dominated towns on steep-bluffed winding rivers, through its broad traditions in painting and music that marked one of the high points of the northern Renaissance in the 16th century, to its contributions to the arts of the 20th century and its maintenance of the folk cultures of past eras. The Belgian landscape has been a major battleground of Europe for centuries, notably in modern times during the Battle of Waterloo and World Wars I and II. Given its area and population, Belgium today is the most heavily industrialized nation in Europe. It is a member of the Benelux customs union formed in 1944 with The Netherlands and Luxembourg and a member of the European Economic Community (EEC).

PHYSICAL AND HUMAN GEOGRAPHY

Topographical character

The land. *Relief.* Belgium generally is a low-lying country, with a broad coastal plain extending from the North Sea and The Netherlands and rising gradually into the Ardennes hills and forests of the southeast, where a maximum height of 2,277 feet (694 metres) is reached at Botrange.

The main physical regions are the Ardennes and Ardennes foothills; the Anglo-Belgian Basin to the north comprising the Central (Bas) Plateaus, the plain of Vlaanderen (Flanders) and the Kempenland (Campine); and the intrusion of the Paris Basin on the west and southwest known as Belgian Lorraine.

The Ardennes region is part of the Hercynian Belt, reaching from western Ireland into Germany and formed during the late Paleozoic Era well over 225,000,000 years ago. It is a plateau cut deeply by the Meuse River and its tributaries. Its higher points have poor drainage and are more favourable for peat bogs and upland mossy ground than for crops.

A large depression known east of the Meuse as the Famenne, and as the Fagne west of it, separates the Ardennes from the geologically and topographically complex foothills to the north. The principal part of the area is the Condroz, a plateau more than 980 feet in altitude comprising a succession of valleys hollowed out of the limestone between sandstone crests. Its northern boundary is the Sambre-Meuse Valley, which crosses Belgium from east to west.

Situated south of the Ardennes and cut off from the rest of the country, the Belgian Lorraine is a series of hills with north-facing scarps. Half of it remains wooded, and in the south is a small region of iron-ore deposits.

A region of sand and clay soils lying between 160 and 650 feet in altitude, the Central Plateaus cover northern Hainaut, southern Brabant, and the Hesbaye plateau region of Liège. It is dissected by several rivers that enter the Schelde (Escaut) River; it is bounded on the north by a series of outlying hills. The Brussels region lies within the Central Plateaus.

Bordering the North Sea from France to the Schelde, the low-lying plain of Vlaanderen has two main sections. Maritime Vlaanderen, extending inland for five to 10 miles (eight to 16 kilometres), is a region of reclaimed land, or polders, protected by a line of dunes and dikes and having largely clay soils. Interior Vlaanderen comprises most of Oost- and West-Vlaanderen and has sand-silt or sand soils. At 70 to 160 feet in altitude, it is drained by the Leie, Schelde, and Dendre rivers flowing northeastward to the Schelde estuary. Several shipping canals interlace the landscape connecting the river systems.

Covered by pasturelands and industry and lying between 160 and 330 feet in altitude, the Kempenland forms an irregular watershed between the Schelde and Meuse systems.

Climate. Belgium has a temperate, maritime climate predominantly influenced by air masses from the Atlantic. Rapid and frequent alternation of different air masses separated by fronts gives Belgium considerable variability in weather. Frontal conditions moving from the west produce rainy weather, with rainfall heavy and frequent, averaging around 30 to 40 inches (750 to 1,000 millimetres). Winters are damp and mild with frequent fogs; summers, rather cool. The annual mean temperature is around 50° F (10° C). Brussels, for example, roughly in the middle of the country, has a minimum of 36° F (2.2° C) in December and a maximum of more than 62° F (16.6° C) in July.

Regional climatic variations

Regional climatic differences are determined by altitude and distance inland. Farther inland, maritime influences become weaker, and the climate becomes more continental, characterized by greater seasonal extremes of temperature. The Ardennes region, the highest and farthest inland, is the coldest. In January, mean temperatures are lower, frost occurs on about 100 days, and snow falls on 30 to 35 days. In summer, the altitude counteracts the effect of distance inland, and July temperatures are the lowest in the country. Because of the topography, the region has the highest rainfall in Belgium. In contrast, the Vlaanderen region enjoys generally higher temperatures throughout the year. There are fewer than 60 days of frost, and fewer than 15 of snow. On the seacoast these figures are reduced to below 50 and 10, respectively. There are a few very hot days, especially on the coast, and the annual rainfall is the lowest in the country.

Plant and animal life. Belgium lies within the area of deciduous forestation. The dominant tree is the oak, and other trees include beech, birch, and elm. For about a century, however, coniferous plantings have modified the character of the forests. The two most wooded regions are the Kempenland, where planted coniferous forests predominate, and the Ardennes, with deciduous and coniferous forests. The animal population, greatly reduced by man's activities, is Eurasian. Wild boar, wildcats, and deer roam the forests of the Ardennes.

Settlement patterns. The biological resources of the several natural regions and the consequent variations in land use have been major factors in determining patterns of rural settlement. The nature of the urban developments imposed on this landscape is derived mainly from the patterns of mining, manufacturing, commerce, and related enterprise throughout the country.

Man's uses of the land

Forest and grassland dominate the landscape south of the Sambre-Meuse Valley: meadows, with orchards, monopolize the land in the Fagne and in the Plateau de Herve, while forest occupies two-thirds of the area along both edges of the Ardennes and in the heart of Lorraine. Forage crops, oats, potatoes, and even wheat are grown everywhere but especially in Lorraine and in the Condroz. The region is one of striking contrasts: in the Condroz, farms often exceed 125 acres (50 hectares), whereas in the Ardennes half are less than 25 acres. The population is sparse in the Plateau de Herve in the east and western Entre-Sambre-et-Meuse in the southwest, but it is more concentrated throughout the remainder of the region.

The open countryside of Hainaut, Brabant, and Hesbaye is given over to intensive diversified agriculture devoted to pasture, wheat, sugar beets, and oats; local variations include orchards in northern Hesbaye. Settlement is thick, and farms, with their closed courts, sometimes exceed 250 acres in area.

The open landscape of maritime Vlaanderen and the lower Schelde, intersected by dikes and canals, is dotted with scattered farms, their buildings loosely grouped around a court. Over one-third of the area of farms between 50 and 125 acres is under pasture, while the remainder is cultivated, with wheat and sugar beet the dominant crops. Inland Vlaanderen is a region of scattered habitation and large market towns, its widespread scrub undergrowth giving it a wooded appearance. It is devoted to grazing. Grasslands are sparse. Intensive cultivation is confined to gardens and small farms, one-third of them under 10 acres. Oats, rye, and potatoes are the chief crops, and wheat and such varied industrial crops as sugar beets, chicory, hops, and flax are found in the southwest.

Coniferous forests and grasses dominate the Kempenland countryside, and oats, rye, and potatoes are also important. Farms are small, and most of the population lives in villages. Coal seams lie beneath portions of the region.

In the Walloon coalfields, roughly in and to the north of the Meuse Valley across south central Belgium, coal mining, glass manufacture, iron production, zinc metal-

The major cities of Belgium

lurgy, and the chemical and electrical industries in the 19th and 20th centuries gave rise to heavily populated areas with widely varying urban characteristics. Liège, the regional capital since the Middle Ages, dominates and coordinates the chief metropolitan area of the Walloon region. Charleroi, the heart of a large metropolitan area, is a newer and smaller city and is predominantly commercial. La Louvière, founded during the 19th-century industrial development, is a developing metropolitan centre. The Borinage, an area of high population density without a central city, comes under the influence of Mons.

The ancient city of Antwerp and its metropolitan area, the second largest in the nation, extend along the east bank of the Schelde. The port is formed by the base of the estuary and the concave riverbank. The existence of the port has favoured the establishment of important and diverse industries: petroleum refining, chemical and metallurgical industries, food processing, and electronics manufacturing. Gent (Ghent, or Gand), a historic university town, is Belgium's second largest port. A centre of the textile industry, Gent is experiencing an industrial regeneration, characterized especially by heavy steel production along the Gent-Terneuzen Canal, connecting the port to the Schelde.

Brugge (in French Bruges, "bridge"), the capital of West Vlaanderen, is a city of medieval aspect, resplendent with cathedrals and ancient homes. As its name implies, the city has many bridges spanning the several canals and the canalized Rei River. Mentioned as early as the 7th century, it became a trading centre for the Hanseatic League and reached its greatest period during the 15th century, when the dukes of Burgundy held court there.

Leuven (in French, Louvain), 16 miles east of Brussels, is the site of the first university founded (1425) in the Low

MAP INDEX

Political subdivisions

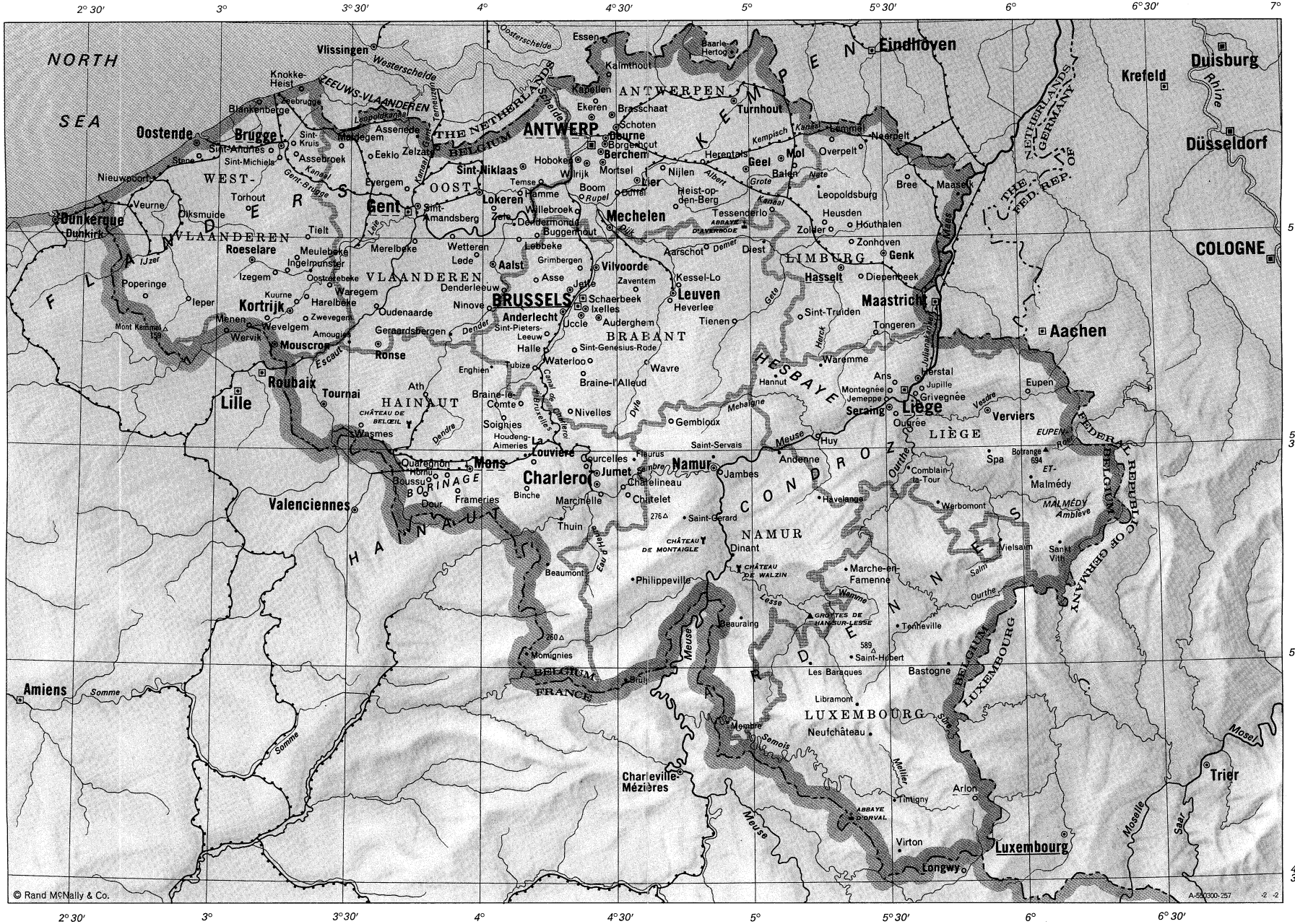
Antwerpen.....51·10n 4·50e
Brabant.....50·45n 4·30e
East Flanders, see Oost-Vlaanderen
Hainaut.....50·30n 3·50e
Liège.....50·30n 5·30e
Limburg.....51·00n 5·30e
Luxembourg.....50·00n 5·30e
Namur.....50·20n 4·50e
Oost-Vlaanderen..51·00n 3·45e
West Flanders, see West-Vlaanderen
West-Vlaanderen..51·00n 2·43e

Cities and towns

Aalst.....50·56n 4·02e
Aarschot.....50·59n 4·50e
Amougies.....50·45n 3·29e
Andenne.....50·29n 5·06e
Anderlecht.....50·50n 4·18e
Ans.....50·39n 5·32e
Antwerp.....51·13n 4·25e
Anvers, see Antwerp
Arlon.....49·41n 5·49e
Asse.....50·55n 4·12e
Assebroek.....51·12n 3·16e
Assenede.....51·14n 3·45e
Ath.....50·38n 3·47e
Auderghem.....50·49n 4·26e
Baarle-Hertog.....51·27n 4·56e
Balen.....51·10n 5·09e
Bastogne.....50·00n 5·43e
Beaumont.....50·14n 4·14e
Beauraing.....50·07n 4·58e
Berchem.....51·12n 4·26e
Bergen, see Mons
Binche.....50·24n 4·10e
Blankenberge.....51·19n 3·08e
Boom.....51·05n 4·22e
Borgerhout.....51·13n 4·26e
Boussu.....50·26n 3·48e
Braine-l'Alleud....50·41n 4·22e
Braine-le-Comte..50·36n 4·08e
Brasschaat.....51·17n 4·27e
Bree.....51·08n 5·36e
Bruges, see Brugge
Brugge.....51·13n 3·14e
Brûly.....49·58n 4·31e
Brussels.....50·50n 4·20e
Bruxelles, see Brussels
Buggenhout.....51·01n 4·12e
Charleroi.....50·25n 4·26e
Châtelet.....50·24n 4·31e
Châtelineau.....50·25n 4·31e
Comblain-la-Tour.....50·27n 5·34e
Courcelles.....50·28n 4·22e
Courtrai, see Kortrijk
Denderleeuw.....50·53n 4·04e
Dendermonde....51·02n 4·07e
Deurne.....51·13n 4·28e
Diepenbeek.....50·54n 5·24e
Diest.....50·59n 5·03e
Diksmuide.....51·02n 2·52e
Dinant.....50·16n 4·55e
Dixmude, see Diksmuide
Dour.....50·24n 3·47e
Duffel.....51·06n 4·31e
Eeklo.....51·11n 3·34e
Ekeren.....51·17n 4·25e
Enghien.....50·42n 4·02e
Essen.....51·28n 4·28e
Eupen.....50·38n 6·02e
Evergem.....51·07n 3·42e
Fleurus.....50·29n 4·33e
Frameries.....50·24n 3·54e
Gand, see Gent
Geel.....51·10n 5·00e
Gembloux.....50·34n 4·41e
Genk.....50·58n 5·30e
Gent.....51·03n 3·43e
Geraardsbergen..50·46n 3·52e
Ghent, see Gent
Grimbergen.....50·56n 4·23e
Grivegnée.....50·37n 5·36e
Halle.....50·44n 4·13e
Hamme.....51·06n 4·08e
Hannut.....50·40n 5·05e
Harelbeke.....50·51n 3·18e
Hasselt.....50·56n 5·20e
Havelange.....50·23n 5·14e
Heist-op-den-Berg.....51·05n 4·43e
Herentals.....51·11n 4·50e
Herstal.....50·40n 5·38e
Heusden.....51·02n 5·16e
Heverlee.....50·52n 4·42e
Hoboken.....51·10n 4·21e
Hornu.....50·26n 3·49e
Houdeng-Aimeries.....50·29n 4·08e
Houthalen.....51·02n 5·22e
Huy.....50·31n 5·14e
Ieper.....50·51n 2·53e
Ingelmunster.....50·55n 3·15e
Ixelles.....50·50n 4·22e
Izegem.....50·55n 3·12e
Jambes.....50·28n 4·52e
Jemeppe.....50·37n 5·30e
Jette.....50·52n 4·20e
Jumet.....50·26n 4·25e
Jupille.....50·39n 5·38e
Kalmthout.....51·23n 4·28e
Kapellen.....51·19n 4·26e
Kessel-Lo.....50·54n 4·45e
Knokke-Heist....51·21n 3·17e
Kortrijk.....50·50n 3·16e
Kuurne.....50·51n 3·17e
La Louvière.....50·28n 4·11e
Lebbeke.....51·00n 4·08e
Lede.....50·58n 3·58e
Leopoldsburg.....51·07n 5·15e
Les Baraques.....50·01n 5·13e
Leuven (Louvain)..50·53n 4·42e
Libramont.....49·55n 5·23e
Liège.....50·38n 5·34e
Lier.....51·08n 4·34e
Lokeren.....51·06n 4·00e
Lommel.....51·14n 5·18e
Louvain, see Leuven
Luik, see Liège
Maaseik.....51·06n 5·48e
Maldegem.....51·13n 3·27e
Malines, see Mechelen
Malmédy.....50·25n 6·02e
Marche-en-Famenne.....50·12n 5·20e
Mechelen.....51·02n 4·28e
Marcinelle.....50·24n 4·26e
Membre.....49·52n 4·54e
Menen.....50·48n 3·07e
Merelbeke.....51·00n 3·45e
Meulebeke.....50·57n 3·17e
Mol.....51·11n 5·06e
Momignies.....50·02n 4·10e
Mons.....50·27n 3·56e
Montegnée.....50·39n 5·31e
Mortsel.....51·10n 4·28e
Mousoron.....50·44n 3·13e
Namur (Namen)..50·28n 4·52e
Neerpelt.....51·13n 5·25e
Neufchâteau.....49·50n 5·26e
Nieuwpoort.....51·08n 2·45e
Nijlen.....51·10n 4·39e
Ninove.....50·50n 4·01e
Nivelles.....50·36n 4·20e
Oostende.....51·13n 2·55e
Oostrozebeke.....50·55n 3·20e
Ostend, see Oostende
Oudenaarde.....50·51n 3·36e
Ougrée.....50·36n 5·32e
Overpelt.....51·13n 5·25e
Philippeville.....50·12n 4·32e
Poperinge.....50·51n 2·43e
Quaregnon.....50·26n 3·51e
Roeselare.....50·57n 3·08e
Ronse.....50·45n 3·36e
Saint, see also under Sint, Sankt
Saint-Gérard.....50·21n 4·45e
Saint-Hubert.....50·01n 5·23e
Saint-Servais.....50·28n 4·50e
Sankt Vith.....50·17n 6·08e
Schaerbeek.....50·51n 4·23e
Schoten.....51·15n 4·30e
Seraing.....50·36n 5·29e
Sint-Amandsberg.51·04n 3·45e
Sint-Andries.....51·12n 3·10e
Sint-Genesius-Rode.....50·45n 4·21e
Sint-Kruis.....51·13n 3·15e
Sint-Michiels.....51·11n 3·12e
Sint-Niklaas.....51·10n 4·08e
Sint-Pieters-Leeuw.....50·47n 4·14e
Sint-Truiden.....50·48n 5·12e
Soignies.....50·35n 4·04e
Spa.....50·30n 5·52e
Stene.....51·12n 2·55e
Temse.....51·08n 4·13e
Tenneville.....50·06n 5·32e
Tessenderlo.....51·04n 5·05e
Thuin.....50·20n 4·17e
Tielt.....51·00n 3·19e
Tienen.....50·48n 4·57e
Tintigny.....49·41n 5·31e
Tongeren.....50·47n 5·28e
Tournai.....50·36n 3·23e
Torhout.....51·04n 3·06e
Tubize.....50·41n 4·12e
Turnhout.....51·19n 4·57e
Uccle.....50·48n 4·19e
Verviers.....50·35n 5·52e
Veurne.....51·04n 2·40e
Vielsalm.....50·17n 5·55e
Vilvoorde.....50·56n 4·26e
Virton.....49·34n 5·32e
Waregem.....50·53n 3·25e
Waremme.....50·41n 5·15e
Wasmes.....50·33n 3·32e
Waterloo.....50·43n 4·23e
Wavre.....50·43n 4·37e
Werbomont.....50·23n 5·41e
Wervik.....50·47n 3·02e
Wetteren.....51·00n 3·53e
Wevelgem.....50·48n 3·10e
Willebroek.....51·04n 4·22e
Wilrijk.....51·10n 4·24e
Ypres, see Ieper
Zaventem.....50·53n 4·28e
Zeebrugge.....51·20n 3·12e
Zele.....51·04n 4·02e
Zelzate.....51·12n 3·49e
Zinnik, see Soignies
Zolder.....51·01n 5·18e
Zonhoven.....50·59n 5·21e
Zwevegem.....50·48n 3·20e

Physical features and points of interest

Albert Kanaal, *canal*.....50·39n 5·37e
Amblève, *river*....50·28n 5·36e
Averbode, Abbaye d', *abbey*.51·02n 4·59e
Ardennes, *physical region*...50·10n 5·45e
Belœil, Château de, *palace*.....50·35n 3·44e
Borinage, *historic region*....50·25n 3·50e
Botrange, *mountain*.....50·30n 6·08e
Charleroi à Bruxelles, Canal de.....50·51n 4·19e
Condroz, *historic region*....50·20n 5·00e
Demer, *river*.....50·58n 4·42e
Dender (Dendre), *river*.....51·02n 4·06e
Dijk, *river*.....51·04n 4·25e
Dyle, *river*.....51·04n 4·25e
Eau d'Heure, *river*.....50·18n 4·24e
Escaut (Schelde), *river*.....51·20n 4·15e
Eupen-et-Malmédy, *historic region*....50·30n 6·05e
Flanders, *historic region*....51·00n 3·00e
Gent-Brugge, Kanaal, *canal*....51·03n 3·43e
Gent-Terneuzen, Kanaal, *canal*.....51·04n 3·44e
Gete, *river*.....50·57n 5·07e
Grote Nete, *river*..51·07n 4·34e
Hainaut, *historic region*....50·30n 3·50e
Han-sur-Lesse, Grottes de, *cave*..50·08n 5·15e
Herck, *river*.....50·58n 5·07e
Hesbaye, *historic region*....50·35n 5·15e
IJzer, *river*.....51·09n 2·43e
Julianakanaal, *canal*.....51·05n 5·50e
Kemmel, Mont, *hill*.....50·47n 2·50e
Kempen, *historic region*....51·10n 5·20e
Kempisch Kanaal, *canal*....51·10n 4·49e
Leie, *river*.....51·03n 3·43e
Leopoldkanaal, *canal*.....51·14n 3·46e
Lesse, *river*.....50·14n 4·54e
Maas (Meuse), *river*.....51·10n 5·52e
Mehaigne, *river*...50·32n 5·13e
Mellier, *river*.....49·43n 5·32e
Meuse (Maas), *river*.....51·10n 5·52e
Montaigle, Château de, *castle*.....50·18n 4·49e
North Sea.....51·20n 2·30e
Orval, Abbaye d', *abbey*.....49·38n 5·22e
Ourthe, *river*.....50·38n 5·35e
Roer, *river*.....50·32n 6·13e
Rupel, *river*.....51·07n 4·19e
Sambre, *river*.....50·28n 4·52e
Salm, *river*.....50·22n 5·52e
Schelde (Escaut), *river*.....51·20n 4·15e
Semois, *river*.....49·52n 4·52e
Sûre, *river*.....49·50n 5·45e
Vesdre, *river*.....50·37n 5·37e
Walzin, Château de, *castle*.....50·13n 4·55e
Wamme, *river*.....50·10n 5·16e



BELGIUM

Size of symbol indicates relative size of town • ○ ◉ ▣ ■

Elevations in metres

0 10 20 30 40 50 km

0 10 20 30 40 mi

Countries. The institution was damaged severely during both world wars, but it was rebuilt and its libraries restocked by aid from many nations.

Brussels, the capital, has suburbs that spread over more than half of Brabant. It is the centre of commerce, industry, and intellectual life in Belgium.

The people. *Ethnic groups and religion.* The population of Belgium is divided into three linguistic communities. The Dutch-speaking Flemings in the north constitute well over half the population and are increasing relative to the French-speaking Walloons in the south and west. The German-language region in eastern Liège Province, containing fewer than 1 percent of the Belgians, consists of more than 20 communes around Eupen and St. Vith. The Brussels area comprises 19 bilingual, one Walloon, and six Flemish communes. The French-speaking population is by far the larger, however, in the capital region.

Consequences of linguistic diversity

Historically, Belgium's linguistic division has created a social and economic fabric in which the managerial, professional, and administrative ranks were filled almost entirely by the French-speaking segment of the population, which at one time was also numerically superior. The Flemings long protested against what they felt was the exclusion of the average, non-bilingual Fleming from effective participation even in everyday dealings concerning law, medicine, and industrial employment. Along with their gradually increasing numerical strength, the Flemings have made gains in many areas, and since 1898 Belgium has been officially a bilingual nation. Many disputes and much rancour remain, however, and the future direction of the controversy is uncertain.

The majority of Belgians are Roman Catholic, but regular attendance at religious services is variable. Although it is marked in the Flemish region, regular attendance at church has decreased in the Walloon industrial region and in Brussels. The Protestants live mostly in Hainaut, particularly in the Borinage, and in Brabant. The Jewish population is concentrated in the general areas of Brussels and Antwerp.

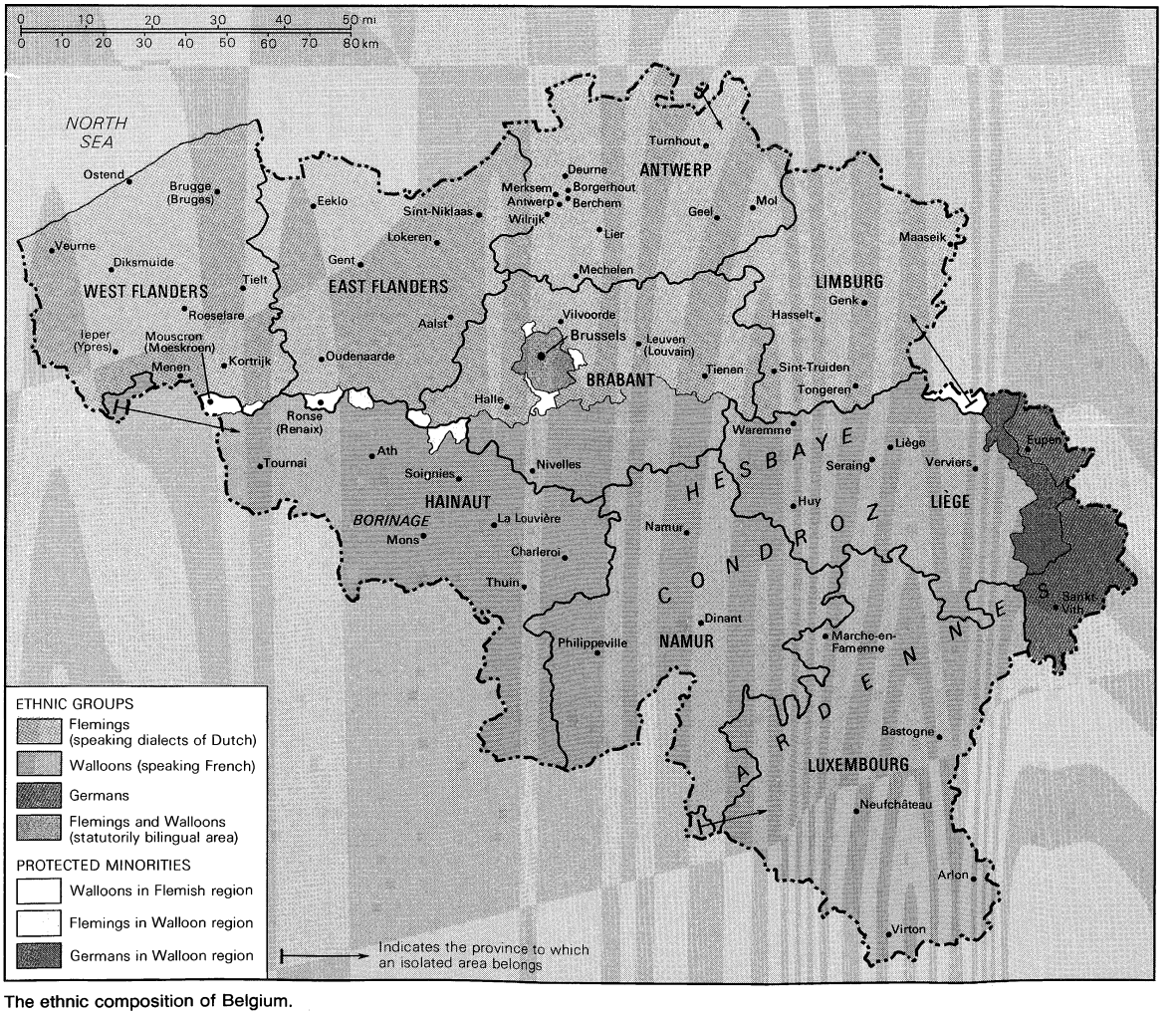
Linguistic regions and demographic phenomena

Demographic trends. The annual growth rate of the Belgian population is very low, for overall birth rates and immigration exceed death rates and emigration only slightly. Marked differences exist between the Dutch-language region and the remainder of the country. Its natural increase—*i.e.*, preponderance of births over deaths—and net immigration give Flemish Belgium a high growth rate. In contrast, the French- and German-language regions together have a negative rate of natural increase that would reduce the population progressively were it not for a very small net immigration that raises these regions' total rate of population increase.

Since World War II, the foreign-born population has increased continuously at a rate higher than that of Belgian nationals: in addition to its continuing inflow, it has also had a high birth rate. The largest concentrations of foreigners are found in the Walloon mining and industrial areas and in Brussels. The foreign workers are largely of Mediterranean (mostly Italian) and North African origin.

General picture of the economy

The economy. A small percentage of Belgium's active population is engaged in agriculture, suggesting the great role of industry and commerce in the national economy. National prosperity is dependent mainly on Belgium's role as a fabricator and processor of imported raw materials and on the subsequent export of finished goods. The country became a major steel producer in the early 19th century, with factories centred in the southern Walloon coal-mining region. After World War II drastic monetary reform aided postwar recovery and expansion, particularly of the Flemish light manufacturing and chemical industries that developed rapidly in the north, and Belgium



The ethnic composition of Belgium.

was one of the first European countries to reestablish a favourable balance of trade. By the late 20th century, however, coal reserves were becoming exhausted, the aging steel industry had become inefficient, labour costs had risen dramatically, and foreign investment (about half of the country's industrial assets are controlled by multinational companies) had declined. The government, in an effort to reverse the near-depression levels of industrial output, subsidized ailing industries, particularly steel and textiles, and offered tax incentives, reduced interest rates, and capital bonuses to attract foreign investment.

A geographic analysis of the gross domestic product (GDP) by province reveals that the highest levels of GDP per capita are in the Brabant and Antwerp areas. The analysis of the GDP according to economic activities confirms the importance of services and the weakness of agriculture. The provinces of Luxembourg, Namur, and West-Vlaanderen have the most highly developed agriculture, whereas industry is the major economic activity of Hainaut, Limburg, Liège, and Oost-Vlaanderen. The Antwerp region dominates in transportation.

Mining. Coal long has been Belgium's only important mineral resource. The iron-ore deposits of the Sambre-Meuse belt, in decline since the latter part of the 19th century, have now been worked out, and only small deposits of nonferrous ores exist.

The Borinage coal-mining area in the Sambre-Meuse Valley lies in a narrow band across south central Belgium from the French border through Mons, Charleroi, Namur, and Liège. Because the easily extractable coal reserves have been exhausted, most of the mines in the region have been closed. Although many mines in the Kempenland field in northern Belgium also are worked out or abandoned as uneconomic, this second region provides almost all of the domestic coking and slowburning coal for industry. Difficult mining conditions add to costs, and the yield per miner is among the lowest of any EEC country. Although the industry operates as a private enterprise, it is a considerable drain on public finances, for the government subsidizes the industry and assists new industries in areas hit by mine closings.

Agriculture, fishing, and forestry. The principal crops are cereal grains, potatoes, and sugar beets, with livestock the major agricultural export. Horticulture is being developed slowly. Ostend, the main fishing port, sends a modest fleet of trawlers to the North Sea fishing grounds. Reforestation is extensive, and the lumbering industry, aided by mechanization, has increased its output.

The steel, chemical, and metallurgical industries

Industry. The manufacturing sector, despite a sagging economy in the late 20th century, accounts for about one-third of national income. The steel, metallurgical, textile, and chemical sectors are the major components of the Belgian industry. Belgium is one of the world's leading processors of cobalt, germanium, and radium. Refineries, located principally in the Antwerp area, process crude petroleum. The lace for which Belgium has been known around the world long was part of a slowly dying industry that depended heavily on the handwork of elderly women. The authorities and individuals in Brugge, the centre of the industry, gave new impetus to the tradition, however, establishing specialized schools for training younger workers and arranging expositions in various countries.

Finance. The economic importance of the financial sector has increased significantly in recent years. The General Savings and Pensions Fund is the major financial organization for the collection of savings, while the Société Générale de Belgique dominates the banking sector. The National Bank is charged with issuing currency and has considerable autonomy.

Trade. The main imports are raw materials (including petroleum, the nation's main source of energy) and food products. The major exports are chemicals, machinery, motor vehicles, food and livestock, and iron and steel. The principal trade partners of Belgium are the member nations of the EEC.

Management of the economy. Free enterprise, which dominates the entire economy with the exception of the air and rail transportation sectors and broadcasting, faces scarcely any government competition. Government activity is aimed primarily at maintaining overall economic equilibrium. Ordinary expenditures are covered by tax revenues, while such extraordinary expenditures as public works projects are financed through loans. During the late 20th century government tax revenues declined following the economic downturn, and the government was forced to borrow heavily from abroad to finance foreign trade, to sustain its generous social welfare system (which was burdened more than normally by the large number of unemployed) and to assist its subsidization of the steel and other industries. The level of public debt is high, and the interest on it is a significant item in the annual budget.

Workers are represented mainly by three trade-union organizations: the General Federation of Belgian Workers, the Confederation of Christian Trade Unions, and the Belgium General Federation of Liberal Trade Unions. The first two enjoy a virtual monopoly of trade-union membership. The Federation of Belgian Industries and the Federation of Non-Industrial Enterprises of Belgium represent employers.

Transportation. The extensive system of main roads in Belgium is supplemented by modern roadways that extend from Brussels to Ostend by way of Gent and Brugge, from Brussels to Antwerp, and from Antwerp to Aachen, West Germany, by way of Hasselt and Liège. Other expressways include those from Antwerp to Kortrijk by way of Gent, from Brussels to the Ardennes by way of Namur, and from Brussels to Paris through Mons and Charleroi.

The railway network, a state enterprise, is one of the densest in the world. Brussels is the heart of the system, the centre of a series of lines that radiate outward and link the capital to other cities both inside and outside the country. The heaviest traffic is between Brussels and Antwerp.

Antwerp handles about 80 percent of the country's foreign trade through its port. Other important ports are Brugge-Zeebrugge, Ostend, Gent, and Brussels. Navigable inland waterways include the Meuse and Schelde, which are navigable throughout their length in Belgium. The Albert Canal, which can take barges up to 2,000 tons, links Antwerp with the Liège region. A canal from Charleroi to Brussels, carrying barges up to 1,350 tons, links the basins of the two main rivers through Ronquieres lock. A maritime canal connects Brugge and Zeebrugge, another connects Gent and Terneuzen, on the Schelde estuary, and a third links Brussels and Rupel.

The Brussels international airport is the centre of Belgian air traffic, while smaller international facilities are maintained at Antwerp, Liège, Charleroi, and Ostend. SABENA, the national airline, was created in 1923; although 90 percent government owned, it is incorporated as a private corporation. It serves most parts of the world and has domestic flights.

Administrative and social conditions. *Government.* Belgium is a constitutional monarchy. Under the terms of the Belgian constitution, ministers are appointed and dismissed by the monarch. Since the monarch's person is inviolable, royal acts must be countersigned by a minister, who in turn becomes responsible for them to the parliament. Legislative power is shared by the monarch with the Chamber of Representatives and the Senate; the members of both these bodies are elected for four-year terms. Except for a few points that were clarified by the constitutional revision of 1970, the constitution is silent on the subject of governmental structure. The structure thus has been created by custom and has varied through the course of time.

Ministerial and legislative bodies

Apart from the secretaries of state, who were introduced in the revision of 1970 and who are not part of the Council of Ministers, there has never been a constitutionally established hierarchy among the ministers. The position of prime minister was created in 1919 and that of vice prime minister in 1961. Before linguistic equality was formally established, an unwritten rule prevailed that, except for the prime minister, the government must include as many Dutch as French-speaking ministers. The provinces, divided into administrative districts (*arrondissements*) and further subdivided into *communes*, are each under the authority of a governor, with legislative power exercised by the provincial council. The Députation Permanente

(Permanent Deputation), elected from the members of the provincial council, provides for daily provincial administration. Each *commune* is headed by a burgomaster, and the communal council elects the deputy mayors (*échevins*).

Political life. Communal elections take place every six years, while provincial and national elections are held every four years. Deputies and communal representatives are elected directly, as are certain senators, while other senators are either designated by the provincial councils or are selected by the elected senators and the provincial senators.

The traditional parties are, in order of importance, the Christian Social Party (PSC), the Belgian Socialist Party (PSB; with both French- and Dutch-speaking wings), the Party of Freedom and Progress (PLP), and the Communist Party (PC). In addition, a few new parties have emerged, notably the People's Union (VU) in Flanders, the Walloon Assembly (RW) in the French-speaking area, and the French Democratic Front (FDF) in Brussels.

Participation in decision making is exercised on various levels by the representatives, party members, and the electorate, in the latter case through the compulsory vote. Citizens are informed of political events through the press, but, with press ownership increasingly concentrated in fewer hands, many persons consider the medium to be unamenable to the expression of a wide range of opinion. Radio and television often organize debates and discussions that provide political information. In spite of these efforts, a marked disaffection exists among the citizens with regard to politics, and most national governments are formed by party coalitions.

Justice. Judges are appointed for life by the monarch; they cannot be removed except by judicial sentence. At the cantonal, or lowest, judicial level, justices of the peace decide civil and commercial cases, and police tribunals decide criminal cases. At the district level, judicial powers are divided among the tribunals of first instance, which are subdivided into civil, criminal, and juvenile courts and commercial and labour tribunals. At the appeals level, the courts of appeal include civil, criminal, and juvenile divisions that are supplemented by labour courts. Courts of assizes sit in each province to judge crimes and political and press offenses. These are composed of three judges and 12 citizens chosen by lot.

Structure of the judiciary

The Supreme Court of Justice is composed of three chambers: civil and commercial, criminal, and one for matters of social and fiscal law and the militia. The last court does not deal with cases in depth, but regulates the application of the law throughout all jurisdictions. The military jurisdictions judge all cases concerning offenders responsible to the army and, in time of war, those concerning persons accused of treason. The State Council arbitrates in disputed administrative matters and gives advice on all bills and decrees.

Armed forces. Under the constitution, the monarch is required to "maintain national independence and territorial integrity," is charged with "the command of land and sea forces" and the making of treaties of alliance and of peace, and is given the power to confer ranks in the army. Belgium's defense policy is, in fact, determined by the Council of Ministers, although the constitution limits the government's powers in military matters by requiring the intervention of the legislature to determine the method of recruitment into the army, the size of the annual conscription draft, and the promotion, rights, and obligations of military personnel. The legislature is also required to decide military pensions and retirement ages, as well as on the deprivation of military ranks, honours, and pensions.

Belgium is a member of the North Atlantic Treaty Organization (NATO).

Education. Freedom of education is a constitutional guarantee in Belgium, but conflicts between public and confessional (*i.e.,* Roman Catholic) schools date almost to the founding of the kingdom and remain among the most delicate problems within the social fabric. Currently, a dual system of state and "free" schools exists on the primary and secondary levels, the latter subsidized by the state to compensate for the abolition of fees in 1958. The language of instruction is either French or Dutch, depending on the region. Secondary schools are graded into a lower category, staffed by graduates from teachers' colleges, and institutions staffed by university graduates and offering either a classical or modern curriculum.

In addition to numerous specialized institutions for advanced training, Belgium has several universities. The Catholic University of Louvain (1425) and the Free Brussels University (1834), both formerly bilingual, were each divided into independent Dutch- and French-speaking universities (thereby creating four universities) in 1970. The University of Liège (1817) and the State University of Mons (1965) teach in French, the State University of Gent (1817) in Dutch.

Health and welfare. The great improvement in health conditions during the middle of the 20th century has been due as much to the programs of social insurance, covering about 95 percent of the population, as to advances in medical science. In addition to the many hospitals, hundreds of centres offered specialized help in medical, psychological, and geriatric areas as well as in rehabilitation of the handicapped. Under a 1925 statute, each commune has a commission of public assistance that is represented on the communal council; it acts to provide aid to the indigent.

Housing. Building is encouraged in a number of ways, including government-guaranteed mortgage loans that have lower interest rates. The National Housing Society oversees public housing construction for low-income families, and the state also is in charge of eliminating slum conditions in urban areas or where immigrant workers are forced to seek housing.

Cultural life. The cultural and artistic heritages of the Belgian people are long and rich, reaching high points in the painting of Pieter Bruegel the Elder; the music of Josquin des Prêz, Orlando di Lasso, and César Franck; the dramas of Maurice Maeterlinck and Michel de Ghelderode; and in the many palaces, castles, and cathedrals of the Belgian cities and countryside.

The cultural milieu

A basic cultural diversity exists between the Flemish and Walloon sectors. In general, the Walloons are attuned to the vibrations of Paris, whereas the Flemings are more concerned with maintaining and broadening their own cultural identity without strong reference to The Netherlands. Cultural unity is stressed throughout the country by means of the institutions known as "houses of culture," a use to which the former royal palace in Antwerp, for example, has been converted. The Walloons have emphasized half-cultural, half-touristic activities of the "poetry festival" type.

A variety of trends has resulted from this cultural diversity. Various schools of art have arisen, particularly in Vlaanderen. In music, avant-garde tendencies have become influential in Brussels, Liège, Ghent, and Antwerp; while Hainaut remains the centre of the classical and popular traditions. Literary works produced in Vlaanderen have a style peculiar to the region; whereas, in the Walloon area and in Brussels, most authors are trying to write for a readership that is more inclined to French, and especially Parisian, works. Moreover, some of these same French works are by Belgian authors living in France, and still others are by writers living in Belgium but considered French.

Political decentralization and its corollary, regional autonomy, both of which characterize Belgium, have engendered their cultural counterpart. Brussels has lost its former dominance in the arts to certain provincial capitals, such as Antwerp, Liège, and Gent, which have become cradles of contemporary artistic creation, although the National Orchestra in Brussels remains one of the country's most important symphonic ensembles. Some of the provinces have their own music festivals, such as the Flanders Festival, as well as their own orchestra.

The renowned Queen Elisabeth of Belgium International Music Contest attracts talented young artists from all over the world. The international String Quartet Competition in Liège and the Young Musicians' Competition were first held in Belgium.

The traditional culture is based on the celebration of the seasons. In the Walloon areas there are joyous spring festivals, such as the carnivals of Binche and Stavelot;

summer festivals, such as the procession of giants at Ath and the dragon battle in Mons; and the winter festivals of St. Nicholas, Christmas, and New Year. In Vlaanderen, these festivals have become folkloric celebrations with a religious or historic character. Notable festivals include the Festival of Cats in Ieper and the Procession of the Holy Blood in Brugge. Finally, marionette shows survive in the Toone Theatre in Brussels. This traditional folk culture is in strong contrast to modern forms of popular culture, which as everywhere in the West are dominated by television and cinema and exemplified by the pop-music festival held first at Comblain-La-Tour and then at Amougies.

The artistic heritage is preserved in the many museums located throughout the country and in the very structure of countless buildings and parks. The Musée Instrumental du Conservatoire Royal in Brussels has a fine collection of musical instruments. The framework for cultural and artistic activities is provided by fine arts palaces, operas, the "houses of culture," and other cultural centres in Brussels and the provinces and by private and subsidized theatres and the five royal conservatories of music in Antwerp, Brussels, Ghent, Liège, and Mons.

Newspaper groups

The many daily newspapers published in Belgium are controlled by press consortiums. Each paper or chain within these larger groups preserves its autonomy, with cooperation limited to technical and advertising matters. A German-language daily is published in Eupen. The majority of newspapers have some political affiliation, but only those of the Socialist press are linked to a political party. Belgium has several periodicals, but these face strong foreign competition. Nevertheless, the women's and the children's press are more widely distributed abroad than in Belgium.

Radio broadcasting was born in Belgium. As early as 1913, weekly musical broadcasts were given from the Laeken Royal Park. Radio-Belgium, founded in 1923, was broadcasting the equivalent of a spoken newspaper as early as 1926.

Belgian Television Broadcasting (RTBF), which broadcasts in French, and Belgian Radio and Television (BRT), in Dutch, were created as public services. They are both autonomous and managed by an administrative council of 10 members, of whom eight are named by Parliament and, in fact, represent the traditional parties. The two systems share a service institute, which assures administrative and financial coordination, and are financed by a tax on receivers. Unlike neighbouring countries, Belgium forbids advertising on radio and television. The country's small size and linguistic divisions limit the resources available to each of the systems. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.* (Ar.D.)

HISTORY

It is sometimes suggested that Belgian nationality dates from 1830–31. After the Burgundian regime in the Netherlands and the abdication of the emperor Charles V, the country was subject first to Spain, then to the Austrian Habsburgs, then to France, and finally to the kingdom of The Netherlands, union with which ended with the 1830 revolution. But the Belgian provinces preserved their ancient institutions and traditions, and it was only for a short period, under the First French Republic and Napoleon, that integration with an alien system could be enforced.

The Burgundian period (1363–1477), from Philip II the Bold to Charles the Bold, was one of political prestige and economic and artistic splendour. The finest monuments of the country and its most famous paintings date from the 15th century, and the "Great Dukes of the West," as the Burgundian princes were called, were rightly considered as national sovereigns, their domains extending from the Zuider Zee to the Somme. The death of Charles the Bold (1477) and the marriage of his daughter Mary to the archduke Maximilian of Austria proved fatal to the independence of the Low Countries by bringing them more and more under the sway of the Habsburg dynasty.

After the death of Maximilian's son the archduke Philip the Handsome (1506), the provinces were wisely governed by Margaret of Austria, Philip's sister, the aunt of the young prince Charles who was to become king of Spain as Charles I in 1516 and Holy Roman emperor as Charles V in 1519.

The Spanish Netherlands (1579–1713). In Brussels on October 25, 1555, Charles V abdicated the Netherlands to his son, who in January 1556 became king of Spain as Philip II. Under Spanish rule discontent increased in the Netherlands, and revolution broke out in 1567, but the union between Catholics in the south and Protestants in the north could not be maintained after the first years of conflict.

The formation of the Union of Arras (January 7, 1579) by the Catholic provinces of Artois and Hainaut enabled the Spanish commander Alessandro Farnese (later the duke of Parma) to resume the war against the Protestants of the Netherlands. William I of Orange, who was supported by the Union of Utrecht (January 23, 1579), at first tried to set up François, duc d'Anjou and brother of the King of France, in opposition to the King of Spain; but the assistance of the ambitious French prince proved to be useless. He arrived in Antwerp early in 1582 and attempted in the following year to gain possession of the city by force (January 17, 1583). This act of treachery put an end to his pretensions and promoted the success of Farnese. On November 30, 1581, the latter had captured Tournai. After a series of sieges he made himself master of Ypres, Dunkirk, Brugge (Bruges), Gent (Ghent), Brussels, and finally Antwerp, which capitulated on August 17, 1585. From this time onward the whole of the southern part of the Netherlands once more recognized Philip II as its sovereign. In 1598 Philip II granted the sovereignty of the Netherlands to his daughter Isabella and her husband, Archduke Albert of Austria.

The northern provinces were never recovered, and Spain allowed only a nominal independence to the south. In 1609 Albert was forced to make a 12-year truce with the United Provinces. He died in 1621, the same year that the war was resumed, and Isabella was, from that time on, nothing more than a governor general. During the course of the war with the United Provinces, northern Brabant, as well as the region to the east of the Meuse, and Zeeland were lost (1621–48). In addition, France, which was allied to the United Provinces, seized Artois (1640). Philip IV of Spain agreed to the new northern boundary of the Spanish Netherlands in the Peace of Münster (January 30, 1648), which also confirmed the closing of the Scheldt River to foreign shipping, destroying the commercial importance of Antwerp. Hostilities between Spain and France continued during most of the latter part of the 17th century. The treaties that interrupted this conflict were marked, in almost every instance, by new losses of territory by the Spanish Netherlands. The Treaty of the Pyrenees (1659) confirmed the loss of Artois and provided for a series of French fortifications to be built along the border of what is now Belgium. By the Treaty of Aix-la-Chapelle (1668), France occupied part of Flanders. In 1700 the Spanish Habsburg dynasty died out with Charles II, who had turned his holdings over to Philippe d' Anjou, grandson of Louis XIV of France. The Spanish Netherlands was thereafter under French rule until it was occupied by the British and Dutch in 1706. By the Treaty of Utrecht (1713), ending the War of the Spanish Succession, the territory comprising present-day Belgium and Luxembourg passed under the sovereignty of the Holy Roman emperor Charles VI, head of the Austrian branch of the House of Habsburg.

Territorial losses

Administration. The government of the Spanish Netherlands, though not independent, was characterized by a large degree of autonomy. The king was represented in Brussels by a governor general who was, more often than not, a prince of the Spanish royal family. The three councils that assisted the governor (the Council of State, the Privy Council, and the Council of Finances) were composed almost entirely of local leaders, though the Council of State did include a certain number of Spanish nominees. The president of the Privy Council became a sort of first minister. During the 17th century the principal presidents were Pieter Roose (1632–53) and Charles de

Hovyne (1653–71). These men did not hesitate to show independence of Madrid in order to protect their interests, but they were essentially supporters of absolutism and regularly tried to assert the authority of the royal government at the expense of regional and local rights. After 1664 the Council of Finances, under its chief official, the treasurer general, began to act as a sort of ministry of economic affairs.

Internal autonomy

These councils gained considerable autonomy in internal affairs, but, with respect to foreign policy, they were controlled less by the governor general than by a Spanish official in Brussels called the secretary of state and war. In Madrid, there was a council of state for the Netherlands, the members of which were natives of the Belgian provinces. In this delicately balanced administration, divided between Brussels and Madrid, the Spanish element tended to predominate, especially in the field of foreign affairs.

There were representative bodies in the Spanish Netherlands. The most important were the provincial estates (assemblies) whose power lay in their authority to levy and collect taxes. This authority enabled them to see that a considerable portion of the revenue was spent within the country. A permanent deputation drawn from the estates supervised public works.

Before and during the revolt against Spain, the States-General, composed of delegates from all the provincial estates, had enjoyed great influence. When the country returned to Spanish rule, the role of the States-General was diminished; after 1632 it no longer met.

During the 16th century regionalism was deep-rooted in the provinces, but in the 17th century a wider feeling of unity developed. The nobles, angered by the government's centralizing policies, revolted in 1631–34; the aristocratic provincial governors involved in this revolt were forced to flee the country because the towns did not support them. Some were not replaced, and by 1700 only Hainaut, Luxembourg, Namur, Limburg, and south Gelderland, all of which had proved their loyalty, still had provincial governors.

In judicial matters, the supreme authority was still the Great Council of Malines, founded in 1504; but it had to defend its jurisdiction against the encroachments of the Privy Council. The provincial courts of justice were the councils of Flanders, Brabant, Namur, Luxembourg, south Gelderland, Hainaut, and (until 1659) Artois. The most important of these was the Council of Brabant, to which the king had originally granted autonomy in conformity with the provincial liberties of Brabant. Nevertheless, after 1603, he was represented in Brabant by financial officials under a procurer general. These magistrates, in accordance with a practice common at that time, had, in addition to their judicial duties, administrative functions that increased during the 17th century.

In the 17th century, the administration of the country was made difficult by constant warfare. Government finances, weakened by the loss of revenues from the northern provinces and the loans contracted during the 16th century, suffered further because of increased military expenses. Sale of public offices became a common practice.

Taxation

At the end of the 16th and the beginning of the 17th centuries, import and export duties provided a new source of revenue. Prior to this, the shipping of goods was taxed only as *tonlieu* (goods transported within the country). This tax remained but only supplied regional or local finances. Taxes on foreign trade originated from permits allowing trade with the rebellious northern provinces. At first these were sold to individuals, but they were later replaced with fixed duties that, by 1654, had become customs tariffs. There were then customs houses not only at the frontiers but in Brabant and Flanders as well.

In the country's military organization, the fortresses of Antwerp, Ghent, Ostend, and Charleroi were manned by Spanish troops; other armed forces were raised locally.

The principality of Liège (in present eastern Belgium) was ruled as a separate territory by its bishops, as it had been since the Middle Ages. During the revolt against Spain, it maintained a strict neutrality and continued to do so during most of the 17th century. Its institutional development was, in the main, parallel to that of the Spanish Netherlands.

Economy. The economy of the country had, of course, suffered greatly during the struggle against Philip II, but recent research has shown that economic recovery was more rapid and more far-reaching than had previously been thought.

The fortunes of the important linen industry can be traced by studying tax fluctuations. In the region of Oudenaarde, excise taxes almost doubled between 1570 and 1579. The figure for 1579 represents the highest point of the century; the low point occurred in 1589. Slightly earlier, a similar pattern occurred in Courtrai. Thus the decade 1570–80 was a favourable one for the linen industry. A period of crisis followed, but after 1590 there was a marked recovery. In 1600, Oudenaarde again reached the standard of 1561, while Courtrai surpassed that of the preceding century. The 17th century, formerly thought to have been a time of unrelieved calamity, now appears to have known considerable prosperity. Between 1609 and 1610 production doubled in Courtrai; in 1642 it was eight times that of the best years of the 16th century.

A similar situation existed in other branches of the textile industry. In the sayette (serge) mills at Hondschoote, the 10 years following 1570 were still a period of great prosperity; the crisis did not begin until about 1580. In this instance the lowest point also occurs between 1580 and 1590. Historians formerly believed that there was a massive emigration of Flemish weavers to the northern Low Countries during the period of turmoil. Recent research has shown that this emigration was on a much smaller scale than had previously been thought and that in 1588 Hondschoote, Bruges, and Ypres were again very active.

The economic crisis in Flanders, the most advanced region in the southern Netherlands, was so short-lived that its existence was not generally realized. It coincides with the reconquest of Alessandro Farnese, in 1585. When this is understood, evidences of prosperity in the 17th century become more comprehensible. After 1600, the war between Spain and the United Provinces and repeated invasions by the French did cause havoc, but the basis of the economy was not seriously threatened until the War of the League of Augsburg (1689–97) and the War of the Spanish Succession (1701–13). In the second half of the century, despite the fact that the Dutch were able to close Antwerp to sea traffic, commerce was kept alive by transporting goods overland from the ports of Holland and Zeeland. Other goods entered the country by the ports of Ostend, Nieuwpoort, and Dunkirk. The digging of new canals facilitated internal transport, and overland trade with France, Germany, and Italy remained active.

The populations of the important cities steadily increased during the 17th century and declined after 1700. In the small towns and the countryside, the rate of expansion continued into the 18th century. The development of the rural linen industry is significant in this connection. The increase in urban population during the 17th century can be explained by the prosperity of the export trade, mostly in the various branches of the textile industry. In Ghent, silks, velvets, and blended weaves of wool, flax, silk, and cotton were all produced in increasing quantities. In Bruges, serge and fustian (a cotton-linen weave) were manufactured; Ghent specialized in luxury fabrics, Bruges in cloth for everyday use. The products of these towns had an important place in the European and Spanish-American markets.

Ghent's textile industries

Linen dyers were especially numerous in Ghent. They processed large quantities of linen, which were shipped mainly to Spain and England. The majority of looms belonged to a restricted number of manufacturers. This concentration of the means of production in the hands of a few large entrepreneurs was increased by the mid-18th century to a point that presages the beginning of the Industrial Revolution. For a time Ghent rivalled Antwerp and Brussels as a centre of tapestry manufacture and trade, mainly because of the immigration of merchant entrepreneurs from Oudenaarde. In 1687 one manufacturer employed 200 labourers. In addition to simple floral designs, Ghent tapestries featured religious, mythological,

and historical subjects in the tradition of the great tapestry centres of Flanders. These goods were exported to France, Germany, Denmark, and the United Provinces.

Bruges in the mid-17th century had a population of about 35,000; some 1,500 adult labourers were employed there in the textile industry as weavers and in related jobs. If the wool combers and spinners living in the villages around Bruges are also counted, the total comes to more than 10,000.

The prosperity evident in the Flemish textile industry was not equalled in other areas of the economy, although metallurgy and mining in the principality of Liège were of considerable importance. Agricultural production was high enough to avert famine but increased too slowly to provide a base for much industrial expansion.

At Antwerp, the economic recovery that marked the end of Philip II's reign continued until at least the middle of the 17th century. The town had 42,000 inhabitants in 1589; 57,338 in 1644; and 65,711 in 1698. Although it lost much of its maritime commerce to other cities, Antwerp continued to prosper as an industrial centre. Goods made from Italian silk were exported to France, England, and Holland; ribbon and thread went to Spain and other parts of western Europe. English cloth continued to be distributed through Antwerp, thanks to a certain number of local entrepreneurs and to English merchants settled there. The city produced a great deal of lace for export; during the second half of the century a single lace manufacturer employed 300 persons. This lace was, for the most part, sent to Spain and Latin America. Antwerp was also a centre for the distribution of artworks throughout Europe and America. The paintings of Flemish artists were sent everywhere, and those of Italian masters were often distributed throughout western and central Europe via Antwerp.

Foreign trade at Antwerp

The same was true in printing and the book trade. Printers were important and numerous; the firm of Moretus was the largest industrial concern in Antwerp and one of the principal firms in Europe. The diamond industry and trade had an impressive development. Antwerp had 104 diamond cutters in 1608; by 1674 their number was 200. People came there from Frankfurt, Hamburg, London, and Paris to perfect their craft. The highly speculative diamond trade—the value of diamonds being contingent on their scarcity—was well suited to a city that dealt mainly in foreign commerce. Commercial firms of international importance were still centred in Antwerp. A recent estimate lists 46 of these in the period 1620–40, some Flemish and some foreign, a number of them controlling large amounts of capital. One of these firms, newly established and moderately prosperous, had commercial ties with Rouen, Calais, Lille, London, Dover, Canterbury, Amsterdam, Leiden, and Hamburg. It sold sugar and other colonial products, Spanish wool, and textiles. Within 30 years this company passed from ordinary trade to banking. It is clear, then, that the economic depression formerly attributed to the 17th century did not really begin until the latter part of that century and the beginning of the 18th.

The relative prosperity of the southern Netherlands, and particularly of Antwerp, no doubt partly explains its flourishing artistic life during this period. This was chiefly evident in the works of the Flemish School of 17th-century painters—among them Peter Paul Rubens, Anthony Van Dyck, and Jacob Jordaensz. Rubens (1577–1640), court painter to Isabella and Archduke Albert, made Antwerp one of the cultural capitals of Europe. In the area of scholarship, the "Bollandists," a group of Antwerp Jesuits, made valuable contributions to historical methodology.

The Austrian Netherlands (1713–95). Under the Austrian, as under the Spanish, Habsburgs, the southern Netherlands enjoyed political autonomy. But its prosperity had been ruined by the wars of the late 17th and early 18th centuries, and discontent was widespread. By 1719 there was opposition in Brussels to taxes levied by the new government. The emperor Charles VI attempted to relieve the economic distress of his new possession by founding the Ostend Company to trade with India (1722), but Britain and the United Provinces, wishing to avoid unwelcome competition, forced him to abandon the company in 1727.

Economic decline

At the death of Charles VI (October 20, 1740), the southern Netherlands passed to his daughter Maria Theresa without opposition. The War of the Austrian Succession, however, led to a new French occupation in 1744. Austrian rule was restored by the Treaty of Aix-la-Chapelle in 1748.

The regime of the empress Maria Theresa of Austria enjoyed popularity as the economic situation began to improve again toward the middle of the 18th century. As in contemporary England, an increase in agricultural productivity stimulated a population increase, especially in rural areas. This, in turn, spurred the development of various industries. The agricultural transformation occurred mainly on the small farms of Flanders; one of its main features was the spread of potato cultivation, which added an important element to the diet of the rural population. In addition, in the French-speaking part of the country, a number of landed proprietors invested in mining enterprises, notably in the area between the Sambre and the Meuse rivers, which belonged to the principality of Liège. In the southern Netherlands, urban merchants and manufacturers had more in common with the rural landowning class than was usual in continental European countries in the 18th century. As in the case of Britain, this created an atmosphere favourable to the development of industrial capitalism. During this period, Ghent, Antwerp, and Tournai had factories with more than 100 workers; wages, however, were poor. Verviers, in the principality of Liège, was an important centre for woolen manufactures, Ghent for cotton goods.

Under the influence of the Enlightenment, the Austrian regime sponsored a number of reforms in the Netherlands. An academy was founded in 1772 and, when the Jesuits were suppressed in 1773, the government replaced their colleges with secular institutions. The government was not inclined to support the remnants of feudal privilege that caused so much discontent in 18th-century France.

Not all of its innovations were popular, however. The emperor Joseph II (reigned 1780–90) suppressed 163 religious houses, granted freedom of worship by his Edict of Toleration (1781), and closed the episcopal seminaries, replacing them with two state seminaries (1786). These measures alienated the Roman Catholic clergy. In 1787 the administration was modernized by replacing the traditional provinces with districts (*cercles*) governed by intendants. This caused dissatisfaction among the groups that controlled the provincial estates. The discontent of the privileged classes was increased by the suppression of the traditional councils and courts of justice, which, like the estates, were looked upon as safeguards of provincial autonomy.

Joseph's sister, Maria Christina, duchess of Saxony-Teschen, who was governor general, was reluctant to implement the unpopular edicts abolishing these traditional bodies. But other members of the administration, including Ferdinand von Trauttmansdorff, the Emperor's minister plenipotentiary, and Gen. Richard d'Alton, commander of the troops, insisted upon their enforcement. The estates of Hainaut and Brabant reacted by refusing to pay taxes (November 1788). In the following year, influenced by the outbreak of revolution in neighbouring France, conservatives led by Henri van der Noot and progressives led by Jean-François Vonck united in opposition to the Emperor. Raising a corps of volunteers, Vonck defeated an Austrian force at Turnhout (October 24–27, 1789). Nervous because of the situation in France, the Austrians retreated to Luxembourg, and van der Noot made a triumphant entry into Brussels, where he and his "Statists," were supported by the Estates of Brabant, this time not only against the Emperor but also in opposition to the progressive Vonck, who favoured the convening of a national assembly. They successfully incited the people of Brussels against Vonck, who had to take refuge in Lille.

The Brabant Revolution

This "Brabant Revolution" (so called because most of its leaders came from Brabant) had widespread support in the towns; but the peasants, having little in common with the middle class revolutionaries, generally supported the Austrians. This was particularly evident among that segment of the rural population that lived, at least in part,

by domestic work for the new manufacturers in the towns. Thus, in Hainaut, the province where the estates had first rebelled, the peasants of the countryside of Chimay rose up in favour of the Emperor and against the Statists.

Thus, when Joseph II died, and his successor, Leopold II, decided to reestablish imperial authority, he encountered no opposition from the mass of people. On December 2, 1790, Brussels was reoccupied by Austrian troops. Nevertheless, in order not to provoke the Statists, the Emperor did not reintroduce his predecessor's reforms. The discontented now looked to revolutionary France for support. When the French general C.F. Dumouriez occupied the Austrian Netherlands after his victory at the Battle of Jemappes (November 6, 1792), he was very well received. Enthusiasm languished, however, when it became clear that France was not going to allow the Netherlands to be independent. Austrian rule was briefly restored in 1793–94; but France's victory at Fleurus, on June 26, 1794, was the prelude to French annexation. This proved more of a jolt for the Austrian Netherlands itself than for the principality of Liège, where a revolution against the prince-bishop had prepared the country for assimilation into the French Republic.

French rule. On October 1, 1795, the French National Convention voted to annex Belgium and the principality of Liège, which was thenceforth amalgamated with it. Under French rule, there was no autonomy as there had been under the Spanish and Austrian regimes; the reforms of Joseph II were revived, although in a more radical form. The administration was centralized and the territory divided into departments. Aristocratic privileges were abolished and the church persecuted. Military conscription, introduced in September 1798, provoked a peasants' revolt in the same areas that had opposed the conservative Statists during the Revolution of 1789. Repression was extremely harsh.

Under the Napoleonic Consulate and Empire (1799–1814), the French civil code and the decimal metric system were introduced, and the position of the clergy was regulated by Napoleon's concordat with the papacy. A politically united Europe offered vast outlets for expanding Belgian industry.

The beginning of industrialization

The period of the Napoleonic Empire, recent research has revealed, may be considered the "take-off" stage of the Industrial Revolution in Belgium. The Belgian departments produced almost half the empire's coal; the department of Jemappes (the former county of Hainaut) alone accounted for 30 percent. The blast furnaces and forges of the Belgian departments provided 25 percent of the empire's iron production. In relation to its population, Belgium produced 450 kilograms (992 pounds) of coal per capita. At the same time the corresponding total in France was 40 kilograms (88 pounds). That this was the beginning of the Industrial Revolution may be seen by a comparison with contemporary England, where this revolution was in progress, and with France, where it had not begun. The 450 kilograms (992 pounds) of coal per capita are comparable to the English 500 (1,100 pounds) but not to French production, which was one-tenth that amount. Figures for iron production were similar: 11 kilograms (24 pounds) per capita in Belgium, 20 (44 pounds) in England, and 4 (9 pounds) in France. The Verviers textile industry was second in the empire. In Flanders, linen and cotton occupied an important place; the same was true for paper, leather, and glass. Thus, Belgium was the first country on the Continent to go through the process of industrialization. However, during the later years of French rule, it became evident in Belgium, as it had to all the peoples subjugated by Napoleon, that military domination had had its day. The allied occupation of Belgium in 1814 was welcomed.

Kingdom of The Netherlands. The allied powers were determined not to leave France in possession of Belgium. It could not be given back to the Austrian emperor, who in any case did not desire it. Under the influence of Great Britain it was decided to unite it in a single state with the old republic of the United Provinces and thus to constitute a new barrier better capable of resisting future French expansion than that of 1715. The Kingdom of The Netherlands, the existence of which was confirmed by the Treaty of Vienna (June 1815), was thus established for the convenience of Europe, regardless of the wishes of the Belgians and the Dutch who would have to live side by side in one country. Prince William of Orange ascended the throne on March 16, 1815, under the title of William I; he was crowned September 27.

Thus the northern and southern Netherlands, which had been one country until the 16th century, were reunited. But in the two intervening centuries they had developed in markedly different ways. The north was commercial, the south increasingly industrial; the north Protestant, the south Catholic. Under the Dutch House of Orange, the north was to be predominant; the fundamental law of the new kingdom gave Belgium and Holland the same number of representatives in the States-General despite the fact that the population of Belgium was 3,500,000 and that of the former United Provinces 2,000,000. Belgian representatives rejected the constitution, but it was promulgated by the King over their objections.

In the beginning, the economic situation continued to be favourable. The textile industries of Ghent and Verviers continued their progress, while the coal mines and forges of Liège and Hainaut prospered. King William I created three state universities: Ghent and Liège, which were new, and Louvain, which he put under state control to remove it from Catholic influence. In secondary education, the secular *athénées* were established, and schools controlled by the church were subject to state inspection. An attempt to interfere with the training of priests (1825) brought clerical dissatisfaction with the government to its height.

Dissatisfaction with Dutch rule

After 1820 the conflicting interests of north and south also created an economic split. The commercial north, with little industry, desired free trade; the south needed tariff protection to allow its industry to compete against British prices. This gave the middle classes a grievance against the government. Progressives and clericals joined forces in 1828; the programs were different, but the enemy was the same: the Dutch regime. Following the July Revolution in Paris (1830), rioting broke out among the workers in Brussels on August 25. The bourgeoisie armed themselves, and, when the working class uprising was put down, they turned their arms against the Dutch troops. The Dutch, led by the Prince of Orange, were expelled from Brussels on September 3 and tried unsuccessfully to reoccupy it on September 23. Two days later a provisional Belgian government was established. This revolutionary government proclaimed the country's independence on October 4. Meanwhile, the Belgians occupied Antwerp; Dutch troops, besieged in the citadel, bombarded the city on October 27, further increasing ill-feeling among the Belgians.

A National Congress was elected and met on November 10. It reaffirmed Belgian independence on the 18th and on the 24th voted to exclude the House of Orange from the Belgian throne. William I prepared for war, but on December 20 the great powers intervened, imposing an armistice on both sides. On January 20, 1831, a conference in London recognized Belgium as an independent, neutral state, its neutrality to be guaranteed by the powers.

Independent Belgium. *The election and reign of Leopold I.* The National Congress had decided that Belgium should be a monarchy, but finding a king proved difficult. On February 3, 1831, the congress elected Louis, duc de Nemours, son of the new French king Louis-Philippe, but the British, fearing French expansionism, opposed this, and Louis declined. The next choice was Prince Leopold of Saxe-Coburg, who was related to the British royal family. Because Leopold married Louis-Philippe's daughter shortly afterward, he was acceptable to the French also. On July 21, 1831, he promised to support the constitution drawn up by the National Congress, which gave the greater part of the governing power to a parliament elected by property owners. Some days later, the Dutch army invaded Belgium. The Belgians, who had no regular army, were defeated, but the London Conference agreed to intervention by the French army, which forced the Dutch to retire. The conference then decided (November 15, 1831) to divide the provinces of Limburg and Luxembourg,

Conflict with The Netherlands

awarding part to Belgium and part to The Netherlands. William I, to the surprise of the powers, refused to accept this settlement, although he did conclude an armistice. The Belgians, therefore, continued to occupy Dutch Limburg and the future Grand Duchy of Luxembourg until William finally relented in 1838. He then became the first grand duke of Luxembourg. The western half of Luxembourg became a Belgian province. The new Belgian king Leopold I tried to increase his role in the government by asking Parliament for the right to appoint and dismiss the members of the provincial and local administrations. Parliament, however, rejected this proposal in 1836.

The economy continued to develop. The first railroad, between Brussels and Malines, was inaugurated in 1836. Capital investments rose from 37 million francs in 1834 to 145 million in 1836. One of the most powerful corporations was the Société Générale, which now controlled a number of industries. At the same time, the traditional Flemish textile industry, which was slow to mechanize, was undergoing a severe crisis. But by 1848, the decisive leap to industrialization was made in this area also.

After Dutch recognition of Belgium in 1839, the coalition that had backed the independence movement showed signs of breaking up. Disputes arose between the liberals (middle-class anticlericals) and the Catholics, and a separate Liberal Party was formed in 1846. Between 1847 and 1857, Catholics and Liberals alternated in office; the period 1857–70 was one of Liberal rule.

Leopold I proved to be a prudent ruler, and when revolution spread across Europe in 1848, Belgium remained peaceful. But the downfall of Louis-Philippe and, to an even greater extent, the creation of the Second French Empire under Napoleon III (1852) eventually caused problems. French republicans sought refuge from political persecution in Belgium, where they spread propaganda against Napoleon. This caused the French government to adopt a hostile attitude toward Belgium; in 1853, Leopold asked that Parliament allow Antwerp to be fortified and the nation to be prepared to meet a possible invasion, which, however, failed to materialize.

Further economic progress

In the economic sphere, industrial development allowed the gradual introduction of a free-trade policy beginning in 1849. A national bank having the right to issue currency was created in 1850. Municipal tolls on the entrance of foodstuffs into the cities were abolished in 1860. Finally, in 1863, the prosperity of the country permitted the redemption of The Netherlands' right to levy charges on shipping entering the Scheldt River. After more than two centuries, Antwerp was a free port again. Leopold I died on December 10, 1865, leaving Belgium a strong and united country.

The reign of Leopold II. Leopold II (reigned 1865–1909) found himself immediately faced with serious problems in foreign affairs. Prussia had become the dominant power in Germany, and Napoleon III wished to extend French control over Belgium and Luxembourg in compensation. A French rail company, with government support, attempted to acquire the railways of the province of Luxembourg in 1868; in 1869, the Belgian chamber reacted by forbidding the sale of railways. Only British support permitted the liberal ministry of Walther Frère-Orban to avoid hostilities with France.

When the Franco-German War broke out in 1870, Belgium remained neutral. Napoleon III's defeat momentarily removed the French threat, but a unified Germany posed an equally serious problem on the eastern frontier. The subsequent rivalry between France and Germany in the period 1871–1914 constituted a continual danger to Belgium. In 1887 Leopold II asked Parliament for authority to erect two new fortifications, one to the east, at Liège, and the other to the south, at Namur.

Meanwhile, the process of industrialization continued. Coal and iron production increased; Belgium's population was the densest in Europe and, because there was no labour movement before 1870, labour was cheap. The Société Générale, one of the most powerful economic organizations in continental Europe, played an increasing role in financing industry.

The Belgian bourgeoisie increased both in number and wealth, and electoral property qualifications favoured their predominance. This class was the backbone of the Liberal Party. A Catholic reaction materialized in 1869, with the formation of a Catholic party opposed to the liberal Catholic deputies who, until then, had represented Catholic opinion in Parliament and in the government. The new party represented a militant and conservative position; in order to achieve its goals, it favoured eliminating property qualifications for voting, a measure also favoured by the more advanced Liberals. Socialism gained some adherents in Belgium after 1848, and the formation of the First Socialist (Marxist) International in 1867 increased its influence among industrial workers. Belgian socialism remained fragmented until the foundation of the Belgian Workers' Party (POB), the ancestor of the later Belgian Socialist Party (PSB), in 1885.

Church–state conflict

The opposition between Catholics and Liberals became acute in 1879, when the Liberal Frère-Orban ministry sponsored a law removing religion entirely from primary education. Controversy over this issue was extremely violent and even divided families. The church's opposition to the law led to a break in relations with the Vatican. In 1884 the Catholics won a majority in both chambers, and the Catholic party remained in power without interruption until 1914. Aside from the education controversy, the biggest factor in the Liberals' defeat was probably their advocacy of free trade, which was favoured by manufacturers but exposed farmers to ruinous foreign competition. In the early 1880s the Belgian market was flooded by American grain; the Catholic party became the champion of the rural classes by promising to protect agriculture. It also espoused the cause of the nascent "Flemish" movement that sought to expand opportunities for Dutch-speaking Belgians in a country until then dominated by a French-speaking upper bourgeoisie.

The working class did not have the right to vote because of the system of electoral property qualifications. The last years of the 19th century and the first of the 20th were years of violent social tension. In 1886 there was a disturbance among workers in Liège, followed by unrest in other industrial areas. The Catholic government of Auguste Beernaert suppressed this movement harshly, but beginning in 1889, a series of laws were passed regulating workers' housing, limiting labour by women and children, and providing workmen's compensation. The legislature revised the constitution in 1890, and, in 1893, universal suffrage was adopted for men 21 and over. Though this measure was modified by giving a plural vote to electors fulfilling certain conditions of income, age, and education and to heads of families, it resulted in the election of the first Socialist deputies to the legislature. Belgian industry, dominated by powerful financial groups, began to assume worldwide importance and was active in Asia and Latin America, as well as in Europe. In Africa, King Leopold II acquired the state of the Congo as a personal possession in 1885, but his subjects showed little interest in the area, leaving him to mismanage it until 1898, when the Matadi-Léopoldville railroad was completed, facilitating access to the interior of the Congo Basin. Then Belgian banks began to push for annexation, which eventually came in 1908. In 1909 the army recruitment system, which until then had favoured the wealthy by allowing them to hire substitutes for military service, was finally reformed.

Annexation of the Congo

Belgium from 1909 to 1944. In 1909, Leopold II died and was succeeded by his nephew Albert, who became king as Albert I (December 23). Social unrest, which had distinguished the last years of Leopold's reign, continued at the beginning of Albert's, but a general strike in 1913 caused no change in the parliamentary majority.

The international situation became increasingly tense at the end of July 1914, on the eve of World War I. The neutrality of Belgium and Luxembourg had been guaranteed by the great powers under the treaties of 1839. On July 30 the British government demanded assurance from France and Germany that Belgian neutrality would be respected in the event of the outbreak of war. Germany, after refusing to respond to the English request, invaded Luxembourg on August 2 and sent an ultimatum to Belgium demanding free passage of troops. Belgium refused

and was invaded on the morning of August 4. Liège fell on August 16, Namur on August 23. The Germans occupied Brussels on August 20, but Antwerp was not evacuated until October 6. The Belgian Army then retired behind the Yser (IJzer) in the west of Flanders and held this position until 1918.

During the war, the Belgian government sat at Le Havre, France, while the King, as commander in chief of the army, remained with his troops in unoccupied Belgium. After the invasion, the Catholic government named both a Liberal and a Socialist minister of state, and, during the war, ministry offices were assigned to Liberals and Socialists. Upon their return to Brussels, the King and Prime Minister announced the introduction of absolute universal suffrage (that is, the abandonment of plural votes). The elections of 1919 brought to power a government composed of Catholics, Liberals, and Socialists, which voted an eight-hour day, progressive taxes, and old-age insurance. The question of schools, which had divided Liberals and Catholics so strongly during Leopold II's reign, gave rise to a compromise whereby subsidies were granted to Catholic institutions.

The Flemish movement

Flemish nationalism had gained ground, and, in 1898, the Dutch language was finally recognized along with French as an official language of the country. Henceforth, the movement also took on an economic and social character because in the coal-rich Walloon provinces industrialization was advanced. In contrast, economic development had been slow in the Flemish region after the Industrial Revolution. But, until 1914, the principal goal of the Flemish movement had been the creation of a university in which instruction would be given in Dutch. During World War I, Germany attempted to profit from Flemish–Walloon antagonism and supported the Activists—partisans of Flemish autonomy. Most Flemings, however, were resolutely hostile to collaboration with the enemy and refused to recognize either the Council of Flanders or the Flemish university, both founded during the occupation. Shortly after the liberation of the territory, the Belgian government made the University of Ghent, first partially, then completely, Flemish.

The Treaty of Versailles (1919) abolished Belgium's obligatory neutrality and returned the cantons of Eupen and Malmedy to its territory. In 1920 a treaty of military assistance was signed with France, and in 1921 an economic union was concluded with Luxembourg. The eastern frontier was guaranteed by the Pact of Locarno (1925). In Africa, Belgium received the mandates for Ruanda and Burundi, former German colonies that Belgian colonial forces had occupied during the conflict.

The decade following the Treaty of Versailles was distinguished by a considerable increase in industrial and commercial wealth and by prosperity among the workers and peasants. The middle classes, caught between these two forces, lost much of their wealth and political influence.

Postwar coalitions

The result was the retreat of the Liberal Party and the aggrandizement of the Catholic and Socialist parties. The three-party government born of the war did not survive the end of 1921; it was followed by a coalition of Catholics and Liberals until 1925, when the Socialists allied themselves with the Christian Democrats, then becoming the most important wing of the Catholic party.

In 1930 Parliament voted a series of laws transforming Belgium into two linguistic areas with different administrations. Henceforth, the Flemish provinces were administered solely in Dutch, the Walloon solely in French. In the face of the Great Depression the Socialist Party advocated a program of economic planning in accordance with the ideas of the socialist theorist Hendrik de Man. At the same time there emerged two Belgian fascist parties: a strictly Flemish party that enjoyed little success and the broader based Rexists under the leadership of Léon Degrelle. The latter party won 21 seats, more than 10 percent of the chamber, in the elections of 1936. Strikes broke out in the same year and led the tripartite government of Paul van Zeeland to establish paid holidays for workers and a 40-hour week for miners.

Meanwhile, King Leopold III, who succeeded his father, Albert I, in 1934, faced an increasingly tense international situation. Leopold advocated a policy of neutrality aimed at keeping Belgium from the seemingly inevitable conflict. Although this policy was approved by Parliament, Belgium, in its determination to resist all aggression, constructed a line of defense from Namur to Antwerp.

On May 10, 1940, Hitler invaded Belgium along with Luxembourg and The Netherlands. The Netherlands capitulated after six days, Belgium after 18. France, which along with Britain had sent troops to Belgium, had to lay down arms three weeks later. The British troops, covered by the Belgian army, retreated from Dunkirk in particularly dramatic circumstances. The four years of ensuing Nazi occupation were distinguished by a growing resistance organization. When the Allied forces reached Belgium on September 3, 1944, the Belgian underground army was able to prevent the destruction of the port of Antwerp, which served as a provisioning point for Allied troops for the remainder of the war.

Belgium after World War II. While the economy recovered rapidly after the war, political stability deteriorated, notably over the "royal question." The King had been taken by the Germans and confined to his palace at Laeken; the government, on the other hand, had taken refuge in London. In 1944, at the time of the Allied offensive, the Germans transferred the King to Austria, where he was kept until 1945. The government, upon returning to Brussels, conferred the regency on the King's brother, Prince Charles. Roughly, the Flemish were the King's partisans and the Walloons his opponents and the Catholics for, the Socialists and Liberals against, his return. In 1950 a referendum showed 57 percent of the voters to be in favour of the return of the sovereign, but the King's return signalled near civil war in the Walloon country. Leopold abdicated in favour of his eldest son, Baudouin I, who on August 11, 1950, took an oath of loyalty to the constitution. Because he was not yet of age, he took the title of prince royal until he became king on July 17, 1951. The composition of the government continued to fluctuate, although the Catholics achieved an absolute majority after the enfranchisement of women in 1949. The government was frequently composed of a leftist alliance of the Socialists and Liberals until the two most important parties, the Catholics and Socialists, formed a coalition that was often in power.

The "royal question"

In 1947 Belgium joined with The Netherlands and Luxembourg in the economic union of Benelux, and in 1949 the country became a signatory of the North Atlantic Treaty Organization (NATO); independence was given to the Congo (now Zaire) in 1960 and to Ruanda-Urundi (now the nations of Rwanda and Burundi) in 1962.

In the second half of the 20th century Belgian politics have increasingly revolved around two issues: the push toward regional autonomy of the Dutch- and French-language communities and, with accelerating urgency, severe economic problems. Concern over the economy has at times occluded and at times exacerbated the linguistic problem.

Tension between the language communities has provoked a number of governmental crises, including a six-month crisis in 1979, when four successive prime ministers-designate were unable to form a new government. It contributed to the dissolution of Parliament in 1985 and in 1986 to Prime Minister Wilfried Martens' offer to resign (rejected by King Baudouin).

In its efforts to alleviate the severe economic crisis, the government has repeatedly tried to avoid becoming embroiled in the divisive linguistic issues. This, however, is not always easy because various austerity measures taken by the government, such as funding for ailing industries in some sections, invariably affect one language community more than the other. Nonetheless, since 1982 several austerity measures have been adopted and imposed, and by 1986 the economy began to show signs of improvement.

Moderate constitutional reforms toward regional autonomy were achieved in the early 1970s, and in 1973 linguistic parity was ensured in the central government. The case of Brussels, however, presented a strong obstacle toward further progress. The country's capital, though located in the Flemish region, has a predominantly Francophone

population. It was only in 1980, the year of Belgium's 150th anniversary celebrations, that a major step toward the long-promised regionalization was taken. Two laws were passed approving the reorganization of the country's structure on the basis of three regions: Flanders, Wallonia, and Brussels.

For later developments in the political history of Belgium, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Nine, Division II, especially Section 923; and Division VI, especially Sections 961 and 963. (J.C.V./Ed.)

Luxembourg

The Grand Duchy of Luxembourg is a tiny sovereign state of only 998 square miles (2,586 square kilometres) surrounded by Belgium, France, and West Germany. A natural and historical meeting point of, and crossroads for, these countries, Luxembourg has come under the control of many states and ruling houses in its long history, but it has been a separate, if not always autonomous, political unit since the 10th century. The ancient Saxon name of its capital city, Lucilinburhuc ("little fortress"), symbolized its strategic and long impregnable position as "the Gibraltar of the north," astride a major military route linking Germanic and Frankish territories.

Strategic location

The heterogeneous peoples of Luxembourg and their language, economic pursuits, and ways of life reflect the grand duchy's multiplicity of common interests with its neighbours as well as the close historical relations. The farming villages and agricultural landscapes that abound across the broad and high northern tableland and the lower plateaus of the south delight the tourists, but in the few urban centres the smokestacks of industry signal an economic diversity. Luxembourg was an original member of the Benelux customs union, established in 1944, that linked its economic life with that of The Netherlands and of Belgium and that formed the core of the European Economic Community (EEC), or Common Market, a broader union created in 1958 that initially included France, Italy, and West Germany as well.

PHYSICAL AND HUMAN GEOGRAPHY

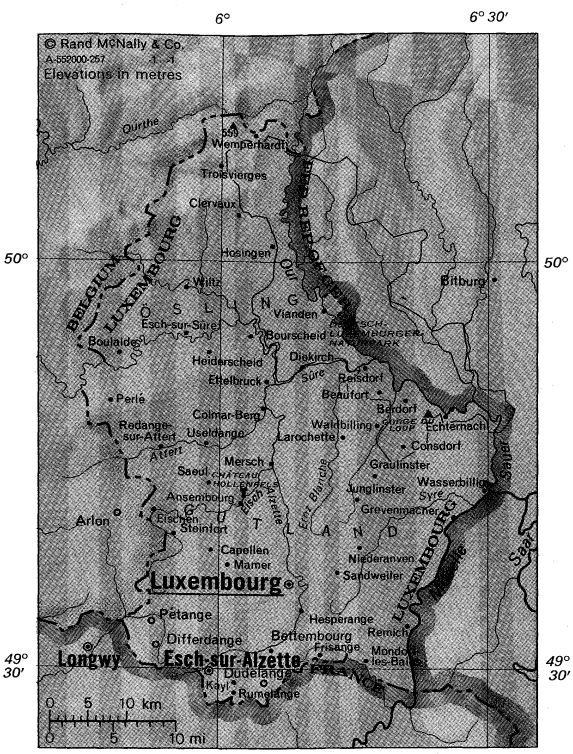
The land. *Relief and soils.* The northern third of Luxembourg, known as the Oesling (Ösling), comprises a corner of the Ardennes Mountains, which lie mainly in southern Belgium. A very level plateau 1,300 to 1,600 feet (400 to 490 metres) in altitude, it is a region of broad vistas and few forests. Its rocky valleys support the ruins of numerous castles, a major attraction for its many tourists. The fertility of the relatively thin mountain soil of the Oesling region was greatly improved with the introduction in the 1890s of a basic-slag fertilizer, which is obtained as a by-product of the grand duchy's steel industry.

Main regions

The heavily wooded south, known as the Gutland, or Bon Pays (German and French: Good Land), has more fertile soil supporting rich pasturelands and vineyards along the Moselle and lower Sûre valleys. At an average altitude of 900 feet, the Gutland contains the capital city, Luxembourg, and industrial centres such as Esch-sur-Alzette. In the east central section lies a great beech forest, the Muellertal (Müllerthal), and the small town of Echternach, to which an annual dance procession, with roots deep in the past, attracts many visitors. Luxembourg's iron mines are located in the south near the French border.

Climate. From both the Oesling and the Gutland, the rivers flow into the Moselle, a tributary of the Rhine. Most farming is mixed, including both gardening and animal raising. The mean temperatures in Luxembourg City range from 37° F (3° C) in January to 67° F (19° C) in July, but in the Oesling both extremes are slightly lower. The Oesling generally has more precipitation than the Gutland, but the greatest amount, about 40 inches (1,000 millimetres), and the least, about 27 inches, fall in the southwest and southeast, respectively.

Settlement patterns. Ethnologists have identified an array of types in human dwellings and in the physical organization of villages throughout Luxembourg. Many date from ancient Celtic and Roman times to the circular villages of the Germanic and Frankish settlements after about AD 400. In addition, many castle villages continue to thrive, centuries after the castles themselves fell to ruin.



LUXEMBOURG

MAP INDEX

Cities and towns

Ansembourg	49·43n	6·02e
Beaufort	49·51n	6·18e
Berdorf	49·50n	6·21e
Bettembourg	49·32n	6·02e
Boulaide	49·54n	5·49e
Bourscheid	49·55n	6·04e
Capellen	49·38n	5·59e
Clervaux	50·04n	6·01e
Colmar-Berg	49·49n	6·05e
Consdorf	49·46n	6·20e
Diekirch	49·53n	6·10e
Differdange	49·32n	5·52e
Dudelange	49·28n	6·05e
Echternach	49·48n	6·26e
Eischen	49·41n	5·52e
Esch-sur-Alzette	49·30n	5·59e
Esch-sur-Sûre	49·55n	5·55e
Ettelbruck	49·52n	6·05e
Frisange	49·32n	6·12e
Graulinster	49·45n	6·18e
Grevenmacher	49·42n	6·20e
Heiderscheid	49·53n	5·54e
Hesperange	49·34n	6·09e
Hosingen	50·01n	6·05e
Junglinster	49·43n	6·15e
Kayl	49·29n	6·02e
Larochette	49·47n	6·15e
Luxembourg	49·36n	6·09e
Mamer	49·38n	6·02e
Mersch	49·46n	6·06e
Mondorf-les-Bains	49·31n	6·16e
Niederanven	49·39n	6·16e
Perlé	49·48n	5·46e
Pétange	49·34n	5·52e
Redange-sur-Attert	49·46n	5·54e
Reisdorf	49·53n	6·15e
Remich	49·33n	6·22e
Rumelange	49·28n	6·02e
Saeul	49·44n	5·59e
Sandweiler	49·37n	6·13e
Steinfort	49·40n	5·55e
Troisvierges	50·08n	6·00e
Useldange	49·47n	5·59e
Vianden	49·57n	6·11e
Waldbilling	49·47n	6·18e
Wasserbillig	49·44n	6·30e
Wemperhardt	50·09n	6·05e
Wiltz	49·48n	5·55e

Physical features and points of interest

Alzette, *river*	49·22n	6·09e
Attert, *river*	49·52n	6·05e
Deutsch-Luxemburger Naturpark, *park*	49·44n	6·12e
Hollenfels, Château, *castle*	49·43n	6·03e
Eisch, *river*	49·45n	6·07e
Ernz Blanche, *river*	49·52n	6·17e
Gutland, *historic region*	49·40n	6·10e
Loup, Gorge du, *gorge*	49·47n	6·23e
Moselle, *river*	49·43n	6·30e
Ösling, *historic region*	49·55n	6·00e
Our, *river*	49·53n	6·18e
Sûre, *river*	49·44n	6·31e
Syre, *river*	49·35n	6·08e

The people. Luxembourg has been one of the historic crossroads of Europe, and myriad peoples have left their bloodlines as well as their cultural imprints on the grand duchy. The Celts, the Ligurians and Romans from Italy, the Belgic peoples known as the Treveri, and especially the Franks were most influential. The dialect spoken by all of Luxembourg's inhabitants is Letzeburgesch, a member of the Germanic language group. Most Luxembourgers speak French and German as well, and all three are des-

National languages

ignated as official national languages. There is a strong sense of national identity among Luxembourgers despite the prevalence of these foreign influences, which can be seen in their country's past as well as its present.

Almost all of Luxembourg's people are Roman Catholic, with a small number of Protestants, mainly Lutherans, and Jews. The several thousand foreigners in the country include many workers in the mining and steel industries or in foreign firms and international organizations located in the capital. Almost two-fifths of the working force is engaged in industry, more than one-half in various service occupations, and the rest in agriculture. There has been a continued trend away from the countryside to urban areas. This has led the government to locate some industries in rural areas, which in turn has altered somewhat the character of rural life, most noticeably in housing and modern amenities of living.

The economy. *Industry.* The production and export of steel have traditionally played major roles in Luxembourg's economy. During the 1970s, however, the demand for steel diminished, causing the steel industry's portion of Luxembourg's domestic product to fall. With the decline of steel production, Luxembourg's economy has become more dependent on the expanding banking sector and on the economic health of U.S. multinational companies operating in the country. The U.S. companies primarily produce tires, chemicals, and fabricated metals.

In 1929 the government began a policy of encouraging the formation in Luxembourg of holding companies, large corporations that control a number of subsidiary companies but are heavily taxed in many areas of the world. The liberal tax climate produced by this policy led many financial corporations as well to maintain offices, often as their European headquarters, in Luxembourg City. The main offices of the European Investment Bank are there, as are the representatives of many banking institutions from around the world who keep in contact with the EEC.

Agriculture. About four-fifths of the arable land in Luxembourg is in farms of less than 125 acres (50 hectares). Grains, root vegetables, and animal fodders are produced in about equal quantities; wine grapes are also grown. Mechanization has replaced animal and human labour in most areas, but the farmer still looks for further help from Common Market agreements. Cattle raising has increased in importance, whereas pig and sheep raising are less important than in former years.

Energy. Luxembourg meets virtually all of its energy needs with imports. Its only domestic source of energy is hydroelectricity.

Transportation. Internal road systems are not extensive but are well maintained, and several highways link Luxembourg with its neighbours. A harbour on the Moselle and a canal connecting it with the system of Rhine waterways provide an avenue for the international movement of goods.

The state has operated the railroads since World War II. They are modern, electrified, and mostly double decked. A major portion of international transportation to and from Luxembourg is by train, and the country is connected with its neighbours by a large number of lines. Findel Airport outside Luxembourg City has become a major European air terminal served by the lines of many countries. Luxair is the national airline.

Administrative and social conditions. *Government.* The grand duchy is a constitutional monarchy with hereditary succession. The constitution, which vests sovereign power in the people and provides guarantees of numerous liberties and freedoms, was revised in 1956 to anticipate the possible transfer of certain sovereign powers to future supranational organs linking several European nations.

Division of government responsibilities

Executive power lies with the grand duke and is executed through a ministerial council, or cabinet. It is responsible to the Chamber of Deputies, whose members are elected by popular vote to five-year terms. Voting by all adult citizens, begun in 1919, is compulsory. In addition, a council of state named by the grand duke functions essentially as a second legislative chamber. It is consulted on all draft legislation and can postpone the enactment of bills. Six official groups, representing agriculture, handicrafts, commerce, civil servants, private employees, and labour, are consulted before the passage of legislation affecting their particular segment of the national working force. Justice is in the hands of magistrates appointed by the grand duke, the final appeal lying with the Superior Court of Justice. In the criminal court of assizes, six magistrates sit as jury as well as judge. The army, under command of the grand duke, is voluntary. Its main functions are to recruit municipal and state police.

Public works, health, and education are among the responsibilities of the communes throughout Luxembourg, which are governed by an elected council. The mayor and aldermen are appointed by the grand duke. These bodies also maintain liaison with the central government and act as its local agents.

Welfare. After World War I a broad system of social security and health services was introduced. Sickness benefits, in which patients pay only a small part of medical costs, as well as birth, family, and unemployment payments, are included in the plans. Existing facilities that require further development include nursing homes for the aged and mental institutions. Housing conditions are generally comparable to those found in other developed European countries. There has been some difficulty, however, in assimilating the many thousands of foreign workers and their families.

Education. Education is compulsory from age six to 15. Primary and secondary education are run by state and local governments and by religious institutions. Great stress is laid on language studies. Instruction is generally given in German in the early years, with an increasing use of French later. There are no four-year universities in the grand duchy, so many young Luxembourgers study abroad.

Cultural life. The major cultural institution of Luxembourg is the Institut Grand-Ducal, which has sections devoted to history, natural sciences, medical sciences, languages and folklore, arts and letters, and moral and political sciences. It functions as an active promoter of the arts, humanities, and culture in general rather than as a conservator. The Musées de l'État have exhibitions on the fine arts, history, and the natural sciences. There is considerable public use of the Bibliothèque Nationale, the Archives de l'État, and the Conservatoire de Musique de la ville de Luxembourg. Luxembourg maintains cultural agreements with several European and other nations that provide the grand duchy with the finest in the musical and theatrical arts. The orchestra of Radio Luxembourg is considered outstanding. There is an extensive market for works of painting and sculpture, both traditional and modern.

A small publishing industry exists, printing literary works in French, German, and Letzeburgesch. The grand duchy's newspapers express diverse political points of view—conservative, liberal, Socialist, and Communist. The government operates all postal and telegraph services and has some control in the private corporation running Radio Luxembourg, which is heard throughout Europe. Some of its multilingual programming is intended for the resident foreign workers. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.* (V.Bi./J.-P.E.)

HISTORY

Ancient and medieval periods. At the time of the Roman conquest, between 57 and 50 BC, the area of what is now Luxembourg was inhabited by a Belgic tribe, the Treveri. After AD 400 Germanic invaders began to penetrate the region, and it became part of the Frankish kingdom of Austrasia and later of Charlemagne's empire. By the Treaty of Verdun (843) the area was included in the central division of the three nations into which that empire was split, but it became an independent entity in 963 under Siegfried, count of Ardennes. Conrad (died 1086), a descendant of Siegfried, took the title of count of Luxembourg about 1060, and his successors enlarged their possessions, chiefly by marriage. Conrad's great-granddaughter, the countess Ermesinde (1196–1247), granted charters of franchise to most of the important towns of the county. Her great-

Treaty of Verdun

grandson Henry IV became German king in 1308 and Holy Roman emperor as Henry VII, and the Luxembourg dynasty was later continued on the imperial throne in the persons of Charles IV, Wenceslaus, and Sigismund (died 1437). In 1354 the emperor Charles IV made the county a duchy. In 1441–44 Sigismund's niece Elizabeth of Görlitz ceded Luxembourg to Philip III the Good of Burgundy.

(K.C.E.)

Habsburg and French rule. With the Netherlands, Luxembourg later passed to the Habsburgs. Spanish domination began in 1555–56, with the abdication of the emperor Charles V. In the revolt of the Low Countries against Philip II, Luxembourg took no part. In 1684 the French under Louis XIV captured Luxembourg but returned it 13 years later to Spain by the Treaty of Rijswijk. In 1713 Luxembourg with Belgium passed from the Spanish to the Austrian Habsburgs. It was occupied by the French republic in 1795 and became a *département* of France under Napoleon.

Personal union with The Netherlands. The Congress of Vienna in 1815 gave Luxembourg as a grand duchy to William I, king of The Netherlands. Districts east of the Moselle, Sûre, and Our were handed over to the king of Prussia. Luxembourg had the legal position of an independent state, in personal union only with The Netherlands, and was included within the German Confederation. William I, however, succeeded in uniting the grand duchy de facto and de jure with his kingdom as its 18th province. When the Belgian provinces revolted against William I, most of the country, with the exception of the town of Luxembourg itself (then garrisoned by Prussians), sided with them. In 1831 the powers divided the country into two parts, of which the larger (now the Belgian province of Luxembourg) went to Belgium and the smaller, in the east, was retained by William I, who, however, refused to accept the decision until 1839. From that year until 1867 the grand duchy was administered separately from The Netherlands. After the dissolution of the German Confederation in 1866 the link that had bound Luxembourg to Germany was broken, but the Prussian garrison remained in the capital. In 1867, by the Treaty of London, this garrison was withdrawn; the neutrality of the grand duchy was guaranteed by the powers and its sovereignty was vested in the House of Nassau.

Independent Luxembourg. On the death of William III of The Netherlands in 1890 without a male heir, the grand duchy passed, in accordance with the Nassau Succession Agreement (1783), to Adolphus, duke of Nassau-Weilburg (died 1905), who was succeeded by his son William (died 1912). In 1914 the neutrality of Luxembourg was violated by Germany. On January 15, 1919, after the duchy's liberation, William's daughter, the grand duchess Marie Adélaïde, who had shown pro-German sympathies, abdicated. Her sister Charlotte took over and exercised all her constitutional powers. Her abdication in favour of Prince Jean was announced in 1964.

Royal successions

Luxembourg was occupied again by the Germans from 1940 to 1944, during World War II. Though it suffered some damage in the fighting of 1944, the industrial capacity of Luxembourg was not seriously impaired, and the country soon regained its prosperity. The 1948 revision of the constitution abolished the "perpetually neutral" status of the country. In 1949 Luxembourg became one of the original members of the North Atlantic Treaty Organization (NATO). The country also holds membership in a number of organizations dedicated to European integration, such as the Benelux Economic Union and the European Economic Community. Despite some economic difficulties in the early 1980s, which prompted the government to impose austerity measures, the country over the past decades has enjoyed economic prosperity and political stability.

For later developments in the political history of Luxembourg, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Nine, Division II, especially Section 923; and Division VI, especially Sections 961 and 963. (K.C.E./Ed.)

Major former principalities

BRABANT

In the early Middle Ages there existed a *pagus* Bracbantensis, whose name was later given to the much larger duchy that included it. The oldest centre of the duchy was in the region of Louvain and Brussels, and the counts of Louvain and Brussels built up the duchy into an important territorial principality.

The history of Brabant is connected with that of the duchy of Lower Lorraine, which in the course of the 11th century was split into a number of small feudal states. The counts of Hainaut, Namur, Luxembourg, and Limburg asserted their independence; and the territory of Liège passed to the bishops of that city. The remnant of the duchy, united since 1100 with the margraviate of Antwerp, was conferred in 1101 by the Holy Roman emperor Henry IV, with the title of duke of Lower Lorraine, upon Henry, count of Limburg. On the Emperor's death his son, Henry V, who in 1106 had rebelled against his father, gave the title to Godfrey I, the Bearded, count of Louvain and Brussels; and for three generations the representatives of the rival houses contested the possession of the ducal dignity in Lower Lorraine. The issue was decided in favour of the House of Louvain by Duke Godfrey III in 1159. His son, Henry I, the Warrior (1183–1235), abandoned the title of duke of Lower Lorraine and assumed in 1190 that of duke of Brabant. Strictly speaking this date marks the beginning of the duchy of Brabant.

Beginning of the Duchy

Godfrey's successors were Henry II (1235–48), Henry III (1248–61), and John I (1261–94). These were all able rulers. Their usual place of residence was Louvain. In 1283, John I bought the duchy of Limburg from Adolf of Berg and secured his acquisition by defeating and slaying his competitor, Henry of Luxembourg, at the Battle of Woeringen (June 5, 1288). From then onward there was a personal union between the two duchies, which conserved their individuality and separate institutions. The Battle of Woeringen was an important episode in the expansionist policy of the dukes of Brabant, who tried to secure control of the territories along the great commercial axis of Brugge-Cologne, a policy that was thwarted by the counts of Flanders in the west and the prince-bishops of Liège in the east.

John II (died October 27, 1312) granted the charter of Cortenberg on September 27, 1312. By it the imposition of grants (*beden*) and taxes was strictly limited and regulated, and its execution was entrusted to a council, appointed by the duke for life, whose duty it was to consider all complaints and to see that the conditions laid down by the charter concerning the administration of justice and finance were not infringed. After the rule of his son, John III (1312–55), the ducal dignity passed to John III's daughter and heiress, Johanna (died 1406). Wenceslas of Luxembourg, her husband, assumed the style of duke of Brabant in right of his wife. On their accession Johanna and Wenceslas granted the famous charter of liberties, known as the Joyeuse Entrée (January 3, 1356). This great constitutional charter gave Brabant an exceptional position among the principalities of the Low Countries and allowed it to play an eminent role in later centuries in the resistance against absolutism.

The Joyeuse Entrée (1356)

Johanna's title was, however, disputed by Louis II of Male, count of Flanders (died 1384), who had married her sister Margaret; and a war broke out in 1356 between Wenceslas, supported by the guilds, and Louis, who upheld the burgher-patrician party in the Brabant cities. After the death of Wenceslas (1383) his widow continued to rule over the two duchies of Luxembourg and Brabant but was obliged to rely on the support of the House of Burgundy in her contests with the turbulent city guilds and with her neighbours, the dukes of Jülich and Gelderland. In 1390 she ceded her duchies to Philip II the Bold of Burgundy, the husband of Johanna's niece Margaret of Flanders (died 1405), keeping only the usufruct to herself. Of the three sons by this marriage, John the Fearless succeeded to Burgundy and Anthony to Brabant on the death of Johanna in 1406. On the extinction of the line of Anthony (1430), the duchy of Brabant became the inheri-

tance of the elder branch of the House of Burgundy in the person of Philip III, the Good, of Burgundy, son of John. From this time onward the history of Brabant is merged with that of other territories of the Netherlands, united by a personal union under the dukes of Burgundy and their successors. The duchy conserved, however, its own central institutions and laws.

Philip III's granddaughter Mary (died 1482), daughter and heiress of Charles I the Bold (died 1477), married the archduke Maximilian and so brought Brabant with the other Burgundian possessions to the House of Habsburg. Brussels, the chief city of Brabant, became under the Habsburg regime the residence of the court and the capital of the Netherlands.

In the course of the Dutch war of independence (1568–1648), the province of Brabant became separated into two portions. In the southern and larger part Spanish rule was maintained, and Brussels continued to be the seat of government. The northern (smaller) part was conquered by the Dutch under Maurice and Frederick Henry of Orange (1597). At the Peace of Münster (1648) this portion, which was known as Generality Land and now forms the Dutch province of North Brabant, was ceded by Philip IV to the United Provinces and was placed under the direct government of the States-General. The southern portion, now divided into the provinces of Antwerp and Brabant, remained under the rule of the Spanish Habsburgs until the death in 1700 of Charles II, the last of his line.

FLANDERS

The county of Flanders was a powerful medieval principality in the southwest of the Low Countries, which occupied territories that are now divided between the French *département* of Nord, the Belgian provinces of East Flanders and West Flanders, and the Dutch province of Zeeland. The name appears in the 8th century and is believed to mean "flooded land." The *pagus* Flandrensis was at first an inconspicuous Frankish district, but a remarkable line of Flemish counts succeeded in erecting a quasi-independent state on the borders between the French and the German kingdoms out of French and German fiefs. As the Treaty of Verdun (843) had made the Scheldt River the dividing line between West and East Francia, the count of Flanders came to be a vassal of the French king for what he held west of the Scheldt (Crown Flanders, or Kroonvlaanderen, the most important part of his countship) and a vassal of the German king for what he held east of it (Aalst, etc.; collectively called Imperial Flanders, or Rijksvlaanderen, as part of the Holy Roman Empire). Linguistically also Flanders was a borderland, including Romance-speaking and Germanic-speaking peoples.

The counts of Flanders in the Middle Ages. In the Carolingian period the *pagus* Flandrensis was the region around Bruges. The rise of the Flemish dynasty began when the count or official administrator of the *pagus,* Baldwin I Iron-Arm, or Ferreus, married Judith, daughter of the Frankish king Charles II the Bald, in 862 and was established as count with control over a number of *pagi* between the North Sea and the Scheldt. His son and successor, Baldwin II (count from 879 to 918), who married the English king Alfred the Great's daughter Aelfthryth, was the founder of the de facto autonomous county and reigned independently over a vast territory along the coast of the North Sea; he fought the Danish raiders (invasion of 879–883).

His son, Arnulf I the Great (918–965), continued the policy of independence and territorial expansion, extending the southern frontiers of Flanders to the Canche River; the main enemy at this time was Normandy. Arnulf's son Baldwin III, however, who had been associated with him as count from 959, died before him (962), and when Arnulf himself died an internal crisis broke out under the rule of his grandson Arnulf II (965–988); but authority was restored under the latter's son Baldwin IV (988–1035), who consolidated his internal position and resumed the expansionist policy. This count looked eastward, coveting imperial territory, and succeeded in conquering lands in the Scheldt estuary and also Valenciennes, an important town on the Scheldt; clearly he thought of the Scheldt as the natural frontier for Flanders on the east. His son Baldwin V (1035–67), an equally able ruler, continued the eastward expansion, adding imperial territories such as Aalst and Termonde, east of the Scheldt, to his dominions. His son and successor Baldwin VI (1067–70), by marrying Richildis, heiress of the neighbouring Hainaut, and thus becoming count of that principality himself, was the first to establish personal union between Hainaut and Flanders; but the succession of his son Arnulf III in 1070 was contested by Robert I the Frisian (died 1093), another son of Baldwin V, who became count in 1071.

Eastward expansion

Robert the Frisian was a powerful ruler and a stern upholder of public order, who also took useful initiatives for the economic development of the county. His son Robert II, who had been associated with him as count from 1087, was succeeded in 1111 by his son Baldwin VII (died 1119), called Baldwin with the Ax because of his repression of violence and crime. On his death the male line of the first dynasty of Flemish counts became extinct.

The next count, designated by Baldwin VII, was his cousin Charles, son of the Danish king Canute IV by Count Robert II's sister Adela. Surnamed the Good for his wise administration. Charles was assassinated at Brugge in 1127 and left no children. The French king Louis VI then intervened and had his candidate accepted as count—namely William Clito, son of Robert Curthose, duke of Normandy (whose mother was a daughter of Baldwin V), and thus an enemy of Henry I of England, against whom the move was directed (1127). This intervention showed that the French crown was emerging from its period of weakness and consequently that Flanders would no longer enjoy undisturbed independence. William Clito, however, soon proved unacceptable to the Flemish population, and on his death (1128) the county went to Thierry, or Diederik (Dietrich), of Alsace, the son of Gertrude, another sister of Robert II. Thierry and his son Philip (1168–91) consolidated the counts' position and won great international prestige for the county, especially since Philip married Elizabeth of Vermandois and thus reigned over territories that stretched far south toward the region of Paris.

Philip's death without issue and the strong policy of Philip II Augustus of France brought Flanders to a perilous position; but the King failed to annex the county itself, to which Baldwin V of Hainaut, a brother-in-law of Philip of Alsace, succeeded as Baldwin VIII (1191–95), though considerable territory was lost in the south (Vermandois and Artois). Then Baldwin VIII's son Baldwin IX (VI of Hainaut), who became Latin emperor of Constantinople in 1204, was succeeded in 1205 by his five-year-old daughter Joan. The French king obtained custody of Joan and made her marry Ferdinand (Ferrand), a younger son of Sancho I of Portugal, in 1212; but in spite of his calculations Ferdinand took sides against him under Flemish pressure and joined the anti-French coalition formed by the emperor Otto IV and John of England. On the defeat of the coalition in the Battle of Bouvines in 1214, Ferdinand was taken to Paris as a prisoner, but Joan was permitted to administer Flanders by the grace of the French crown. She died childless in 1244 and was succeeded in Flanders and Hainaut by her sister Margaret.

Margaret's reign (1244–80) was unfortunately marked by the struggle between the houses of Avesnes and Dampierre—the former being children of her marriage with Bouchard of Avesnes, which was annulled, the latter of her subsequent marriage with William of Dampierre (died 1241). Finally, Louis IX of France adjudged Flanders to Dampierre and Hainaut to Avesnes, but the rivalry between the two houses went on. It became rather dangerous for Flanders when the House of Avesnes obtained the county of Holland, on the northern flank of Flanders, in 1299, just when Margaret's younger son Guy of Dampierre, count of Flanders from 1278 to 1305, had to face a major crisis in Franco-Flemish relations. The French king Philip IV the Fair was in fact trying to bring Flanders directly under royal control and was finding support in the town oligarchies (such as the 39 *échevins* of Gent), who were the Count's political opponents. Guy, however, had the support of the rest of the population, especially in the towns, who felt themselves oppressed

Struggle between the houses of Avesnes and Dampierre

by the oligarchies. Against the French threat, Guy allied himself with Edward I of England (1294 and 1297), but, when Edward had to withdraw his forces from Flanders in order to deal with the Scots, the county was open to invasion. The French overran it in 1300 and took the old count prisoner. The people, however, detested the foreign rule, and on May 18, 1302, the French in Brugge were attacked by surprise and many of them killed (the "Matins of Brugge"). A powerful French army under Jacques de Châtillon then marched on Flanders for revenge, but the Flemings inflicted a disastrous defeat on it in the Battle of Courtrai, or Battle of the Golden Spurs (July 11, 1302). This saved Flanders from French occupation, but the war dragged on until the Treaty of Athis-sur-Orge in 1305. This treaty recognized Flemish independence, but it imposed heavy financial burdens on Flanders and consequently led to considerable loss of territory in the south (Lille, Douai, Orchies) under the Treaty of Pontoise (1312) between Philip IV and Guy's son and successor Robert of Béthune, now Robert III of Flanders (1305–22).

The 14th century brought a new political problem: the big towns, especially Gent (about 50,000 inhabitants), tried to establish communal autonomy against the count and went far toward founding independent city-states. In consequence the counts looked for support to the king of France. Under Louis of Nevers, who succeeded his grandfather Robert of Béthune as count of Flanders in 1322, the outbreak of the Hundred Years' War led to another crisis: the Count took the side of his suzerain, the King of France, but the Flemish towns, under the leadership of Jacob van Artevelde, refused to follow him and took the English side instead, as English wool was indispensable for Flemish industry and prosperity. Louis of Nevers was killed fighting for the French in the Battle of Crécy (1346). His successor Louis of Male (1346–84) established peace and followed an independent policy toward France, but his last years were obscured by the rising of Gent under Philip van Artevelde and by the defeat of the Gent militia by a French royal army in the Battle of Roosebeke (West Rozebeke) in 1382.

The Burgundians and the Habsburgs. Louis of Male left only a daughter, Margaret. Her first husband (1357), Philip of Rouvre, duke of Burgundy, had died in 1361, but in 1369 she had been married again, to Philip the Bold, brother of Charles V of France and, from 1363, himself duke of Burgundy. Thus when Louis died, Philip the Bold succeeded to the county of Flanders. This was the starting point of an evolution that was to result in the unification of the Low Countries under the dukes of Burgundy and the House of Habsburg. At first the Burgundians—Philip II, the Bold (count from 1384 to 1404), John the Fearless (1405–19), and Philip III the Good (1419–67)—acted as French royal princes and strove to play a role in French politics, taking sides against the Armagnacs and with the English. This changed under Charles the Bold (1467–77), who turned against France and attempted to build a Netherlandish-Burgundian realm between France and Germany and to give it strongly centralized institutions. On his death his young daughter Mary, duchess and countess from 1477 to 1482, faced an internal and external crisis and found support by marrying Maximilian of Austria, thus bringing Flanders under the Habsburgs.

Beginning of the unification of the Low Countries

During the religious upheavals of the 16th century, Protestantism won many adherents in Flanders, but the military occupation of the country by the Spaniards undid this development. The Counter-Reformation triumphed, and Flanders remained with the other southern provinces of the Netherlands under Spanish rule in the 17th century (during which it suffered severe territorial losses to the France of Louis XIV) and then under Austrian rule until it disappeared as a political entity in the course of the French Revolutionary Wars. The title count of Flanders (comte de Flandre) remains in use, however, for princes of the Belgian royal family.

Institutions. In the early Middle Ages Flanders shared the institutions of the Frankish kingdom. With the decay of these institutions under the later Carolingians, the country became feudalized and was ruled by powerful princes—wealthy landowners—who sought to expand their territory by force of arms and to safeguard it by building castles. The population, though united under a common rule, was far from homogeneous. In the southernmost area it was mainly Gallo-Roman and Romance-speaking; farther north the Frankish settlement had been denser so that the language was Germanic; and the coastal areas had been settled with people of Saxon and Frisian origin, of a less civilized way of life. The counts effectively united these peoples into one nation. An important new development set in from the 12th century onward, as the counts substituted for the old feudal structure an orderly administration and fiscal organization with salaried officials (*baillis* instead of *châtelains* in the basic administrative circumscriptions), set up a centralized judicial system (using professional judges instructed in Roman law), and began extensive legislation. One of the main starting points of this development was the safeguarding of public order by the counts (*pax comitis*). The role of Philip of Alsace was capital in this field. He gave to a great number of towns charters (*keuren*), the contents of which varied little and which all went back to the charter of Arras of 1157–63. At the same time, but quite distinctly, there developed a powerful communal movement. This led to the establishment in the numerous and wealthy towns of a town government with a considerable measure of independence. The towns were ruled by aldermen called in Flemish *schepenen*, or in French *échevins*, who were administrators and judges at the same time. The constitutional history of Flanders is largely that of the struggle for supremacy between comital and communal authority. Finally the central authority was victorious, and in the 16th century monarchical authority was firmly established. Representative institutions (estates) had grown up around the counts, especially from the 14th century onward; they were largely controlled by the three main towns of Gent, Brugge, and Ypres.

Feudalization

Ecclesiastically, Flanders was divided between three bishoprics, none of which had its seat in an important Flemish city: Thérouanne and Tournai (in the metropolitan province of Reims) and Utrecht (in the metropolitan province of Cologne). In the 14th century Flanders was on the side of Rome against Avignon in the great schism of the papacy.

Economic life. At the outset the Flemish economy was agricultural, but there are very early signs of clothmaking of more than local importance. This was linked with the fact that the country raised large herds of sheep. Around the 12th century Flemish trade and industry became of real international importance. There also took place a crisis in the old manorial organization of agriculture and an expansion of money economy together with the rise of towns as centres of trade and industry. The role of the counts in fostering urban development was important. The cloth industry, which was soon working mainly with English wool and producing quality goods, had its largest centres at Gent and at Ypres. International trade, however, was also very important. Until the 13th century the Flemish merchants conducted their trade abroad, especially at the fairs of Champagne, but later merchants of all nations came to Flanders. The seaport of Brugge became a centre of world commerce. Flanders profited from its geographical situation, being an intermediary between the Mediterranean and the Scandinavian and Baltic countries and also between England and the Rhineland (especially Cologne). Toward the end of the Middle Ages the Flemish economy suffered for two main reasons: (1) the silting of the seaway to Brugge and the competition of Antwerp and (2) the attempts by the towns to conserve artificially an industrial monopoly against normal development in the rest of the country. Even so, Flanders was a rich country whose contribution was of capital importance to the revenues of the Burgundian dukes.

Arts and sciences. Flemish architecture found its outstanding expression in civic buildings, belfries, and town halls, though there were also some fine cathedrals (the Gothic style penetrated early in the 13th century). In literature, satire and the didactic genre (Jacob van Maerlant) flourished particularly. The names of the Flemish painters—the Van Eycks, Rogier van der Weyden, and Hugo van der Goes—belong to world art history, and Jean

d'Okeghem was an outstanding 15th-century composer. For the sciences, the 16th century was brilliant in the humanities, law, and mathematics. (R.C.V.C.)

FRISIA

In prehistoric times the Frisians inhabited the coastal regions from the mouth of the Rhine at Katwijk (north of The Hague) to the mouth of the Ems. Much of the land then was covered by lakes, estuaries, and swamps and exposed to the incursions of the sea, and the inhabitants lived mostly on man-made mounds, the so-called terps. Slowly, as the shifting of the waters allowed, they brought the lower-lying land under cultivation and protected themselves against the sea by building more terps (dikes were not practicable). Most of these were in the modern provinces of Friesland and Groningen.

These Frisians subsequently came under Roman, Frankish, and Carolingian rule. After the partition of Charlemagne's empire, Frisia (from 925) belonged to the German kingdom. The people's subsequent history is that of the three regions into which the country was divided: (1) West Frisia, from the mouth of the Rhine to the Vlie and Lake Flevo; (2) Middle Frisia, from the Vlie to Lauwers; and (3) East Frisia, from Lauwers to the Jade estuary, where the region east of the Ems was penetrated by Frisians as colonists between 600 and 1200—as also was the country north of the Elbe estuary, the so-called North Frisia. The boundary between West and Middle Frisia was soon formed by the floods of Lake Flevo, which developed into the Zuider Zee.

Three regions

West Frisia fell to the counts of Holland, who occupied first the south and the coastal area and, after 1250, the northeastern part. In East Frisia the city of Groningen, with its surrounding country, was ruled by the bishop of Utrecht, and part of the district between the Ems and the Jade became a countship (under the Cirksena family from 1454 to 1744, when it passed to the kingdom of Prussia). Middle Frisia kept itself free of overlords, whether native or foreign, except for short periods, until the end of the Middle Ages.

The Middle Frisians, jealous of their independence, repeatedly expelled the counts of Holland who sought to conquer them. Their country thus remained in "immediate" dependence on the Holy Roman emperor throughout the Middle Ages. Feudalism was never accepted there; hence the adage "Every Frisian is a nobleman." This gave rise to the idea of Frisian liberties that found expression in the so-called Privilege of Freedom, allegedly granted by Charlemagne but in fact a forgery, from the second half of the 13th century. Rejecting foreign interference, these Frisians settled all affairs themselves, keeping legislation, administration, and the judiciary in their own hands and themselves maintaining their dikes, roads, and canals. The Union of the Upstalbeam, by which they consolidated themselves politically, included some of the East Frisians. This degree of liberty is rare in medieval Europe; parallels are found only in the Swiss confederacy and in the administration of Dithmarschen.

The worst external enemies of the Frisians in the Middle Ages were the counts of Holland. The internal security of the country, however, was sometimes threatened by the unruliness of the *hoofdlingen* (rich farmers, who lived in stone-built houses). On the other hand, the monasteries, being rich landowners, played an important role in the country's development.

Enmity of the counts of Holland

As late as 1457 the Holy Roman emperor Frederick III acknowledged the "immediate" status of the Frisians, but in 1498 the emperor Maximilian I presented Frisia to Albert, duke of Saxony, with the title *gubernator in Frisia.* The Saxon dukes, however, could not hold their own; the Frisians, helped by Charles of Gelders, ousted them. Then, in 1524, the Frisian lands between Lauwers and the Zuider Zee fell to the emperor Charles V. Thenceforth they were joined to the Burgundian portion of the Habsburg heritage. This was the first foreign rule to bring the free Frisians a central administration that they accepted.

Friesland adopted the new doctrines at the Reformation and participated in the revolt of the northern Netherlands against Philip II of Spain. It thus became a province of the republic of the Netherlands, as constituted by the Union of Utrecht (1579). A junior branch of the House of Nassau filled the office of stadholder. In 1815 Friesland was absorbed into the Kingdom of The Netherlands.

HOLLAND

The origins of the county of Holland in the Netherlands may be traced to the 9th century, though the name Holland (perhaps derived from Holtland, a property near Leiden) came into use only at the beginning of the 12th. After the destruction of a short-lived Viking rule (885) in which Gerulf, one of the minor Frisian counts, may well have assisted, the coastal area that later grew into the county of Holland came once more under Frankish rule. Gerulf became a vassal of the Carolingians, and he left part of his fief to his son Dirk I (died *c.* 939?), who is considered the ancestor of the House of Holland.

First line of counts. Dirk I ruled as count over Kinhem (Kennemerland), which now is the name of the area between Haarlem and Alkmaar but which then probably comprised the whole coastal belt, extending beyond Leiden and possibly even farther. Only this coastal area was inhabitable; the interior consisted mainly of swamps and lakes and so separated the coastal belt from the diocese of Utrecht. From this nucleus Dirk's successors were to extend their dominion northward, eastward, and southward, favoured by family connections and by their remote situation on the fringe of the Holy Roman Empire, of which the county formed part. Dirk's son Dirk II (died 988) was married to Hildegardis, daughter of Arnulf I of Flanders; one of his sons, Arnulf (died 993), succeeded him as count, while another, Egbert (died 1008), was archbishop of Trier. The aspirations of the counts of Holland in a northerly direction brought them into conflict with the Frisians of the later West Frisia (east of Alkmaar), who were not separated from Middle Frisia by the growing Zuider Zee until the 13th century.

Expansion eastward and southward involved the counts in a series of wars with the bishops of Utrecht and with Germany. A sharp conflict arose during the reign of Arnulf's son Dirk III (died 1039), who after acquiring the Rijn district penetrated as far as the mouth of the Meuse and there built a fortification, Vlaardingen, for the purpose of levying a toll. Adelbold, bishop of Utrecht, though supported by an imperial army under Godfrey of Lorraine, could not oust the count, so that Vlaardingen and its vicinity stayed in the latter's hands. Less fortunate was his son Dirk IV (died 1049), who was forced by the emperor Henry III to abandon his conquests to the diocese of Utrecht. He probably managed to get a firm foothold only in the region of Dordrecht. Dirk V (died 1091), who succeeded his father, Floris I, in 1061, likewise attacked the diocese of Utrecht, supported by his stepfather, Robert the Frisian, but without much result. Dirk V, however, and his successor Floris II (died 1122), who was married to Petronella, sister of the emperor Lothair II, and who was the first of the dynasty to be recognized specifically as count of Holland (in a charter of 1101 from Burchard of Utrecht), did achieve some consolidation of their countship. Under Dirk VI (died 1157) and Floris III (died 1190) further attempts at expansion had little success. Floris III was forced to recognize the Flemish rights on Zeeland (1167); but on the other hand his brother Baldwin, having been elected bishop of Utrecht, gave him the opportunity to extend the influence of Holland into that diocese.

Dirk VII (died 1203) and his brother and successor William I (died 1222) were involved in the contemporary struggle for the German kingship and repeatedly changed sides. William I had difficult times. He had to wrest the countship from his niece Ada, who was married to Louis of Loon; and he subsequently fought several unsuccessful wars with his neighbours. He is important, however, for having started granting privileges to the towns (Middelburg and Utrecht) and for having inaugurated land reclamation and the maintenance of waterways and dikes.

William II, count from 1234 to 1256, the son and successor of Floris IV, extended his grandfather William I's internal policy, promoting municipal development in Mid-

delburg, Delft, Haarlem, and Alkmaar, towns that during the 13th century acquired some importance as trading centres. Externally he cooperated with Brabant. Moreover, in 1247 he was elected German king in opposition to the emperor Frederick II. As king he enjoyed the support of the papacy and of the German towns and thus acquired a position of prominence among the rulers of the Low Countries. He was killed at Hoogwoud in 1256 in battle against the West Frisians, whose independence he was trying to suppress.

The Great Interregnum

With William II's death and the beginning of the Great Interregnum in Germany, the ties between Holland and the empire were loosened considerably; the Netherlands—and Holland especially—went their own way. Under Floris V, who was count from 1256 to 1296, succeeding his father when he was less than two years old, the territorial development of Holland was of major importance. The county was enlarged by the final subjugation of the West Frisians in 1289 and by the acquisition of Waterland (the area between Hoorn and Amsterdam). West Frisia, fertile and protected by castles, made a valuable addition to the revenues of the counts, but efforts to penetrate into Middle Frisia failed. Floris V had further success in the diocese of Utrecht, exploiting the weakness of the bishop-elect, John of Nassau, and forcing several feudal lords in the borderlands to recognize him as their overlord (Amstel, Woerden). When the Gooi area had also been acquired, the county was more or less coextensive with the later provinces of North and South Holland. Careful maintenance of waterways and dikes increased the agrarian importance of the country. Before his successes on his frontiers, Floris V had had to suppress a revolt of the peasants in the north of Holland (1273–74). The grant of a land charter reestablished peace.

It was in Floris V's reign that Holland became involved in Anglo-French quarrels. Edward I of England, seeking allies on the continent, won his support; but Floris in 1296 transferred his alliance to Philip IV of France, probably at the prompting of his cousin John of Avesnes, count of Hainaut, since John and Philip were both hostile to the count of Flanders, Guy of Dampierre, who was also Floris' adversary. This tergiversation was the cause of Floris V's death; discontented nobles, encouraged by the English king, took him prisoner and, when the peasantry prevented them from taking him to England, killed him (1296).

House of Avesnes and Wittelsbachs. John of Avesnes then acted as the governor of Holland until the death of Floris V's son John I (1299), last of the old house of the counts of Holland. Then John of Avesnes himself became count of Holland as John II, when English interference had been repelled and an adverse faction of the nobility had been suppressed. Holland and Hainaut were now united under one count. Interference by the German king Albert I was rejected, and Holland and Hainaut became closely linked to French policy. Holland's influence in Utrecht was strengthened by the election of Guy of Avesnes to the bishopric.

An invasion by the Flemings, who had beaten the French at the Battle of Courtrai (Kortrijk) in 1302, threatened the position of the House of Avesnes in Utrecht, where a rebellion brought the guilds into the government of the town; but help from France enabled John II and, after his death in 1304, his son William III (died 1337) to overcome their difficulties. In 1323 William III made Flanders waive its claims to Zeeland, which thus fell at last to the counts of Holland.

Economic prosperity was promoted by extensive land reclamation. The towns also profited by growing trade. William III played an important role in the affairs of western Europe. He exploited the Anglo-French dissensions leading up to the Hundred Years' War in order to increase his power and prestige. His prominence is shown by the marriages of his daughters: Margaret married the Wittelsbach German king and future emperor Louis IV the Bavarian; Philippa married Edward III of England. Toward the end of his reign William III was head of the anti-French coalition in the Low Countries. In the diocese of Utrecht he acquired great influence; its southern part eventually came under his jurisdiction.

William III's successor, William IV (died 1345), tried to establish his authority in Middle Frisia but fell in the ill-fated Battle of Warns, which put an end to this venture. His death without male issue brought about great disorder and party strife between the factions of Hooks (Hoeken) and Cods (Kabeljauwen) because his sister Margaret (who had been enfeoffed with Holland, Zeeland, and Hainaut by her husband, Louis IV) and her son William V could not agree about their status. The Cods, mainly bourgeois, supported William; the Hooks, mainly nobles, appealed to Margaret. A final settlement was reached only in 1354, under the pressure of a part of the nobility and the rising towns; William V was acknowledged as count of Holland and Zeeland and shortly afterward also of Hainaut. In this way the House of Wittelsbach entered into the Low Countries. As William became insane, his brother Albert ruled as regent from 1358 until William's death in 1389, after which he was himself count until his death in 1404.

The Cods and the Hooks

Albert strove to end the quarrels between Hooks and Cods but was only partly successful. Renewed efforts to conquer Middle Frisia also were without permanent result (1396–1402). In several towns of Holland during Albert's reign, a patrician government was established. His son and successor, William VI, count from 1404 to 1417, sought a *rapprochement* with France and Burgundy. His only daughter, Jacoba, was married first, in 1415, to John of Touraine, second son of Charles VI of France; but John died in April 1417 and William VI himself in May. Though Jacoba's rights had been recognized by the estates of Holland and Zeeland in 1416, difficulties now arose about the succession, and Jacoba had to face opposition not only at home but also from her uncle John of Bavaria. Neither her second nor her third marriages—to her cousin John IV of Brabant and to Humphrey, duke of Gloucester—served to secure her inheritance. (W.J.Al.)

Burgundian dynasty. John of Bavaria, who died in 1425, left his rights in Holland to Philip III the Good, duke of Burgundy, who was also a cousin of Jacoba and of John IV of Brabant; and John IV of Brabant also ceded his rights to Philip. Jacoba resisted as long as possible but under the Treaty of Delft (1428) was forced to hand over the actual government of her provinces to Philip, who finally himself acquired the title of count of Holland, as Philip I, in 1433, because Jacoba, in contravention of the treaty, had made a fourth marriage. Under the Burgundian line of counts, Holland's material prosperity increased continually thanks to the thriving herring fishery and the development of the carrying trade. Under Philip's son Charles the Bold, count of Holland as Charles I from 1467 to 1477, the Hollanders suffered much from excessive taxation; they also felt outraged at Charles's attempt to establish a more centralized government in the Burgundian Netherlands. They took advantage of the collapse of the central government in 1477—when Charles, killed in the Battle of Nancy, was succeeded by his young daughter Mary—to recapture their ancient rights and liberties. A high court of justice was established for Holland, Zeeland, and Friesland and the use of the Dutch language was made official.

Habsburgs and House of Orange. On the sudden death of Mary in 1482, her possessions, including Holland, passed to her infant son Philip the Handsome (Philip II of Holland) under the guardianship of his father, the Austrian archduke Maximilian. Thus the Burgundian dynasty was succeeded by that of Habsburg. The regency of Maximilian was a time of troubles. The Hooks caused new disturbances but were finally overthrown, and the "Bread and Cheese War," a blind revolt of peasants in northern Holland, was proof of the misery caused by civil discords and oppressive taxation.

In 1494 Philip was declared of age. During his effective reign Holland enjoyed rest, and its fisheries benefitted from a commercial treaty with England. On his sudden death in 1506 his elder son, the future Holy Roman emperor Charles V, became count as Charles II.

When Charles abdicated the Netherlands in 1555, his son, the future Philip II of Spain, became Philip III of Holland; in 1559 he appointed William I of Orange as stadholder of Holland, Zeeland, and Utrecht. In 1572 Holland and Zeeland became under William's leadership the centre of

the revolt of the Netherlands against Spanish rule. After the Estates-General, meeting at The Hague, had abjured Philip II in 1581, Holland and Zeeland were about to offer the now vacant countship to William of Orange, but before the matter was finally decided he was assassinated (1584); in 1585 they appointed his son Maurice only to the position of stadholder. Then Robert Dudley, earl of Leicester, exercised some measure of sovereignty in the seven northern provinces of the old Burgundian state but left the country at the close of 1587. Thenceforward Holland was one of the sovereign republics of the seven United Provinces.

Member of the United Provinces. Holland was the most influential of the Dutch republics, thanks to its rapidly increasing wealth (it paid more than half of the federal taxes) and to the concentration of power and talent in its cities, especially Amsterdam. Hence the importance attached to Holland's vote in the assembly of the Estates-General. Yet the authority of Holland did not remain unchallenged. It was in the first place counterbalanced by the influence of the stadholders, the princes of Orange, who, because they held the stadholdership of various provinces at the same time and were generals of the federal army, were sometimes prepared to protect the interests of the other provinces and those of the federal government against Holland's high-handed ways. Sometimes, however, it was not the defense of the federal principle but the furthering of their personal foreign policy that brought the princes of Orange, basing themselves on the majority of the smaller provinces in the Estates-General, into conflict with Holland. Moreover, the very fact of Holland's preponderance and overriding influence gave some unity to the all too loosely constructed federal government. In any case, the chief crises in the internal history of the Dutch republic are to be found in the struggles for supremacy between Holland and the princes of Orange. In the 1780s the conflict led to a state of latent civil war. The "Patriot" opponents of William V of Orange, originally sprung from the *staatsgezinden* (*i.e.,* the anti-Orangist Holland magistracies led by Amsterdam), developed democratic tendencies. When in 1795 the old republic was overthrown, they welcomed the French Revolutionaries as deliverers, not only in order to do away with the stadholdership and the oligarchic system but also in order to realize the unitary state that neither of the old parties had been able to conceive.

Influence of the princes of Orange

The province of Holland during the 17th and 18th centuries was governed by the estates, of which the full title was De Edele Groot Mogende Heeren Staaten van Holland en Westfriesland (the Noble Great Mighty Lords Estates of Holland and West Frisia). After 1608 this assembly consisted of 19 "members" (delegations), one representing the nobility and 18 the towns. A "grand pensionary" acted as chairman. The Estates of Holland sat at The Hague in the months of March, July, September, and November. During the periods of prorogation the continuous supervision of the business and interests of the province was confided to the college of deputed councillors (*gecommitteerde raaden*). The grand pensionary presided over the meetings of the college, which had the general charge of the whole provincial administration, especially finance; the carrying out of the resolutions of the Estates; the maintenance of defenses; and the upholding of the privileges and liberties of the land.

French regime and Kingdom of The Netherlands. In the period of the French Revolutionary and Napoleonic Wars, Holland shared the fate of the rest of The Netherlands. The province was divided into several parts in 1798 but was restored to its old form in 1801. In 1806 it was again divided, now into a northern and a southern part; and from 1810 to 1813 the northern part was united with Utrecht. In 1814, however, when the Kingdom of The Netherlands was established on the fall of Napoleon, the old province was once again restored, though it lost some territory to Brabant. In 1840 it was divided, this time for good, into North Holland and South Holland. (E.H.K.)

LIMBURG

The name of the duchy of Limburg, a historic territory of the Low Countries, survives in that of the two provinces of Limburg, one belonging to the Kingdom of Belgium, the other to the Kingdom of The Netherlands.

When Walram II, count of Arlon, married Judith, daughter of Frederick of Luxembourg, duke of Lower Lorraine (died 1065), he received from his father-in-law some lands lying on either side of the Meuse River and proceeded to build a castle at Limburg. The castle stood on the Vesdre River east of Liège and southwest of Aachen, and from it he took the title of count of Limburg (as Walram I). His son Henry I of Limburg (died 1119?) was invested as duke of Lower Lorraine by the Holy Roman emperor Henry IV in 1101 but was removed from that dignity by Henry V in 1106 and fell into the hands of the new duke, Godfrey the Bearded of Louvain (died 1139). This Godfrey, however, was in turn deposed from the duchy in 1128 in favour of Henry I's son Walram II of Limburg, but after the latter's death (also 1139) the duchy of Lower Lorraine, or Lothier as it was then called, was restored to Godfrey II of Louvain (d. 1142). Walram II's son Henry II of Limburg (died 1167) failed to wrest the title of duke of Lothier from Godfrey III of Louvain and finally gave him his daughter Margaret in marriage (1155), contenting himself with the title of duke of Limburg. This new duchy marked the end of Lower Lorraine, or Lothier, as an entity: Godfrey III's successor in 1190, Henry I of Louvain, styled himself duke of Brabant.

With Henry II's successors in Limburg, namely Henry III (died 1220), Walram III (died 1226), Henry IV (died 1246/47), and Walram IV (died 1279/80), the direct male line of the House of Arlon came to an end. After the death of Walram IV's daughter Irmingard (1282/83), the War of the Limburg Succession broke out, since her cousin Adolf V of Berg disputed the right of her widower, Reinald I of Gelre, to remain in possession of the duchy. Adolf sold his rights to John I of Brabant, and most of the neighbouring states became involved in the struggle: Holland, Jülich, Cleves, and the city of Cologne took John's side; Flanders, Nassau, and Archbishop Siegfried of Cologne took Reinald's. Reinald eventually sold his own rights to Henry III of Luxembourg; but in the Battle of Woeringen, or Worringen (June 5, 1288), John won Limburg for Brabant.

The duchy of Limburg at this time comprised the bailiwicks of Hervé, Montzen, Baelen, Sprimont, and Wallhorn and the counties of Rolduc, Daelhem, and Falkenberg. In personal union with Brabant, it conserved its separate institutions. In 1396 the widowed duchess Joan of Brabant ceded it to her niece Margaret of Flanders, wife of Philip II, the Bold, duke of Burgundy, whose second son Anthony succeeded to it in 1404, two years before he became duke of Brabant also. On the death of Anthony's son Philip (1430), Limburg and Brabant were united with the rest of the Netherlands under Philip III the Good, duke of Burgundy. In 1482 Limburg passed to the House of Habsburg with the Burgundian inheritance. Maastricht was attached to Limburg in 1530.

In the later phases of the war between the United Provinces of the Netherlands and the Spanish Habsburgs, the Dutch conquered the northern part of Limburg; and at the Peace of Westphalia (1648) Daelhem, Falkenberg, and Maastricht were ceded to them to become one of the so-called "Generality lands" under the joint rule of the States-General. Southern Limburg passed with the rest of the Spanish Netherlands to the Austrian Habsburgs under the Peace of Rastatt at the end of the War of the Spanish Succession (1714). Annexed by the French in 1795 and divided between the *départements* of Ourthe and Meuse-Inférieure, Limburg was reconstituted as a province of the new Kingdom of The Netherlands in 1815. This province, however, comprised not only the old duchy but also a piece of Gelderland and the county of Looz.

At the Belgian revolution of 1830 all Limburg except Maastricht took the Belgian side; and so long as William I of The Netherlands refused to recognize the Treaty of London of November 15, 1831, Limburg remained, together with Luxembourg, under Belgian occupation and sent its representatives to Brussels. At last, in 1838, William decided to accept the treaty, and in April 1839 Belgium accordingly had to renounce parts of Limburg and of Luxembourg. The part of Limburg that lay on the

Limburg under Belgian occupation

right bank of the Meuse, together with Maastricht and a number of communes on the left bank (including a large area in the north), then became a sovereign duchy under the rule of the king of The Netherlands; but in order to compensate the German Confederation for the loss of its rights over the part of Luxembourg that Belgium was allowed to retain, the duchy of Limburg (apart from the communes of Maastricht and Venlo) was declared to belong to the confederation, as did the rest of Luxembourg. When the German Confederation was dissolved after the Seven Weeks' War (1866), a conference met in London to consider the Luxembourg-Limburg question; and in 1867 it was decided to sever all connection of the two duchies with Germany. Thereafter eastern Limburg was integrated with The Netherlands, though the province continued to be described officially as a duchy until 1906. Belgian attempts to secure the cession of eastern Limburg after World War I, chiefly on the grounds that the Germans had been able to violate Dutch neutrality to Belgium's disadvantage during the war, were overruled by the Peace Conference of Paris in 1919.

(Ed.)

BIBLIOGRAPHY

General history: J.A. VAN HOUTTE *et al.* (eds.), *Algemene Geschiedenis der Nederlanden,* 12 vol. (1949–58), an extensive history of both the northern and southern Low Countries, with detailed bibliographies; G. EDMUNDSON, *History of Holland* (1922), an excellent general history in English although better for the republican than the modern period; P.J. BLOK, *Geschiedenis van het Nederlandsche Volk,* 4 vol. (1892–99; Eng. trans., *History of the People of the Netherlands,* 5 vol., 1898–1912), a classic history of the northern territories, now outdated, particularly so on the history of the Middle Ages; HENRI PIRENNE, *Histoire de Belgique,* 7 vol. (1900–32, with many revised editions), a brilliantly written history by the famous Belgian historian, covering for the period of the Middle Ages the southern as well as the northern parts of the Low Countries, although emphasis is on the south; I. SCHOEFFER, *A Short History of the Netherlands* (Eng. trans. 1956), a very useful history of the northern Netherlands; MYRON P. GUTMANN, *War and Rural Life in the Early Modern Low Countries* (1980), studies the effect of war on the Basse-Meuse area between 1620 and 1750; HENRI GRIMAL, *Decolonization: the British, French, Dutch, and Belgian Empires, 1919–1963* (1978; trans. of 3rd French ed., 1975).

History of the Netherlands to 1579: S.J. DE LAET, *The Low Countries* (1958), a thorough survey of the Low Countries in prehistoric times; I.H. GOSSES and R.R. POST, *Handboek tot de staatkundige geschiedenis der Nederlanden,* vol. 1, *De Middeleeuwen* (1959), an accurate handbook of the political history, limited, however, to the Dutch-speaking territories; H.P.H. JANSEN, *Middeleeuwse geschiedenis der Nederlanden* (1965), a fairly thorough compendium of the medieval history of the Low Countries, with attention given to the political as well as to the social, economic, and cultural aspects; W. JAPPE ALBERTS and H.P.H. JANSEN, *Welvaart in wording: Sociaal-economische geschiedenis van Nederland van de vroegste tijden tot het einde van de Middeleeuwen* (1964), a social and economic history of the northern Netherlands during the Middle Ages, with emphasis on the economic aspects (the chapters on the Frisian trade and the trade of the 10th to the 13th century, written by J.F. NIERMEYER, belong to the best parts of the book); E. DE MOREAU, *Histoire de l'Église en Belgique,* 5 vol. with a suppl. vol. containing maps (1945–52), an extensive classic handbook of the church history of the southern Netherlands, containing exhaustive bibliographies; R.R. POST, *Kerkgeschiedenis van Nederland in de Middeleeuwen,* 2 vol. (1957), an extensive handbook of the church history of the northern Netherlands; D.P. BLOK, *De Franken (in Nederland), hun optreden in het licht der historie* (1968), a brilliant synthesis of the early history of the northern Netherlands; P.C.J.A. BOELES, *Friesland tot de elfde eeuw: Zijn vóór- en vroege geschiedenis,* 2nd ed. (1951; with an index by H.T. OBREEN, 1962), an authoritative work on the early Frisian history, with an English summary, pp. 559–598; D. JELLEMA, "Frisian Trade in the Dark Ages," *Speculum,* 30:15–36 (1955), a careful survey of Frisian trade and its background in the Merovingian period, drawing especially on archaeological and numismatic evidence; F.L. GANSHOF, *La Flandre sous les premiers comtes* (1948), a solid account of the history of the important principality of Flanders, tracing its origins and ending with the eventful crisis of 1127–28; H.S. LUCAS, *The Low Countries and the Hundred Years' War, 1326–1347* (1929), a very reliable and richly documented account of political events; R. VAUGHAN, *Philip the Bold: The Formation of the Burgundian State* (1962), *John the Fearless: The Growth of Burgundian Power* (1966), and *Philip the Good: The Apogee of Burgundy* (1970), well-documented studies of the Burgundian dukes and the growth of their political power; J. HUIZINGA, *Herfsttij der Middeleeuwen* (1921; Eng. trans., *The Waning of the Middle Ages: A Study in the Forms of Life, Thought and Art in France and the Netherlands in the 14th and 15th Centuries,* 1924), a classical work of cultural historical value; R.R. POST, *The Modern Devotion: Confrontation with Reformation and Humanism* (1968), an exhaustive work, based on old and new sources, with emphasis on the relationship between Modern Devotion and Humanism and the Reformation; G.N. CLARK, "The Birth of the Dutch Republic," in *Studies in History,* pp. 112–144 (1966), written with solid knowledge about the subject; *Geschiedkundige Atlas van Nederland,* 3 vol. of maps and 15 vol. of text (1911–38).

The Netherlands: An extensive source of information is *The Kingdom of The Netherlands* (1971), an official publication. The *Digests of the Kingdom of The Netherlands* are separate publications on the economy, the system of social security, and the constitutional organization in The Netherlands; the Government Information Service updates these publications annually. A complete survey of geographic conditions is published by the CARTOGRAPHIC INSTITUTE (BOOTSMA) under the title *Complete Geography of The Netherlands* (1971). Specialized works on drainage and water control are JOHAN VAN VEEN, *Dredge, Drain, Reclaim: The Art of a Nation,* 5th ed. (1962); see also two publications of the GOVERNMENT INFORMATION SERVICE, *The Zuyder Zee Works* (1967) and *The Delta Project* (1967). On demographic problems the following works in Dutch are available: PIETER VAN DAALEN, *Dringen in Nederland* (1965); and J.P. VAN ROOIJEN, *Het Bevolkingsvraagstuk* (1961). HILBRAND BOSCHMA, *Fauna von Nederland* (1927); and H. HEUKELS, *Schoolflora voor Nederland* (1886, reprinted 1934), are good, if dated, works on Dutch flora. For a description of economic life the following works should be consulted: *The Textile Industry in Europe* (OECD, 1962); FOLKERT DE ROOS, *De Algemene Banken in Nederland* (1968); C.P.A. GEPPAART, *Onevenwichtigheden in de belastingwetgeving* (1968); *Politique des Transports aux Pays-Bas* (Éditions Techniques et Économiques, 1967). Dutch law and politics may be studied from the before-mentioned *Digests* and also from ERNST VAN RAALTE, *Het Nederlandse Parlement* (1958; Eng. trans., *The Parliament of the Kingdom of The Netherlands,* 1959). R.L. MORLAND, *Local Government in The Netherlands* (1958), can be a guide in that particular field. JOHN P. WINDMULLER, *Labor Relations in The Netherlands* (1969) and *Manpower and Social Policy in The Netherlands* (OECD, 1967), are good introductions to Dutch labour politics; educational developments are well covered by *Educational Policy and Planning in The Netherlands* (OECD, 1967). The office for the documentation of art history publishes a *Bibliography of Dutch Art* (The Hague). *Broadcasting in The Netherlands* (NOS, 1971) explains the Dutch system of radio and television broadcasting.

H. DE BUCK, *Bibliografie der Geschiedenis van Nederland* (1968), the standard guide to the historical literature, primarily of the northern Low Countries; P. GEYL, *Geschiedenis van de Nederlandsche Stam,* 2nd ed., 2 vol. (1948–49; rev. ed., 6 vol., 1961–62; partial Eng. trans., *The Revolt of the Netherlands Against Spain, 1555–1609,* 2nd ed., 1958; and *The Netherlands in the Seventeenth Century,* rev. ed., 2 vol., 1961–64), monumental work in which the linguistic–ethnic unity of the Dutch-speaking Low Countries is emphasized; H.H. ROWEN (ed.), *The Low Countries in Early Modern Times* (1972), providing many original sources not available in English; J. HUIZINGA, *Nederland's Beschaving in de Zeventiende Eeuw* (1941; Eng. trans., *Dutch Civilisation in the Seventeenth Century, and Other Essays,* 1941 and 1969), a masterpiece of condensation and illumination; C.R. BOXER, *The Dutch Sea-Borne Empire, 1600–1800* (1965), a vigorously written and well-informed account; C.H. WILSON, *The Dutch Republic and the Civilization of the Seventeenth Century* (1968), by the outstanding English historian of The Netherlands; J. DEN TEX, *Oldenbarnevelt,* 2 vol. (1973), and H.H. ROWEN, *John de Witt, Grand Pensionary of Holland, 1625–1672* (1978), biographies of outstanding statesmen; S. SCHAMA, *Patriots and Liberators: Revolution in the Netherlands, 1780–1813* (1977), a brilliant, innovative study; E.H. KOSSMANN, *De Lage Landen, 1780–1940: Anderhalve eeuw Nederland en België* (1976; *The Low Countries, 1780–1940,* 1978), a penetrating comparative history of Holland and Belgium; A. LIJPHART, *The Politics of Accommodation: Pluralism and Democracy in the Netherlands* (1968), a sociological study of "pillarization," fundamental for the contemporary period.

Belgium: VERNON MALLINSON, *Belgium* (1969), detailed, well documented, with bibliography; FRANK E. HUGGETT, *Modern Belgium* (1969), thorough and discerning; G.M. ASHBY, *Belgium* (1955), a brief, light account; *Fodor's Belgium and Luxembourg* (annual), a useful travel guide; JAN A. GORIS (ed.), *Bel-*

gium (1945), articles by a number of American and Belgian scholars on all aspects of national life; THERESA HENROT, *Belgique* (1958; Eng. trans., 1961), a popular, informative work about daily life; TUDOR EDWARDS, *Belgium and Luxembourg* (1951), interesting on architectural features; DOROTHY LODER, *Belgium and Her People,* rev. ed. (1967), a popular account; F.J. MONKHOUSE, *A Regional Geography of Western Europe,* 2nd ed. (1964), good coverage on Belgium; JULES TARLIER and ALPHONSE WAUTERS, *La Belgique ancienne et moderne: Géographie et histoire des communes belges,* 4 vol. (1859–87, reprinted 1963); C. VERLINDEN, *En Flandre sous Philippe II: Durée de la crise économique* (1952); *Documents pour l'histoire des prix et des salaires en Flandre et en Brabant, XIII^e–XIX^e siècles,* 5 vol. (1959–73); H. VAN DER WEE, *The Growth of the Antwerp Market and the European Economy in the 14th and 16th Centuries,* 3 vol. (1963), until 1619; J. CRAEYBECKX, "Les industries d'exportation dans les villes flamandes au XVII^e siècle, particulièrement à Gand et à Bruges," in *Studi in onore di A. Fanfani,* vol. 4 (1962); H. COPPEJANS-DESMEDT, *Bijdrage tot de studie van de gegoede burgerij te Gent in de XVIII^e eeuw: De vorming van een nieuwe sociaal-economische stand ten tijde van Maria-Theresia* (1952), a study of the well-to-do middle class in Gent during the 18th century and the formation of a new socioeconomic class under Maria Theresa; P. LEBRUN, *L'industrie de la laine à Verviers pendant le XVIII^e et le début du XIX^e siècle* (1948); P. HARSIN, *La révolution liégeoise de 1789* (1954). (*French and Dutch rule*): J. CRAEYBECKX, *Les Statistiques de la fin de l'Empire et le début de la révolution industrielle en Belgique* (1968). (*Belgium and Luxembourg—1831 to the present*): B.S. CHLEPNER, *Cent ans d'histoire sociale en Belgique* (1956). For political history, see publications of the Centre de Recherches et d'Informations socio-politiques (CRISP; 1958–79); A. VERDOODT, *Les problèmes des groupes linguistiques en Belgique* (1977); W. VINCENT, *Het post-unitaire België* (1977); BARBARA EMERSON, *Leopold II of the Belgians: King of Colonialism* (1980), a biography of the acquirer of the Congo Free State; ADRIEN DE MEEUS, *Histoire des Belges* (1956; Eng. trans., 1962), from prehistory to the aftermath of World War II; HENRI PIRENNE, *Histoire de Belgique,* 7 vol. (1900–32), the standard work, and *Les Anciennes Démocraties des Pays-Bas* (1910; Eng. trans., *Early Democracies in the Low Countries,* 1963), on urban society and political conflict in the Middle Ages and the Renaissance; F.G. EYCK, *The Benelux Countries: An Historical Survey* (1959), includes readings in original sources; MARC SCHREIBER, *Belgium* (1945), a concise history with a short bibliography; B.D. GOOCH, *Belgium and the February Revolution* (1963), a study of the impact on Belgium of the 1848 revolution in France; EMILE CAMMAERTS, *The Keystone of Europe* (1939), the foundation and development of independent Belgium to 1939; SHEPARD B. CLOUGH, *A History of the Flemish Movement in Belgium* (1930, reprinted 1968), scholarly; HERMAN VAN DER WEE, *The Growth of the Antwerp Market and the European Economy (Fourteenth–Sixteenth centuries),* 3 vol. (1963); FRANS VAN KALKEN, *Histoire de la Belgique et de son expansion coloniale* (1954); SALLY MARKS, *Innocent Abroad* (1981), a diplomatic history of Belgium at the Paris Peace Conference of 1919.

Luxembourg: ARTHUR HERCHEN, *Manuel d'histoire nationale,* 7th ed. (1966; Eng. trans. of 5th ed. by A.H. COOPER-PRICHARD, *History of the Grand Duchy of Luxemburg,* 1950), provides a wide-ranging account. JOSEPH PETIT, *Luxembourg, Yesterday and To-day,* (published irregularly); AUGUSTE COLLART, *Am Wege zur Unabhängigkeit Luxemburgs* (1938); and ALBERT CALMES, *Au Fil de l'histoire,* 2nd ed. (1968), are also useful. PAUL MARGUE, *A Short History of Luxembourg* (1970, Eng. trans. 1976), is a briefer account; PIERRE MAJERUS, *L'État luxembourgeois,* 3rd ed. (1970; Eng. trans., *The Institutions of the Grand Duchy of Luxembourg,* 1970), explains the constitution. J.A. VAN HOUTTE, *An Economic History of the Low Countries, 800-1800* (1977), is a highly readable account. GEORGES ALS, *Le Luxembourg: Profil Historique Géographique Économique* (1980), provides an overall view of the country. Another topical treatment is presented in MINISTÈRE DE L'ÉDUCATION NATIONALE, *Atlas du Luxembourg* (1971). Other writings include GORDON L. WEIL, *The Benelux Nations: The Politics of Small-Country Democracies* (1970); K.C. EDWARDS, *Luxembourg: The Survival of a Small Nation* (1967); and W.J. TAYLOR-WHITEHEAD, *Luxembourg: Land of Legends* (1951), an interesting discussion of national lore. TUDOR EDWARDS, *Belgium and Luxemburg* (1966), has a useful section on the country. BERYL MILES, *Attic in Luxembourg* (1956); and ROGER PILKINGTON, *Small Boat to Luxembourg* (1967), are personal accounts. JEAN ROUBIER and JOSEPH DELMELLE, *Benelux: Holland, Belgium, Luxembourg* (1958), is a picture book. The ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Belgium-Luxembourg Economic Union* (annual), provides a general guide to the economy.

Luther

The founder of the 16th-century Reformation and of Protestantism, Martin Luther is one of the pivotal figures of Western civilization, as well as of Christianity. By his actions and writings he precipitated a movement that was to yield not only one of the three major theological units of Christianity (along with Roman Catholicism and Eastern Orthodoxy) but was to be a seedbed for social, economic, and political thought. For further treatment of the historical context and consequences of Luther's work, see PROTESTANTISM.

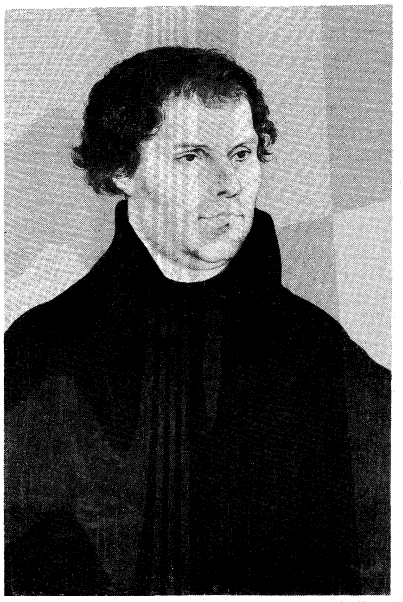
LUTHER AS EDUCATOR AND MONK

Early life and education. Martin Luther was born on November 10, 1483, at Eisleben in Thuringian Saxony (Germany). His parents, Hans and Margarethe Luther, who had moved there from Möhra, soon moved on again to Mansfeld where Hans Luther worked in the copper mines, prospering enough to be able to rent several furnaces and to obtain a position among the councillors of the little town in 1491. Luther's few recollections of childhood that have survived reflect a sombre piety and strict discipline common in that age. His schooling seems to have been unremarkable: the Latin school at Mansfeld, a year at a school in Magdeburg (run by Brethren of the Common Life, a medieval lay group dedicated to Bible study and education) and at Eisenach in his 15th year, where he made valued older friends. In the spring of 1501 he matriculated in arts at the University of Erfurt, one of the oldest and best attended universities in Germany. There he talked long and seriously enough to be nicknamed "the Philosopher," and played the lute. He took the usual arts course and graduated with the B.A. degree in 1502. He took his M.A. in 1505, placing second among 17 candidates. In an age when few students got as far as the master of arts degree, he had fulfilled his parents' hopes. Like many other parents of his time, Hans Luther intended his son to become a lawyer, and he paid cheerfully enough for the expensive textbooks when Martin began legal studies. He was chagrined to learn that his son, without consulting his parents, had decided to enter religion and had sought admission to the house of Augustinian Hermits in Erfurt.

Brother Martin Luther. Evidence on the reason for his decision to enter the religious life is scanty. In his later, not always reliable, *Tischreden* ("Table Talk"), it is related that on July 2, 1505, he was returning from a visit to his parents when he was overtaken by a thunderstorm near the village of Stotternheim and cried out in terror, "Help, St. Anne, and I'll become a monk." In his *De votis monasticis* ("Concerning Monastic Vows," 1521) Luther says "not freely or desirously did I become a monk, but walled around with the terror and agony of sudden death, I vowed a constrained and necessary vow." He sold most of his books, keeping back his Virgil and Plautus, and on July 17, 1505, entered the monastery at Erfurt.

The thunderstorm experience

Augustinian Order at Erfurt. In joining the eremitical order of St. Augustine, Luther had joined an important mendicant order, which by the middle of the 15th century had over 2,000 chapters. As a result of reforms carried through in 1473, the house at Erfurt, to which Luther went, accepted the strict, observant interpretation of the rule. Under Johann von Staupitz, Luther's mentor and



Luther, oil painting by Lucas Cranach, 1526. In the Nationalmuseum, Stockholm.
By courtesy of the Nationalmuseum, Stockholm

vicar general to the order, a revised constitution was made in 1504. Luther made his profession as a monk in September 1506 and was then prepared for ordination. He was ordained priest in April 1507 and his first mass took place at the beginning of May. He had studied a treatise on the canon of the mass by a famous Tübingen Nominalist Gabriel Biel (d. 1495), who, like other "modern" Nominalists, claimed that only named particulars exist and that universal concepts are formed through intuition, and he approached the ceremony with awe. To this occasion his father came with a group of friends, and Luther took this first opportunity to explain personally the imperious nature of his vocation. His father's disgruntled retort, "Did you not read in Scripture that one shall honour one's father and mother?" struck deep into his memory.

Wittenberg University. Luther was selected for advanced theological studies; some of his university teachers were Nominalists of the "modern" way of the English philosopher theologian William of Ockham, whose views undercut the prevailing rationalism of Scholasticism, the school of thought founded in the 11th century in an attempt to reconcile revelation with reason. In 1508 Luther went to the University of Wittenberg (founded 1502), where, though Ockhamism had a foothold, the school of Realism that claimed that universals exist and can be known by reason was championed by scholars such as Martin Pollich. The little town was a contrast to Erfurt, but at least the university was young and forward-looking, and to its comparative remoteness Luther would one day owe his life. The Schlosskirche (called the Church of All Saints) was closely connected with the university, and the elector of Saxony, Frederick III the Wise (1463–1525), lavished generous patronage on both. In March 1509 Luther took the degree of *baccalaureus biblicus* at Wittenberg, returning to Erfurt for his next degree, of *sententiarius,* which involved expounding on the *Sentences*, a medieval theological textbook by Peter Lombard. He had begun his teaching with a course on Aristotle's *Nicomachean Ethics* and now began his career as a theologian with lectures on the *Sentences.* Some of his notes have survived, and if their theology is unexciting there is apparent an acid vehemence at the intrusion of philosophy and above all of Aristotle into the realm of theology.

Influence of Johann von Staupitz

Johann von Staupitz, vicar general of the German Augustinians, was very important in Luther's career as his teacher, friend, and patron. Staupitz seems to have been theologically trained as a Thomist (Realist) and was also influenced by the Augustinian tradition of his order, though his theology shows elements derived from the conflation in the late 15th century of the *devotio moderna* (modern devotion, a term used to describe the spirituality of the Brethren of the Common Life) with German mysticism. His attempt to revive stricter discipline and to unite the observant and conventual Augustinians in Germany led to dispute, and Luther was one of two monks chosen to go to Rome to present the appeal of some dissident houses. He made the journey, the longest of his life, probably late in 1510, and his earnestness was shocked by the levity of the Roman clergy and by the worldliness so evident in high places. The appeal failed, and Luther returned to become a loyal supporter of Staupitz.

Staupitz became interested in his gifted pupil and, perhaps alarmed by his introspectiveness, encouraged him to proceed to his doctorate and to a consequent public teaching career. Luther took his D.Th. on October 19, 1512. The degree was important for Luther, with its implications of public responsibility. He soon took on the duties of a professor in succeeding Staupitz in the chair of biblical theology. This was his lifelong calling, and the exposition of the Bible to his students was a task that called forth his best gifts and energies, one that he sustained until ill health and old age made him relinquish it at the end of his life. In between lectures, in a manner of speaking, he began the Protestant Reformation.

Religious and theological questions. Meanwhile, Luther's own religious and theological difficulties were becoming acute. He had entered into the search for evangelical perfection with characteristic and serious zeal, and sought exactly to fulfill the rule of his order. Nonetheless, he soon found himself in problems difficult for him to understand, struggling against uncertainties and doubts, unhappily bearing a crippling burden of guilt, which neither the sacramental consolations (*e.g.,* the Lord's Supper and penance) of the church nor the wise advice of skilled directors was able to assuage. This distress, which had its centre in his unquiet conscience, brought him into states of anxiety and despair. Nor were his difficulties lessened by the emphases of the Ockhamist theology, which encouraged an extroverted moralism, stressed the human will, and left aspects of uncertainty at the very points where Luther needed most to be reassured. "Temptation" (*Anfechtung*) was to become an important word for Luther's theology, a term that suggests the fight for faith, of which Staupitz could say that such experiences were meat and drink to Martin Luther. These inward, spiritual difficulties were enhanced by theological problems.

Discovery of "the righteousness of God." At the entrance to the world of the thought of St. Paul, Luther was halted—the road blocked by a word that intensified his difficulties to an almost intolerable degree. This was the conception of the "righteousness of God." His sombre childhood piety had made him intensely aware of God's judgment, and as a lecturer in the arts faculty at Wittenberg he had had to expound the Hellenic conception of justice, as he found it in the *Nicomachean Ethics* of Aristotle. Encouraged by the use of *justitia* ("righteousness" or "justice") in the works of several Nominalists, he came to think of God's justice as being primarily the active, punishing severity of God against sinners—*i.e.,* in particular actions. It was for him a final aggravation of his trouble that in Rom. 1:17 it is asserted that the justice of God is revealed in the gospel. Thus, Luther concluded, the divine demand was shown as extending beyond outward obedience to the Law, revealed in the Commandments, to purity of heart, to inward motive and intention, so that grace itself became a demand and an exaction. Such a God could be feared but not loved, could be obeyed out of constraint but never with that happy spontaneity that Luther felt to be of the essence of Christian obedience.

Luther's sense of hypocrisy

Luther's inner conflict. To Luther's sense of failure to obey the Law was added the feeling of hypocrisy, which drove him to the edge of what moral theologians described as "open blasphemy." In 1545, in a celebrated autobiographical fragment that he prefaced to his complete works, he thus described his feelings:

> For however irreproachably I lived as a monk, I felt myself in the presence of God to be a sinner with a most unquiet conscience, nor could I believe that I pleased him with my satisfactions. I did not love, indeed I hated this just God, if

not with open blasphemy, at least with huge murmuring, for I was indignant against him, saying "as if it were really not enough for God that miserable sinners should be eternally lost through original sin, and oppressed with all kind of calamities through the law of the ten commandments, but God must add sorrow on sorrow, and even by the gospel bring his wrath to bear." Thus I raged with a fierce and most agitated conscience, and yet I continued to knock away at Paul in this place, thirsting ardently to know what he really meant.

Thus, the dilemma. Illumination came at last, as in prayer and meditation he pondered the text, examining the connection of the words.

Justification by faith

> At last I began to understand the justice of God as that by which the just man lives by the gift of God, that is to say, by faith, and this sentence, "the justice of God is revealed in the Gospel," to be understood passively, that by which the merciful God justifies by faith, as it is written. "The just man shall live by faith." At this I felt myself to have been born again, and to have entered through open gates into paradise itself.

There has been great controversy about this inner conflict, but it seems certain that there was for Luther just such a crisis as he later described and that it was resolved in the manner he narrates. There has also been argument about the novelty of this discovery. There is in fact a profound difference between the Hellenic conception of distributive justice and the biblical doctrine of the righteousness of God as a divine, saving activity displayed in the field of history and of human experience, and Luther had penetrated deeply into the Pauline vocabulary at this point. The accuracy of Luther's memory about this and, indeed, his integrity have sometimes been impugned, but the verdict of a modern Catholic historian, Joseph Lortz, may stand: that if the discovery were not new, it was at any rate "new for Luther."

Salvation as grace. Had Luther not written this account, it would have been necessary to conjecture something like it to account for the new importance that he gave to justification by faith, a priority it retained in the new theological framework of Protestantism. This became for him the nerve of the gospel, that salvation is to be thought of primarily in terms of grace, and of a divine gift; that God's free, forgiving mercy is displayed in Jesus Christ; that the conscience, forgiven and cleansed, may be at peace, and that the soul, free from the burden of guilt, may serve God with a joyful, spontaneous, creative obedience. In his translation of the Bible Luther came to add "alone" after the word "faith" (*sola fide*) in the verse "For we hold that a man is justified by faith apart from works of law" (Rom. 3:28) because he felt it was demanded by the German language. The word alone or only was retained by the Reformers after him because it seemed to safeguard this important doctrine against such perversions as might seem to make salvation dependent on human achievement or a reward for human merit.

Evaluation of Luther's experience of justification. This experience ought not to be isolated, for Luther speaks of other problems of vocabulary (*e.g.,* the conception of "repentance," *poenitentia*), and it cannot be assumed that this was for him a catastrophic personal experience such as befell St. Augustine, who had a mystical experience of God in the garden at Milan, or the 18th-century founder of Methodism John Wesley, who had a conversion experience at Aldersgate Street, London. About the date of the occurrence there has been much controversy. The publication of Luther's early lectures led naturally to the examination of these firstfruits of the young professor. Though an early view that it must have occurred during the period of Luther's first lectures on the Psalms (1513–15) has been damagingly criticized, Luther's use of the many-sided allegorism of the Middle Ages, which often found three or four levels of meaning in a single text, his concentration on the one historical meaning, and the Christ-centred core of theology of justification have led some scholars to believe that the illumination must have come to him before his lectures on the Letter to the Romans (1515–16).

Something depends on how the discovery itself is assessed: if it was a discovery that justification is a gift, that it is to be taken passively rather than actively, then (as the reference to Augustine's *De spiritu et littera*—"Concerning the Spirit and the Letter"—suggests) Luther was hardly moving beyond the Augustinian framework and it is probably from an early period. If, on the other hand, it was the more mature discovery of the relation of saving faith to the Word of God, then it must be placed later, perhaps in 1518–19. Many scholars now tend in this later direction, and they emphasize how Luther's thinking was stimulated and redirected by the urgent pressure of the church struggle that began in 1517.

The net gain of this chronological discussion has been to demonstrate how important is the whole period of Luther's development from 1509 to 1521, and that his technical vocabulary and the categories of his theology were in movement throughout the whole of this period. Certainly his great courses of lectures on the Psalms (1513–15), on Romans (1515–16), Galatians (1516–17), and Hebrews (1517–18) reveal the growing richness and maturity of his thought.

Luther's growth in theological thought

Luther as preacher and administrator. Meanwhile, his other duties had accumulated. From 1511 he had been preaching in his monastery and in 1514 he became preacher in the parish church. This pulpit became the centre of a long and fruitful preaching ministry wherein Luther expounded the Scriptures profoundly and intelligibly for the common people and related them to the practical context of their lives. Within his order, he had become prior, and, in April 1515, district vicar over 11 other houses. Thus he became involved in a world of practical administration and of pastoral care that gave him valuable experience, standing him in good stead in later years when a large part of his vast correspondence would be concerned with the care of the German churches and the cure of needy souls.

The new University of Wittenberg found it must take sides in an academic crisis that faced the European universities of that day, the tension between an old and a new academic program. Before Luther's advent Martin Pollich, a leading professor at Wittenberg, had shown himself hospitable to Humanist influences, despite his preference for the older Thomism. Now Luther took the lead in inaugurating a new program, involving the displacement of Aristotle and the Scholastic theologians by a biblical humanism that turned to the direct study of the Bible, using as tools the revival of Greek and Hebrew and a renovated Latin and as a dogmatic norm the "old Fathers" (the early Church Fathers, or teachers) and especially St. Augustine. Such a program Luther planned with the help of his senior colleague, Karlstadt, and his young friend Philipp Melanchthon. In February 1517 he penned a series of theses against the Scholastic theologians, which he offered to defend at other universities. Though this attempt to export the Wittenberg program met with no success he could write in May that the battle was won at least in Wittenberg—"our theology, and that of St. Augustine reign." But if his theses remained dormant, a very different fate awaited those that he wrote later in that same year. He could hardly have thought that these would fire a train that would explode the Western Christian world.

LUTHER AS REFORMER

The indulgence controversy. *The nature of indulgences.* The nature and scope of indulgences had been more and more defined during the later Middle Ages, but there was still an element of that dogmatic uncertainty that has been called a theological weakness of the age. Indulgences were the commutation for money of part of the temporal penalty due for sin, of the practical satisfaction that was a part of the sacrament of penance, which also required contrition on the part of the penitent and absolution from a priest. They were granted on papal authority and made available through accredited agents. At no time at all did they even imply that divine forgiveness could be bought or sold, or that they availed for those who were impenitent or unconfessed. But during the Middle Ages, as papal financial difficulties grew more complicated, they were resorted to so often that the financial house of Fugger of Augsburg had to superintend the sacred negotiations involved in them.

The way was open for further misunderstanding when in

1476 Pope Sixtus IV extended their authority to souls in purgatory. The appeal to cupidity and fear, the pomp and circumstance with which these indulgences were attended, the often outrageous statements of some indulgence sellers were a matter of complaint. Luther himself had frequently preached against these abuses, for his patron, the elector Frederick, had amassed a great collection of relics in the castle church at Wittenberg, to which indulgences were attached. But the immediate cause of Luther's public protest was an indulgence that Frederick had prohibited from his lands, though it was available in nearby territory. This was a jubilee indulgence, offering special privileges, the ostensible purpose of which was the rebuilding of St. Peter's basilica in Rome. By a secret arrangement, half of the German proceeds were to go to the young Albert, archbishop of Mainz, who was deeply in debt owing to his rapid promotion to and payment for a number of high ecclesiastical offices.

The immediate cause of the indulgence controversy

The Ninety-five Theses. Of this Luther knew nothing until some time afterward. For him, the provocation lay in the extravagant claims of an old, tried hand at this kind of thing, the Dominican salesman of indulgences Johann Tetzel. With these claims in mind, Luther drew up the Ninety-five Theses, "for the purpose of eliciting truth," and fastened them on the door of All Saints Church, Wittenberg, on October 31, 1517, the eve of All Saints' Day and of the great exposure of relics there. These were tentative opinions, about some of which Luther himself was not committed. They did not deny the papal prerogative in this matter, though by implication they criticized papal policy; still less did they attack such established teaching as the doctrine of purgatory. But they did stress the spiritual, inward character of the Christian religion, and the first thesis, which claimed that repentance involved the whole life of the Christian man, and the 62nd, that the true treasure of the church was the most holy gospel of the glory and the grace of God, showed the author's intention. The closing section attacked the false peace, that "security," which as a young lecturer Luther had so often attacked, of those who thought of divine grace as something cheaply acquired and who refused to recognize that to be a Christian involved embracing the cross and entering heaven through tribulation. Luther sent copies of the theses to the archbishop of Mainz and to his bishop. And here the invention of printing intervened. Copies were circulated far and wide, so that what might have been a mere local issue became a public controversy discussed in ever widening circles.

Reaction to the Ninety-five Theses. The archbishop of Mainz, alarmed and annoyed, forwarded the documents to Rome in December 1517, with the request that Luther be inhibited, at the same time reprimanding the indulgence sellers for their extravagance. At the time, it seemed to many that this was simply another squabble between the Dominicans and the Augustinians. Colour was given to this belief by the counter-theses prepared by a theologian, Konrad Wimpina, that Tetzel had defended before a Dominican audience at Frankfurt at the end of January 1518. When copies of these reached Wittenberg in March they were publicly burned by excited students. At Rome the pope merely instructed Gabriel della Volta, the vicar general of the Augustinians, to deal with the recalcitrant monk through the usual channels, in this case through Staupitz. Luther himself prepared a long Latin manuscript with explanations of his Ninety-five Theses, publication of which was held up until the autumn of 1518; it is a document of some theological importance, and shows how far from superficial Luther's original protest had been. Meanwhile, the chapter of the German Augustinians was held at Heidelberg, April 25, 1518. Luther was relieved of his extra duties as district vicar, in the circumstances a great relief and intended as such. He found great comfort in the support of his friends, and was himself in great form, winning over two young men, Martin Bucer, a Dominican, and Theodor Bibliander.

Counter-theses and student reaction

"Theology of the cross"

At this period Luther's theology was most especially a "theology of the cross"; *i.e.,* a theology that stressed the revelation of Christ on the cross. According to Luther, the "theology of the cross" seems foolishness to the wisdom of the world and is opposed to the natural theology of divine power and majesty, which he attacked as a Scholastic "theology of glory." Important for him at this time was the inward religion preached by the 14th-century German mystic Johann Tauler and a short 14th-century mystical tract, the *Theologia Germanica,* that he himself edited and published (1516–18). In these months, therefore, he lay great stress on the need for the Christian to share the cross of Christ, in suffering and in temptation. Though these stresses were to recede into the background of Luther's developing theology, they were to remain important for the radical Reformation, for which the *Theologia Germanica* would be an important and seminal document.

Involvement of Johann Eck. During Luther's absence, and perhaps catastrophically, his senior colleague, Karlstadt, had taken action that was greatly to widen the scope and publicity of the controversy. The scholar Johann Eck (1486–1543) of Ingolstadt, a man of some learning, and with a zest for disputation, with whom Luther was already in friendly contact through a common friend, became involved in the controversy. He had written some observations on the Ninety-five Theses for his friend, the bishop of Eichstädt, and these manuscript observations, the so-called Obelisks, reached Wittenberg shortly before Luther went off to his chapter at Heidelberg; Luther himself replied with a few "Asterisks," but Karlstadt, concerned to defend the Wittenberg program, sprang into the fray with 379 theses, adding another 26 before publication. In some of these Eck was impugned. The Dominicans continued to press for Luther's impeachment, and proceedings against him for heresy began to move slowly in Rome. Luther himself did not improve matters by publishing a bold sermon on the power of excommunication that made it clear that here was not a man who would accept unquestioned whatever might be decided by the pope in terms of some undefined plenitude of power.

The Augsburg interview, 1518. *Luther before Cajetan.* A papal citation summoning Luther to Rome was sent to the cardinal Cajetan (1468–1534), a renowned Thomist, at Augsburg. But at this perilous moment politics fatefully intervened, and the period during which the Luther affair might have been swiftly, drastically disposed of without wider disaster to the church was eroded by considerations of policy. The elector Frederick, as one of the seven prince electors of the Holy Roman Empire, was most important to the pope, in view of the imminent choice of a new emperor, and the pope could not afford to antagonize him. The result was that Luther was bidden to a personal interview with Cajetan at Augsburg. He arrived there on October 7 with an imperial safe-conduct. The discussion had moved from indulgences to the discussion of the relation between faith and sacramental grace (the unmerited gifts of God in such acts as Baptism and the Lord's Supper), when an argument developed between the two theologians about the meaning of the "treasure" of merits that the papal definition of Sixtus IV said that Christ had acquired, and the incensed cardinal dismissed Luther from his presence, telling him to stay away unless he would unconditionally recant.

Dispute over meaning of merits

Luther's flight from Augsburg. While Luther waited uneasily, the Saxon councillors reported rumours that he would be taken in chains to Rome. Eventually, bundled through a postern by his friends, he fled the city. Now he wrote an appeal from the pope to a general (or ecumenical) council and a full defense of his actions to his prince. Cajetan, meanwhile, lost no time in denouncing Luther to Frederick, who was in something of a dilemma, though it counted much for Luther that he had the admiring friendship of the elector's secretary, the Humanist Georg Spalatin. At this time, too, the Wittenberg theological faculty addressed the prince on Luther's behalf, pointing out that the fate of the university and its reputation would be involved in Luther's disgrace. At one moment, it seemed that Luther might have to depart, perhaps for France or Bohemia. There then appeared Karl von Miltitz, a papal diplomat, who applied "stick and carrot" tactics to the elector, dangling before him at one moment threats against Luther and at the next the signal compliment of the golden rose, symbol of high papal honour and recogni-

tion. The diplomat promised more than he could possibly perform, and after an interview with him at Altenburg, in January 1519, Luther sensed this and came to distrust him. A papal definition about indulgences, issued at Cajetan's request, seemed to show that Luther had indeed put his finger on some fatal ambiguities.

The Leipzig disputation, 1519. *Debate between Luther and Eck.* At Augsburg Luther had been in touch with Eck and arrangements were made for a public disputation at Leipzig in the summer. This was to be in the first place a debate between Eck and Karlstadt, though Luther was Eck's ultimate objective, but the hostility of George, duke of Saxony (the elector Frederick's first cousin), toward the Reformer raised difficulties about Luther's participation. Eventually it was arranged that Eck should debate with the two Wittenberg theologians in turn, in the castle of the Pleissenburg, Leipzig, at the end of July. There was a preliminary pamphlet skirmish. The issue between Eck and Karlstadt was the Augustinian doctrine of grace and free will, and Karlstadt wished to meddle neither with indulgences nor with papal authority. Among the preliminary matters, the origin of the papal power was raised and so Luther turned to a study of church history and Canon Law in the fateful weeks before the debate. A large contingent from Wittenberg attended, and in the presence of theologians from both universities, Duke George and notables of church and state, the debate began. Eck showed some skill in manoeuvring Luther into a position in which he cast doubt on the authority of the great General Council of Constance (1414–18), and also defended some of the propositions of Jan Hus, a Bohemian Reformer who had been declared a heretic at Constance and burned to death at the stake. Leipzig was a part of Germany with a strong feeling against Bohemia, and the admission was received as damaging, giving ground for Eck's loud boast that the disputation had been his personal triumph. Luther, who had earlier said of the debate that it had not begun in God's name and would not end in his name, left Leipzig somewhat shaken and disturbed by Eck's verbal manoeuvring.

Luther's relationship to Jan Hus

Luther's questioning of authority. Eck was able to go off to Rome with new prestige to give sharpness to the process of Luther's official condemnation. Luther had now to examine the further implications of his actions to date, in relation to the authority of the church, of councils, and of Scripture; his correspondence shows that he was reaching something like a crisis in his attitude to papal authority. There had been a small pamphlet war after the disputation that made it plain that there was strong support for Luther among the Humanists in Germany and Switzerland. Luther himself became involved in controversy with diverse theologians of Leipzig, and if he now wrote in the vernacular with increasing power and violence, his polemical writings reveal also his deep perceptions of the issues between himself and contemporary theology. Two Catholic universities, strongholds of tradition, Cologne and Louvain, next condemned Luther's teaching. But polemic was not Luther's main concern, and his *Sermon von den guten Werken* ("Sermon on Good Works"), issued in June 1520, is an important exposition of the ethical implications of justification by faith. As a tract it deserves to be associated with Luther's more famous tract on Christian liberty issued in the next months. On June 15, 1520, there appeared the papal bull (a decree issued under the papal seal) *Exsurge Domine* or "Lord, cast out," against 41 articles of Luther's teaching, followed by the burning of Luther's writings in Rome. Eck and the Humanist diplomat and cardinal Girolamo Aleandro (1480–1542) were entrusted with the task of taking the bull to the cities of Germany.

The Reformation treatises of 1520. Eck and Aleandro were alarmed to discover how swiftly German opinion had moved to Luther's side. In contrast to his treatment the year before, Eck had to seek refuge in Leipzig from physical violence. Aleandro did what he could in agitated correspondence to shock the Curia (papal administrative bureaucracy) into realizing the grave danger facing the church in Germany. Luther's friends, aware of how precarious his position was, sought to moderate his violence, but he now moved well beyond their horizon. In Luther's own opinion of himself, he was far too temperate in view of all the ecclesiastical hypocrisy and offenses. The result was the defiant tracts of the summer of 1520. The first, the real manifesto, was his *An den christlichen Adel deutscher Nation* ("Address to the Christian Nobility of the German Nation"), addressed to the rulers of Germany, princes, knights, cities, under the young emperor Charles V. It argued that in the crisis, when the spiritual arm had refused to take in hand the amendment of the church and the often expressed grievances of the German people against Rome (*i.e.,* the papacy), it was necessary for the secular arm to intervene and call a reforming council. The document was ill arranged and tailed off, but it found deep response among sections of the nation, and in the next months Luther was carried along with the tide of national resentment against Rome.

"Address to the Christian Nobility"

His second treatise, *De captivitate Babylonica ecclesiae praeludium* ("A Prelude Concerning the Babylonian Captivity of the Church"), intended for clergy and scholars, was an act of ecclesiastical revolution. It inevitably estranged many moderate Humanists, for it reduced to only three (Baptism, the Lord's Supper, and penance) the seven sacraments of the church, denied mass and attacked transubstantiation (the doctrine that the substance of the bread and wine is changed into the body and blood of Christ in the sacrament of the Lord's Supper), made vehement charges against papal authority, and asserted the supremacy of Holy Scripture and the rights of individual conscience. The third work, dedicated to the pope, was, as a still, small voice after the uproar, a minor classic of edification, *Von der Freiheit eines Christenmenschen* ("Of the Freedom of a Christian Man"), which made clear the ethical implications of justification by faith, and showed that his thought and his public actions were connected by a coherent theological core. On December 10, 1520, the students lit a bonfire before the Elster Gate in Wittenberg, and as they fed the works of the canonists to the flames Luther added the papal bull (*Exsurge Domine*) against himself with suitable imprecation—"Because you have corrupted God's truth, may God destroy you in this fire."

In January 1521 the pope issued the bull of formal excommunication (*Decet Romanum Pontificem*), though it was some months before the condemnation was received throughout Germany. Meanwhile, the imperial Diet was meeting at Worms, and there was a good deal of lobbying for and against Luther. In the end, Frederick the Wise obtained a promise from the emperor that Luther should not be condemned unheard and should be summoned to appear before the Diet. This enraged Aleandro, who asserted that the papal condemnation was sufficient and that the secular arm had only to carry out its orders. It also alarmed Luther's friends, who did what they could to dissuade him. Luther was firm in his determination to go, and began the journey in April 1521, undeterred by the news, on the way, that the emperor had ordered his books to be burned. What was meant to be the safe custody of a heretic turned out to be something like a triumphal procession, and when Luther entered Worms on April 16 he was attended by a cavalcade of German knights and the streets were so thronged as to enrage his enemies.

Excommunication

The Diet and Edict of Worms, 1521. *Luther's defiance of the Diet.* In the early evening of April 17, 1521, Luther appeared before the notables of church and state and faced the young emperor Charles V, whom he found cold and hostile. A pile of writings lay before him, but when he was formally asked whether he acknowledged them, his legal adviser insisted that the titles be read. In view of the gravity of recantation, Luther asked for time to think, a request that may have taken his enemies off guard. A day's respite was granted, and the following afternoon, in a larger hall, and before an even more crowded assembly, Luther reappeared. This time he could not be prevented from making a long speech. He distinguished between his writings: for the works of edification he need not and ought not to recant, for the violence of his polemic he would apologize, but for the rest he could not recant; and, as he went on to explain why, the demand was brusquely made for a plain, simple answer. This he now

gave in words of unyielding defiance. He would recant if convinced of his error either by Scripture or by evident reason. Otherwise he could not go against his conscience, which was bound by the Word of God. Though evidence is now tilted against the authenticity of the famous conclusion, "Here I stand. I can do no other," it at least registers the authentic note of Luther's reply in a moment that captured the imagination of Europe. There was a moment of confusion with Eck and Luther shouting, and then the emperor cut short the proceedings. Luther strode through his thronging enemies to his friends, his arm raised in a gesture of relief and triumph.

"Here I stand. I can do no other."

There followed a diplomatic flurry. It was evident that Luther had powerful friends; there was some sabre rattling from the knights and the peasant emblem appeared in the streets. There is evidence to support Luther's boast that had he wished he could have started such a game that the emperor's life would not have been safe. The radical Reformer and social revolutionary, Thomas Müntzer, later asserted that had Luther recanted the angry knights would have killed him. At any rate, Luther was now given what he had long asked for in vain, something like a real hearing before reasonably impartial judges, while he was kindly handled by the archbishop of Trier. But he could not now make even minor concessions, and the discussions broke down on the fallibility of councils. He was formally dismissed and departed under his safe-conduct.

Despite his spectacular moral triumph, Luther's enemies, nonetheless, achieved something important at this point when a rump Diet passed the Edict of Worms. It declared Luther to be an outlaw whose writings were proscribed. The edict was to shadow him and fetter his movements all his days. It meant also that his prince must, for a time at least, walk delicately and could not publicly support his protégé. The result was the pretended kidnapping of Luther who was lodged secretly in the romantic castle of the Wartburg, near Eisenach.

Luther at the Wartburg. In this aerie among the trees Luther remained until March 1522. Known as Junker Georg, or Knight George, he dressed as a layman, grew a beard, and put on weight. The lack of exercise and the unwontedly rich diet brought on physical distress, whereas his mind, flung back on itself after months of crisis, knew intense reaction in a period of acute depression of the kind that Luther ranked high among temptations. But he was far from idle. He finished a beautiful exposition of the Magnificat (the song of Mary, the mother of Christ, in the liturgy) and prepared an edition of sermons on the Epistles and Gospels at mass, which he thought was perhaps his best writing. Although away from books, he wrote his ablest controversial piece, *Rationis Latomianae pro Incendiariis Lovaniensis Scholae sophistis redditae Lutheriana confutatio* ("Refutation of the argument of Latomus"—who was a member of the theological faculty of the University of Louvain), containing a luminous exposition of justification. Most important of all, he began to translate the New Testament from the original Greek into German. He did not believe that such work should be left to one mind, and soon enlisted his colleagues, notably Melanchthon, in the enterprise. But Luther's was the controlling genius, and the resulting New Testament (published in September 1522), like the Old Testament, translated from the Hebrew, which followed later (1534), was a monumental work, which had deep and lasting influence on the language, life, and religion of the German people. He had now to deal with some of the practical implications of his revolt. Private masses, celibacy of clergy, religious vows were no theoretical questions, but were themselves entangled in a network of legal, financial, and liturgical affairs. He wrote about these things forthrightly, and Spalatin tried in vain to hold up their publication, for in Wittenberg there were growing difficulties, and the prince, the university, and the cathedral chapter were all, for various reasons, anxious to go slowly.

Influence of Luther's translation of the Bible on the German language

Commotion in Wittenberg, 1521–22. *Radical reform.* There was a lively section of the town and of the university, however, that was determined to force the pace, and there were violent scenes in the streets and churches early in October 1521. Yet Luther, on a secret visit to his friends early in December, was not alarmed, and it was his influence that led the Augustinians to decide, in the new year, that those of them who wished might return to the world. Two radical leaders now appeared, the incorrigible troublemaker Karlstadt and Gabriel Zwilling, an ebullient spellbinder from the Augustinians. When Karlstadt announced his betrothal to a girl of 16, and at Christmas administered Communion in both kinds (bread and wine) while dressed as a layman, attacked images in a violent tract and in innumerable theses denounced vows and masses, and demanded a vernacular liturgy, it was evident that here was a program that in timing and method differed from Luther's. Moreover, its appeal to Scripture was legalistic and made matters of necessity things that for Luther lay within the option of Christian liberty. In the new year, the town council issued a notable and pioneering ordinance regulating religion, public morals, and poor relief, a document that owes much to Luther's teaching and perhaps something to the initiative of Karlstadt. At the end of 1521 confusion was increased by the arrival of the so-called Zwickau prophets, radicals on the run from the town of Zwickau, who spoke impressively of revelations given them through dreams and visions, claiming that the end of the world was near and that all priests should be killed. A flustered and outmanoeuvred Melanchthon wrote urgently for advice to Luther, who sent wise and calm counsel.

Restoration of balanced reform. In the next months the situation worsened and in March 1522 Luther returned to Wittenberg, explaining the reason for his disobedience to instructions in a justly famous letter to his prince. Then, deliberately habited as an Augustinian monk once more, he took charge of his town pulpit and in a powerful series of sermons redressed the balance of reform. In these important utterances, the difference between Luther's conservatism and the radical pattern of reform is made plain. Luther deplored the use of violence, for the Word of God must be the agent of reform. He believed that revolt could not take place without destruction and the shedding of innocent blood; that the real idols are in the hearts of men and if their hearts are changed the images on church walls must fall into disuse. Moreover, the pace of reform must take into account the unconverted, weaker brethren. From this time onward Luther fought a war on two fronts, against the Catholics and against those whom he lumped together as *Schwärmer* ("fanatics"). One result of the Wittenberg crisis was to slow down the practical reforms, and though Luther introduced a reformed rite (*Formula Missae* or "Formula of the Mass," 1523) it was not until 1526 that he provided a vernacular liturgy (*Deutsche Messe,* or "German Mass"). Throughout Germany the evangelical movement continued to grow, and it was apparent that the Edict of Worms would not be everywhere enforced. A Diet at Nürnberg, 1522–23, refused to suppress the evangelical preachers and demanded a reforming, national council; though Catholic pressure was stronger in the following year, the Diet again pressed for a council and would consent only to the enforcement of the edict "as far as possible."

Luther's dislike for violence

The Peasants' War. *Activities of the radical Reformers.* On his journeys to and from Worms Luther had been dismayed by the evident social and political unrest. In the next months he wrote open letters, warning the rulers of Saxony and the councils of such cities as Strassburg of the danger that the new radical teaching would provoke revolution. In 1523 he made his own views of secular government plain in an important treatise *Von weltlicher Obrigkeit* ("Of Earthly Government"), in which he firmly asserted the duty of a Christian prince and the place of secular government within God's ordinances for mankind; he distinguished between the two realms of spiritual and of temporal government, through which the one rule of God is administered, and stressed the duty of civil obedience and the sinfulness of rebellion against lawful authority.

The two realms theory

In Saxony the radical teachers posed a problem for their untheological rulers. In Orlamünde, after having been rebuked at Wittenberg, Karlstadt had converted the community to his own brand of mystical quietism. Luther made a preaching tour of the area at the request of his

prince, and was greeted with hostility and ridicule. Luther himself denounced such social evils as usury, but in Eisenach the fiery preacher Jakob Strauss conducted a violent campaign against usury and tithes. Most formidable of all, in the little town of Allstedt, Thomas Müntzer, an unruly genius, combined his own ingenious liturgical reforms with a program of holy war. Himself a former "Martinian" (or follower of Martin Luther), he not only shared Karlstadt's enthusiasm for the mystics but added an explosive element (perhaps influenced by Hussite teaching) that gave point to Luther's worst fears. Müntzer threatened revolution and claimed that God would rid the world of its shame. Luther's warnings and events themselves forced the rulers to take action, and in the summer of 1524 Müntzer fled and Karlstadt was exiled. Müntzer wrote in a pamphlet that Luther was nothing more than a shameless monk, "whoring and drinking," and called him Dr. Liar. Karlstadt also wrote a series of tracts against his former comrades, denouncing, among other things, the corporeal presence in the Eucharist. Luther replied in a devastating and profound treatise, *Wider die himmlischen Propheten, von den Bildern und Sakrament* ("Against the Heavenly Prophets in the Matter of Images and Sacraments"). He claimed that the radical Reformers sought glory and honour, not the salvation of men's souls.

Luther's response to the Peasants' War. In the summer of 1524 the Peasants' War had broken out in the Black Forest area. Their program was variously motivated. Their demands were for concrete medieval liberties connected with the game and forest laws or with tithes. Some of them drew on Catholic teaching, others on the theology of Zwingli and of Luther, who had set an example of successful defiance of authority, had been no respecter of dignities, and whose teachings about Christian liberty and a priesthood in which all believers shared were plainer than his subtle distinctions between two kingdoms. Thus, both where he was understood and where he was misunderstood, Luther's influence in the Peasants' War has to be taken into account. Some of the moderate peasants included Luther among possible arbitrators. He himself published in May 1525 the *Ermahnung zum Frieden* ("Exhortation for Freedom"), an analysis of the "12 articles" of the Swabian peasants, sympathizing with just grievances, criticizing the princes, but repudiating the notion of a so-called Christian rebellion: "My dear friends, Christians are not so numerous that they can get together in a mob." Luther also claimed that the worldly kingdom cannot exist without inequality of persons.

Luther's view of rebellion

In the spring of 1525 the Thuringian peasants rose, with Thomas Müntzer among their leaders, and at first seemed likely to carry all before them. Faced with imminent political chaos, Luther wrote a brutal, virulent broadsheet, *Wider die räuberischen und mörderischen Rotten der andern Bauern* ("Against the Murdering and Thieving Hordes of Peasants"). The writing was less violent than Müntzer's hysterical manifestos, but it was bad enough. It appeared, however, as an appendix to his moderate tract about the "12 articles." Moreover, words written at the height of the peasant success read very differently after their collapse at the Battle of Frankenhausen, May 15, 1525, and in the bloody reprisal that followed. It was typical of Luther that he refused to climb down, to regain lost popularity, and neither thereafter nor at any time can he be accused of subservience to rulers. As he had once refused to become the tool of the knights, so he had never "taken up" the peasant cause. But he confirmed many peasants in their preference for the radical ideology, which was soon to find more peaceful coherence in the Anabaptist movement.

Watershed year, 1525. *Luther and Erasmus.* In other ways, too, 1525 was a watershed in Luther's career. At the height of the Peasants' War in June 1525, "to spite the devil" he had married Katherina von Bora, a former nun. He certainly needed looking after, and she proved an admirable wife and a good businesswoman. His home meant a great deal to him and was an emblem for him of Christian vocation, so that he included domestic life among the three hierarchies (or "orders of creation") of Christian existence in this world, the other two being political and church life. In the same year there came his open break with the great Humanist Erasmus. The differences between the two men had long been apparent, and Erasmus, who found in Luther the type of violent, dogmatic mendicant theologian he had always detested, liked what he saw of the Reformation less and less. Nonetheless, both men had a common band of admirers and friends and entered the arena with reluctance. Erasmus, in his *De libero arbitrio,* or "Concerning Free Will" (1524), attacked Luther's doctrine of the enslaved will and provoked a resounding reply in Luther's *De servo arbitrio,* or "Concerning the Bondage of the Will" (1525), a one-sided, violent treatise that, nevertheless, includes profundities still fruitfully debated. In that year, too, Frederick the Wise died. The two men had met only once, but Luther owed much to this prince. The new ruler, the elector John, and his successor John Frederick were Luther's devout supporters and with other princes, notably Philip, landgrave of Hesse, and Albert of Brandenburg, formed a coherent group in the imperial Diet.

Luther's marriage

The Diets of Speyer. The hostility of Charles V to the Reformers and his devotion to the Catholic faith never altered, but he had to take account of political exigencies, his quarrels with the Pope and with the king of France, and the need for support against the Turks. At the Diet of Speyer in 1526, the Edict of Worms was suspended, pending a national council; in the interval it was ruled that each prince must behave as he could answer to God and to the emperor. Luther stated that there was no fear or discipline any longer and that everyone did as he pleased. As a result, it was possible to plan the reorganization of the Saxon Church, and a visitation was carried out by jurists and theologians (1527–28). Some scholars have seen a tension between Melanchthon's *Instruktion für die Visitatoren,* or "Instructions for the Visitation" (1528), and Luther's comments, which may reveal his distrust of secular intervention in spiritual affairs; and though he thoroughly approved of the development of the evangelical *Landeskirchen* ("territorial churches"), there were to be aspects of Lutheranism that blurred rather than reflected Luther's theological distinctions. At the second Diet of Speyer in 1529, renewed Catholic pressure led to the reversal of earlier concessions, drawing from the evangelical princes, and from a number of cities, a protest that won them, for the first time, the name Protestant.

The eucharistic controversy. *Doctrinal differences among the Reformers.* Doctrinal differences about the Eucharist broke the common evangelical front. Though all the Reformers repudiated the sacrifice of the mass, they were deeply divided about the nature of the divine Presence. Luther, with simple biblicism, insisted that Christ's words "This is my body" must be literally interpreted, because allegory is not to be used in interpreting Scripture unless the context plainly requires it. Karlstadt's fanciful argument (that the word this referred not to bread and wine but the Lord's physical body) was soon dropped. Zwingli won many to his view that "is" must be taken as "means," and his learned friend, the Humanist John Oecolampadius, brought support from the early Church Fathers for a spiritual Presence and stressed the idea of the 2nd-century Tertullian that "body" meant "sign of the body." Thus, the initial debate was about interpretive principles, about the words of institution, though the scriptural argument moved to the relevance or irrelevance of the Gospel According to John (*e.g.,* "he who eats my flesh and drinks my blood has eternal life" [John 6:54]).

Luther and Zwingli

The debate turned to the intricate matter of Christology (*i.e.,* doctrine of Christ). Zwingli insisted on the distinction between the two natures of Christ and that because it is the property of a human body to be in one place, Christ's human body was not here but in heaven. Luther, on the other hand, stressed the indivisible unity of the one Person of Jesus Christ, the mediator. Without going into a metaphysical doctrine of "ubiquity," or Presence everywhere (which was developed by other Lutherans), he asserted that Christ is present wherever he wills to be and that we are not to think of him in heaven "like a stork in a nest." Martin Bucer and the Strassburg theologians echoed the more positive stresses of the Swiss, and Bucer used the Realist language of the early Church Fathers to

support a true, spiritual Presence. Luther's treatise *Dass diese Worte Christi "Das ist mein Leib" noch fest stehen wider die Schwärmgeister* ("That these words of Christ 'This is my Body' still stand firm against the Fanatics," 1527) showed that in three years of controversy he had not budged. Zwingli's Latin tract *Amica exegesis* ("A Friendly Exegesis" 1527) was far less amicable than the title suggests and brought a great outburst from Luther, the impressive *Vom Abendmahl Christi, Bekenntnis* ("Confession of the Lord's Supper," 1528). This convinced Bucer that he had misunderstood Luther, who did not mean a local, confined Presence; and from then on he intensified his awkward, well-intended attempts to make peace.

The Marburg Colloquy and the Diet of Augsburg. The political advantages of a common front were obvious, not least to the vulnerable Zwingli and Philip, landgrave of Hesse, and the prince invited theologians of both sides to a private colloquy at Marburg in October 1529. Luther began by saying that in his opinion Zwingli did not know much about the gospel. When Zwingli asked if it was permissible for a Christian to ask how Christ could be present in the bread and wine of the Lord's Supper, Luther replied that if the Lord commanded him to eat crab apples and manure, he would do it because it was a command. After three days' debate, there was no agreement about the Eucharist, though the air had been cleared of many misunderstandings. But if the conference failed, there were agreements on other issues, and these might have been fruitful had not the coming imperial Diet caused the Wittenberg theologians to draw away from the Swiss. As an outlaw, Luther could not attend this fateful Diet of Augsburg and had to fidget in the castle of Coburg, leaving the care of the gospel to Melanchthon, who did very well and produced in the Augsburg Confession (1530), one of the great documents of the Reformation as well as a normative confession of Lutheranism.

The Augsburg Confession

Luther used his influence to stiffen the elector against compromise, though from this time onward he could not refuse his consent to political Protestantism as it took a more and more military shape in the Schmalkaldic League, which was established by Protestant princes in preparation for armed resistance to Catholic aggression. The political situation again changed swiftly, however, and, confronted with the Turkish invasion, the Emperor agreed to a truce with the Protestants in the Religious Peace of Nürnberg (1532). This was a valuable breathing space, and its effects are evident in Luther's writings in the next years. Now, more and more, Luther left matters to the action of Melanchthon. Opponents attempted to break up the friendship of the two. Luther said, regarding this matter, that if Melanchthon would allow himself to be won over by their opponents, "he could easily become a cardinal and keep wife and child."

Growth of Lutheranism, 1530–46. *Melanchthon's leadership.* Luther acquiesced in the eucharistic agreement—by which the south Germans reached agreement on the Lord's Supper—that the triumphant Bucer brought off with Melanchthon in 1536 (the Wittenberg Concord), though Bucer was unable to widen the agreement and bring in the Swiss. When an English embassy from Henry VIII arrived to discuss joining the Schmalkaldic League, it was Melanchthon who drew up the theological agenda (the Wittenberg Articles, 1535) with an ambiguous statement of justification of which Luther wrote, "this agrees well with our teaching." But he would not follow Melanchthon when he thought he wrote too irenically about the papacy, and as the papal council loomed near he penned his own uncompromising Schmalkaldic Articles (1537).

Melanchthon's great work in the field of education was to earn him the name preceptor of Germany, but Luther too was important in this matter. His open letter to the councillors of Germany about the need for schools (1524), and his published sermon *Dass man Kinder zur Schulen halten solle* ("On Keeping Children at School," 1530) show how wise and forward looking was his concern for education. He himself composed two important catechetic documents, the lovely classic, *Kleiner Katechismus* ("Small Catechism"), and *Grosser Katechismus* ("Large Catechism," 1529), for teachers and pastors.

In Wittenberg Luther had a group of able colleagues: Justus Jonas, Johannes Bugenhagen, Feliks Krzyzak (Cruciger). In scores of cities his disciples and friends spread the evangelical teaching that formed the Lutheran pattern of church life. Luther, though not pre-eminent as a liturgist, provided orders of worship from which numerous other *Kirchenordnungen* ("church orders") were derived. The influence of Luther's writings was everywhere felt in the Western Christian world. It was in Scandinavia that the Lutheran Church struck its deepest roots and won its most complete ascendancy, but it also had deep influence in Austrian and Hungarian lands. Luther realized the importance of hymns and encouraged his friends to write them. He wrote a score of fine hymns, four of which appeared in his first Protestant hymnbook in 1524. The famous "Ein feste Burg ist unser Gott" ("A Safe Stronghold Our God Is Still" or "A Mighty Fortress Is Our God") became almost an event in European history. During the last decade of his life, John Calvin (1509–64) was the rising portent in Switzerland, though Luther's personal contact with him was slight. He continued to attack bitterly the *Schwärmer* ("fanatics"), who then included besides the Anabaptists a number of radicals such as Kaspar Schwenckfeld, a Reformer who tried to mediate between various groups. Although he maintained to the end his view that error can be conquered only by the Word, Luther came to accept the punishment of the Anabaptists.

Luther's hymns

The affair of Philip of Hesse. In 1540 Bucer and Melanchthon took the initiative in conniving at the deplorable bigamy of Philip of Hesse, but Luther was involved and had he willed could have stopped it. It would have been easy for Philip to remedy his incorrigible incontinence by taking a mistress, but this he refused to do, though his guilty conscience kept him from the sacrament. The desperate device, as a lesser of evils, was to grant him a secret dispensation to take a second wife. When the affair became public Luther angrily threatened to expose the whole story. He himself was so far from lowering moral standards that in the next years he threatened to leave Wittenberg because public morals there were a shame on a city that had known the evangelical teaching so long. After a serious illness in 1537 he was an almost chronic invalid, prematurely aged, seldom free from discomfort, often in pain, and he brought his teaching career to an end with lectures on Genesis. In the last decade of his life he had to witness the recovery of the papacy, which he thought to have been mortally wounded, in the preparations for the Council of Trent (1545–63), and the growing menace of Catholic military might. His last outstanding controversial treatise was *Von den Conciliis und Kirchen* ("Of Councils and Churches," 1539). Among his last writings, *Against the Anabaptists, Against the Jews, Against the Papacy at Rome, Founded by the Devil,* the most violent is the last, coarse and angry but still defiant.

Luther's last activities. Early in 1546 Luther was asked to go to Eisleben to mediate in a quarrel between two arrogant young princes, Counts Albrecht and Gebhard of Mansfeld. He was old and ill, but they were his *Obrigkeiten* ("authorities") to whom he owed obedience, and he set off in the snowy winter, leaving his wife stiff with anxiety. His letters to her teased her, comforted her, and spoke at last of a mission successfully accomplished. But he had overtaxed his strength, and in a few hours the chill of death came upon him. He died in Eisleben, the town where he was born, on February 18, 1546, and his body was interred in the Church of All Saints, Wittenberg. The great funeral orations by Bugenhagen and Melanchthon, who knew him so well, are not to be dismissed as simply panegyric. They witness that his intimates regarded him as a really great man, standing within the historic succession of prophets and doctors of the church, through whose life and witness the Word of God had gone forth, conquering and to conquer.

Luther's death

LUTHER AS THEOLOGIAN

Luther was no systematizer, like Melanchthon or Calvin, though the dissensions among Lutheran theologians after his death, each appealing to one aspect of his thought, testify to the width, coherence, and delicate balance of

Luther's own teaching. The basis of his theology was Holy Scripture; and though the differences between his own and Augustine's thought are important, Augustine must stand next to the Bible among the influences upon his mind. The doctrines of salvation were of prime importance for him, and here the two great, many-sided complex conceptions of the Word and of faith are important. His often subtle doctrine about civil obedience was not always understood by his later followers, and non-theological factors in German history perpetuated and, to a certain extent, even perverted this misunderstanding. His doctrine of Christian vocation in this world and the importance of human life in the world became part of the general Protestant and Puritan inheritance. In other matters—in the room allowed for Christian liberty, in his conception of the part played by law in Christian life, and in his insistence on the Real Presence in the Eucharist—his theology differs from the patterns that emerged in the Reformed (Presbyterian) churches, in Puritanism, and in the sects such as the Anabaptists.

MAJOR WORKS

In Latin

THEOLOGICAL WORKS: *Epistola Lutheriana ad Leonem decimum summum pontificem. Dissertatio de libertate Christiana per autorem recognita* (1519; "Concerning Christian Liberty"); *De votis monasticis* (1521); *De captivitate Babylonica ecclesiae praeludium* (1520; "A Prelude Concerning the Babylonian Captivity of the Church"); *De servo arbitrio* (1525; "Concerning the Bondage of the Will").

CONTROVERSIAL WRITINGS: *B. Martini Lutheri theses Tezelio, indulgentiarum institori oppositas* (1517; Ninety-five Theses); *Rationis Latomianae pro incendiariis Lovaniensis scholae sophistis redditae Lutheriana confutatio* (1521).

EXEGESIS: *Enarrationes epistolarum et evangeliorum, quas postillas vocant* (1521).

In German

THEOLOGICAL WORKS: *Von den guten Wercken* (1520; "Of Good Works"); *Von welltlicher Uberkeytt, wie weytt man yhr gehorsam schuldig sey* (1523; "Of Earthly Government"); *Das diese wort Christi (Das ist mein leib etce.) noch fest stehen widder die Schwermgeister* (1527; "That These Words of Christ 'This is My Body' Still Stand Firm Against the Fanatics"); *Vom Abendmal Christi, Bekenntnis* (1528; "Confession of the Lord's Supper"); *Von den Conciliis und Kirchen* (1539; "Of Councils and Churches").

CONTROVERSIAL WRITINGS: *An den christlichen Adel deutscher Nation* (1520; "Address to the Christian Nobility of the German Nation"); *Widder die hymelischen Propheten von den Bildern und Sacrament* (1525; "Against the Heavenly Prophets in the Matter of Images and Sacraments"); *An die Radsherrn aller Stedte deutsches Lands: Das sie Christliche Schulen auffrichten und hallten sollen* (1524); *Ermanunge zum Fride auff die zwelff Artikel der Bawrschafft ynn Schwaben* (1525); *Wider die mordischen uñ reubischen Rotten der Bawren* (1525); *Wider Hans Worst* (1541); *Wider das Bapstum zu Rom vom Teuffel gestifft* (1545).

TRANSLATIONS AND EXEGESIS: *Das Newe Testament Deutzsch* (1522); *Biblia, das ist, die gantze Heilige Scrifft Deudsch* (1534); *Das Magnificat verteuschet und ausgelegt* (1521).

OTHER WORKS (LITURGICAL): *Deudsche Messe* (1526). (DIDACTIC): *Der kleine Catechismus* (1559; "Small Catechism"); *Deudsch Catechismus* (1529; "Large Catechism"). Among his hymns the most famous is probably "Ein feste Burg ist unser Gott" ("A Mighty Fortress Is Our God").

(E.G.R.)

BIBLIOGRAPHY

Luther's writings: The *Works of Martin Luther,* 6 vol., Philadelphia ed. (1915–32, reissued 1982); *Luther's Works,* American ed., ed. by J. PELIKAN and H.T. LEHMANN, 54 vol. (1955–76), henceforth an indispensable tool for English study. There is a single-volume anthology edited by J. DILLENBERGER, *Martin Luther: Selections from His Writings* (1961); see also ERNEST G. RUPP and W.B. DREWERY, *Martin Luther* (1970). The following are important volumes in the "Library of Christian Classics": vol. 15, *Lectures on Romans,* ed. by W. PAUCK (1961); vol. 16, *Early Theological Works,* ed. by J. ATKINSON (1962); vol. 17, *Luther and Erasmus,* ed. by ERNEST G. RUPP and P.S. WATSON (1969); and vol. 18, *Letters of Spiritual Counsel,* ed. by T.G. TAPPERT (1955). See also *A Commentary on St. Paul's Epistle to the Galatians,* ed. by P.S. WATSON (1953). In German the definitive edition is the Weimar edition by J.C.F. KNAAKE *et al.* (1883–).

Studies: For biographical studies in English, see PETER MANNS, *Martin Luther: An Illustrated Biography* (1982), with emphasis on the religious context; JOHN M. TODD, *Luther* (1982), a popular biography; MARK A. EDWARDS, *Luther and the False Brethren* (1975), detailing the years between the Diet of Worms and Luther's death; HARRY G. HAILE, *Luther: An Experiment in Biography,* (1980) which concentrates on the last 10 years of his life; DAVID C. STEINMETZ, *Luther and Staupitz: An Essay in the Intellectual Origins of the Protestant Reformation* (1980), a study of the influence on Luther of his early confessor and friend; ROLAND H. BAINTON, *Here I Stand!* (1950, reissued 1978); FRANZ LAU, *Luther* (1963; originally published in German, 1959); and W.J. KOOIMAN, *By Faith Alone* (1954; originally published in Dutch, 1946). PRESERVED SMITH, *The Life and Letters of Martin Luther* (1911), is the best of the older studies. For a broad survey, see E.G. SCHWIEBERT, *Luther and His Times* (1950); on the young Luther, see R.H. FIFE, *The Revolt of Martin Luther* (1957); for a brief account, see ERNEST G. RUPP, *Luther's Progress to the Diet of Worms, 1521* (1951). On Luther and his age, see JAMES ATKINSON, *Martin Luther and the Birth of Protestantism,* 3rd ed. (1982); ARTHUR G. DICKENS, *Reformation and Society in Sixteenth Century Europe* (1966); JOSEPH LORTZ, *The Reformation in Germany,* 2 vol. (1968; 5th German ed., 1965); and W. PAUCK, *Heritage of the Reformation,* rev. ed. (1961). On Luther's theology, see GERHARD EBELING, *Luther: An Introduction to His Thought* (1970; originally published in German, 1964); PHILIP S. WATSON, *Let God Be God* (1947); ERNEST G. RUPP, *The Righteousness of God* (1953, reissued 1963); HEINRICH BORNKAMM, *Luther's World of Thought* (1958; originally published in German, 1953); B.A. GERRISH, *Grace and Reason* (1962, reprinted 1979); REGIN PRENTER, *Spiritus Creator* (1953; originally published in German, 1944); and IAN D.K. SIGGINS, *Martin Luther's Doctrine of Christ* (1970). On the influence of Luther, see E.W. ZEEDEN, *Martin Luther und die Reformation,* 2 vol. (1950–52; Eng. trans. of vol. 1, *The Legacy of Luther,* 1954); and EDGAR M. CARLSON, *The Reinterpretation of Luther* (1948), a survey of Scandinavian Luther studies. Important studies in languages other than English include: KARL HOLL, *Gesammelte Aufsätze zur Kirchengeschichte,* 3 vol. (1921–28); EMANUEL HIRSCH, *Lutherstudien,* 2 vol. (1954); RUDOLF HERMANN, *Gesammelte Studien zur Theologie Luthers und der Reformation* (1960); ERNST WOLF, *Peregrinatio,* 2 vol. (1954); J. HECKEL, *Lex Charitatis,* 2nd ed. (1973); E. BIZER, *Fides ex auditu,* 3rd ed. (1966); O.H. PESCH, *Die Theologie der Rechtfertigung bei Martin Luther und Thomas von Aquin* (1967); REINHARD SCHWARZ, *Fides, Spes und Caritas beim Jungen Luther* (1962); and B. LOHSE, *Mönchtum und Reformation* (1963). For psychological studies, see PAUL J. REITER, *Martin Luthers Umwelt, Charakter und Psychose,* 2 vol. (1937–41); and ERIK H. ERIKSON, *Young Man Luther: A Study in Psychoanalysis and History* (1958).

Madrid

Madrid is the capital of Spain, its largest city, and a national centre of arts and industry. With its surrounding province, also called Madrid, it forms one of the new autonomous regions of the post-Franco era. Its capital status reflects the centralizing policy of the 16th-century Spanish king Philip II and his successors. The choice of Madrid as capital, however, was also the result of Madrid's previous obscurity and neutrality, in that it lacked ties with an established, non-royal power, rather than of any strategic, geographic, or economic considerations. Indeed Madrid is deficient in other characteristics that might be thought to qualify it for a leading role. It is not on a major river, in the way that so many European cities are; the playwright Lope de Vega, referring to a magnificent bridge over the distinctly unimposing waters of the Manzanares, suggested either selling the bridge or buying another river. Madrid does not possess mineral deposits or other natural wealth, nor was it ever a destination of pilgrimages, although its patron saint, San Isidro, enjoys the all but unique distinction of having been married to another saint. Even the city's origins seem inappropriate for a national capital, since its earliest historical role was as the site of a small Moorish fortress on a rocky outcrop—part of the northern defenses of what was then the far more important city of Toledo, 40 miles (65 kilometres) to the southwest.

It was in 1607, a whole generation after Philip II took the court to Madrid in 1561, that Philip III officially made the city the national capital, a status it has retained ever since. Under the patronage of Philip and his successors, Madrid developed into a city of curious contrasts, preserving its old, overcrowded centre, around which developed palaces, convents, churches, and public buildings. This combination has created a quintessentially Spanish city, a true capital in character, with a sparkle and vitality all its own.

This article is divided into the following sections:

Physical and human geography 373
- The landscape 373
  - The city site and climate
  - The city layout
- The people 373
- The economy 375
  - Industry, commerce, and finance
  - Transportation
- Administration and social conditions 375
- Cultural life 375

History 376
- The early period
- Development under the Bourbon kings
- Modern Madrid

Bibliography 377

Physical and human geography

THE LANDSCAPE

The city site and climate. Madrid lies almost exactly at the geographical heart of the Iberian Peninsula. It is situated on an undulating plateau of sand and clay known as the Meseta (derived from the Spanish word *mesa,* or "table") at an altitude of 2,100 feet (635 metres) above sea level, making it one of the highest capitals in Europe. This location, together with the proximity of the Carpetovetónica Range, is partly responsible for the weather pattern of cold, crisp winters accompanied by sharp winds. Sudden variations of temperature are possible, but summers are consistently dry and hot, becoming especially oppressive in July and August, when temperatures sometimes rise above 100° F (37.8° C). Average temperatures range between 41° and 75° F (5° and 24° C), while average rainfall varies between a low of less than one-half inch (11 millimetres) in July up to about two inches in October, usually the rainiest month of the year. The temperate times of year are spring and fall, which are also the most attractive seasons for visitors.

Temperature variations

The city layout. Madrid is a city of contrasting styles, reflecting clearly the different periods in which change and development took place. The old centre, a maze of small streets around a few squares in the vicinity of the imposing Plaza Mayor, contrasts with the stately Neoclassical buildings and grand boulevards created by the most eminent architects of their day. Modern office buildings in the centre and swaths of apartment blocks around the outskirts attest to the styles and economic realities of present-day development. Much of Madrid gives the impression of being cramped. When Madrid was first made the capital the king obliged the city's inhabitants to let a floor of their houses to ambassadors and visiting dignitaries, which prompted many people to build structures with only one floor or sometimes (in the so-called *casas a la malicia,* or "spite houses") having two floors but with a facade giving the impression of only one. Subsequent development of the city generated an enormous demand for land, particularly with the extensive construction of public buildings and convents. The last of Madrid's four sets of city walls was built in 1625 and not demolished until 1860 (by which time the population of the city had quadrupled). The situation was not alleviated even when Napoleon's brother Joseph, who briefly interrupted the Bourbon line of kings, demolished the convents to create more open space. Joseph's nickname of "El Rey Plazuelas" ("King of the Small Plazas"; one of the few complimentary ones he was given) derived from the squares he created. They did little to appease the ecclesiastical authorities, whose alienation contributed to his downfall. One of the squares, the Plaza de Oriente, facing the palace of the same name, was cleared of 56 houses, a library, a church, and several convents.

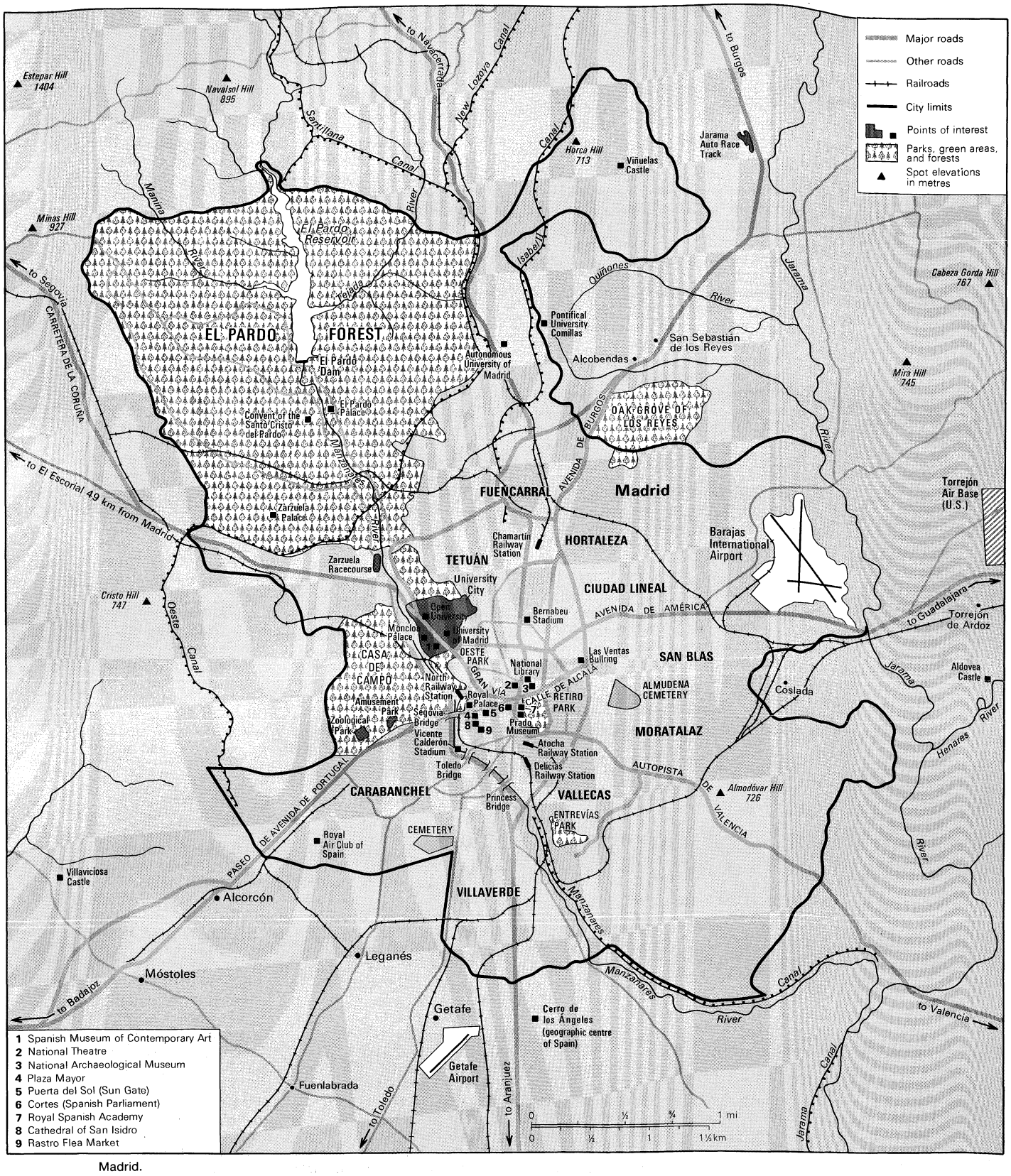
Creation of open space

"Los Madriles" ("the Madrids") is a traditional phrase that acknowledges the fact that each *barrio* (quarter) has developed its own style. There was also a geographical and social distinction among the *barrios altos, barrios centrales,* and *barrios bajos.* The last, spilling downhill from the Plaza Mayor along the Calle de Toledo toward the river, are still poor, albeit picturesque. Later development, also accommodating Madrid's poorer citizens, spread down toward the reclaimed marshland on both sides of the river, where low-cost housing is still to be found. Just over the brow of the hill is the Rastro, the popular flea market. Despite a number of urban development plans Madrid did not spread into the open spaces around it, not even crossing the Manzanares River until 1948. By contrast, however, the city as a whole has some extensive parks, with more open space overall than Paris. Some, like El Pardo or Casa de Campo, are survivals of hunting parks; the Retiro, on the other hand, is the site of a former royal palace.

Madrid has not escaped the problems common to so many modern cities. Pollution can be intense, traffic a problem. Personal safety is not as certain as it once was in the days of the *serenos* (night watchmen). But the city has preserved the charm, character, and vivacity that give it and its inhabitants a style of their own—an important aspect of modern Spain, where each region seeks to express its own identity.

THE PEOPLE

The flow of migration to Madrid, attracted chiefly by the city's expanding industrial belt, has created a modern population representative of the entire Spanish nation. A traditional nickname for the Madrileños is *gatos* ("cats"),



Madrid.

originally coined in the Middle Ages as a reference to the ability of local troops to scale castle walls. It would be no less apt as a reference to the local life-style and the late hours kept by the city's modern inhabitants, although keeping late hours is also common in other parts of Spain, especially in the heat of summer. People eat late, theatres and cinemas begin performances late as a matter of course, and the siesta is by no means dead, although the introduction of modern business methods and the influx of foreign interests have tended to lead to the introduction of the *semana inglesa* (literally "English week"), doing away with the long midday break. The city offers a wealth of cultural events and entertainments. It is not a cosmopolitan city as such, and yet its cultivated people tend to be widely read, while the young are up-to-date with the latest pop music. There is no large foreign population, nor are there distinct ethnic groups, although the number of *casas regionales*—regional clubs catering to people who

Late-hour life-style

have come to work from all over the country—reflects the source of labour for Madrid's burgeoning industrial and commercial sectors. Since the early 1970s the Latin-American population has become more numerous as the historic flow of those migrating to the Americas to escape economic misery or political conflict has been reversed. Madrid is a city that, with its style and flair, absorbs and holds those who live there or know it. Its inhabitants have a reputation for being attached to it; in the words of a local proverb, "From Madrid to heaven, and in heaven a little window from which to see it."

THE ECONOMY

Industry, commerce, and finance. Being the centre for government, finance, and insurance is a factor that has for long contributed to the prosperity of the capital, not to mention activities such as tourism and the city's position as the nation's transportation hub. In the postwar period (which, of course, for Spaniards means post-1939—since that was when the Spanish Civil War ended, and since Spain was neutral in World War II) the city became an important manufacturing centre for the automotive and aircraft industries, and for electric and electronic equipment, and optics, the production of plastics and rubber, and consumer goods. Madrid, with Barcelona, dominates publishing in Spain.

Postwar industrial development

Transportation. The road and rail systems both converge on the capital from all corners of the country. A subway system, the Metro, serves Madrid with various lines that extend throughout the city. The international airport of Barajas lies about eight miles east of the city. A motorway (expressway) system encircles Madrid in a roughly pentagonal shape, coming to a point in the south. Other major motorways radiate from the encircling artery in all directions. There are numerous bus routes operated by municipal and private authorities; minibus service is also available. The road-building programs of the 1960s, when much was sacrificed to the convenience of automobile owners, have since been recognized as less than wholly beneficial. Some of the overpasses introduced to speed up traffic flow have since been dismantled.

ADMINISTRATION AND SOCIAL CONDITIONS

With the return of democracy to Spain in the late 1970s and the development of autonomous regional governments, more emphasis has been placed on local consultation and issues such as the future of the environment. In 1982 the city administration carried out a massive public opinion survey to find out what people really wanted at the neighbourhood level. The resulting General Ordinance Plan (Plan General de Ordenación) was an attempt to establish a long-term, full-scale scheme for future directed growth, aiming not only to modernize the infrastructure of essential services but also to improve the quality of life in the city. Despite the introduction of the autonomous regions, Madrid continues to be the focus of Spain's government. It is also a bishopric, headquarters of the army corps, and residence of the captain general of the first military region. The supreme court and government ministries are also located there, besides the Cortes (the Spanish parliament), which is housed in a 19th-century Neoclassical building. Characterized by the bronze lions (made from melted down Moorish cannon) flanking the entrance, it is one of the smallest parliament buildings in Europe. Local administration is under the direction of a mayor and city council.

The General Ordinance Plan

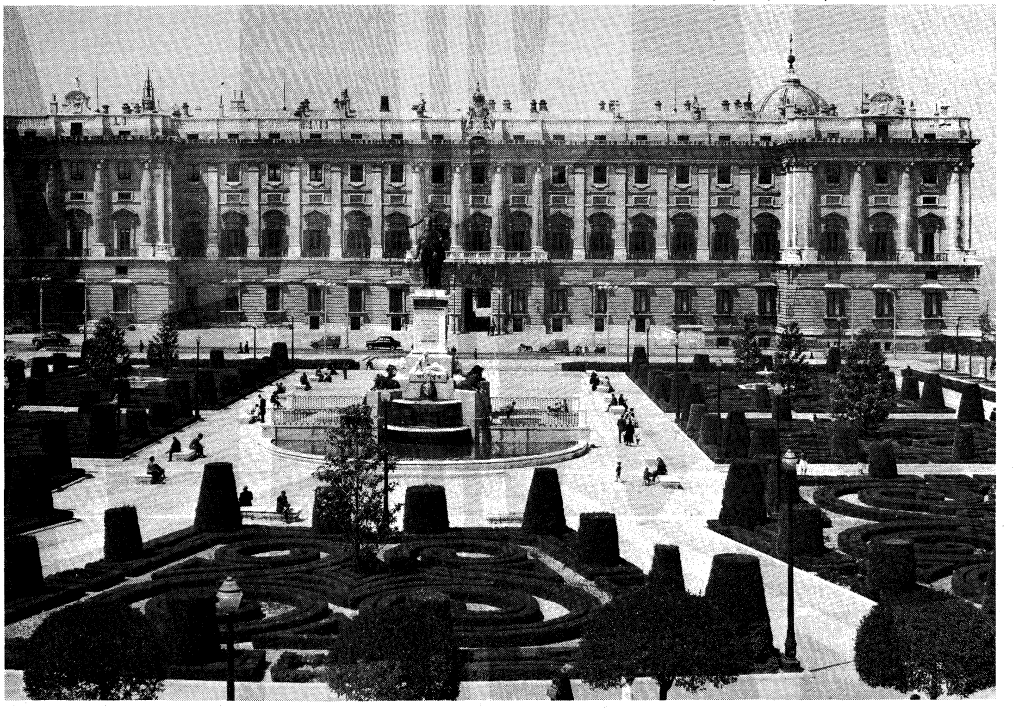
CULTURAL LIFE

Modern pressures have perhaps inhibited the extensive street life for which Madrid has been famous, although people still live very much in the streets, especially during the intense heat of summer when the café terraces fill and people stroll up and down in the evenings. Modern culture, in the form of film, theatre, and music, is extensively represented, as is to be expected in a city with several major universities and academies. But the cultural life for which Madrid was once noted, the *tertulias,* that is to say the informal conversational gatherings and informal societies, have all but faded, along with the elegant cafés that housed them. Madrid's literary traditions, its associations with Lope de Vega, Calderón, Cervantes, Quevedo, Pérez Galdós, Larra, Baroja, and Azorín, continue in the city's varied cultural life, as demonstrated by the fact that it is one of the major publishing centres for the Spanish-speaking world.

Modern Madrid has attractions at all levels. The 23,000-seat bullring, Las Ventas, is the largest in Spain and the place where the novice bullfighters have to display their skills in the *alternativa* (the occasion on which a matador kills his first bull) in order to become established. The season runs from March to October. There are two major football (soccer) teams, with the annual matches against the Barcelona teams among the high points of the year. Important matches are played in two stadia, the

Bullfighting tradition

By courtesy of the Spanish National Tourist Office



Palacio Real and gardens, Madrid.

125,000-seat Santiago Bernabéu and the 70,000-seat Vicente Calderón. The *verbenas,* special fiestas held in each quarter in honour of its patron saint, are regular public events, especially in warm weather, with San Isidro (mid-May) taking pride of place. The *zarzuelas* (light opera of mildly satirical flavour, indulging in topical comment and set by tradition in Madrid) are commonly held in the open air at this time. There are, in fact, more than 40 parks and public gardens, the principal ones being the Retiro, Campo del Moro, Casa de Campo, and Oeste Park, not to mention the curious temple of Debod (an ancient Egyptian temple acquired by Spain at the time of the construction of the Aswān High Dam) near Rosales, with splendid panoramic views over the city.

Museums abound. Unusual ones include those for theatre, railways, and (understandably enough for Spain) *tauromaquia,* the bullfight. The city is richly endowed with artistic masterpieces: tapestries in the Casa de Cisneros (the mayor's residence) and the Royal Palace (Palacio Real); paintings by Bruegel and Titian in the convent of the Descalzas Reales; and Spanish and foreign masters in the Palacio de Liria, home of the dukes of Alba. The most famous collection is housed in the Prado, which displays the artworks collected by the Spanish monarchy over the ages and reflects the pattern of Spain's alliances. Charles V and Philip II were patrons of Venetian art; Philip IV was a great collector in the 17th century; and the accession of the Bourbon family led to an influx of French works. Spain's control of the Netherlands led to a solid Flemish section. El Casón del Buen Retiro is nearby and houses 19th- and 20th-century works, the most famous probably being Picasso's "Guernica." This painting in 1981 was sent to Spain from New York City in accordance with Picasso's directive that the painting be moved there only after democracy had returned to the country.

The Prado

Notable among an abundance of libraries are the prestigious National Library (Biblioteca Nacional) and the Library of the Royal Palace, acclaimed for its historic collection. Madrid is also famous for its secondhand bookshops, and the Feria del Libro (book fair), held in the spring, is a widely heralded event.

Madrid is Spain's foremost centre of higher education and includes several of the country's leading universities, among which are the Open University (Universidad Nacional de Educación a Distancia), the Complutensian University, and the Polytechnical University, all in University City, and to the north the Autonomous University.

History

The early period. The Arab town, or medina, grew around the alcazar (castle) on a promontory overlooking the Manzanares River. The name Majerit first appears in AD 932, when the Christian king Ramiro II of León razed the town's walls, but there are traces of earlier (even prehistoric) habitation. The Christian king Alfonso VI of Castile and León captured the town from the Muslims in 1083, and thereafter a number of kings of Castile spent time there. The parliament (Cortes) was called there as early as 1309. The alcazar was damaged in an earthquake in 1466 and the subsequent medieval palace was extended by various monarchs, notably Charles I and Philip II. In this period the town grew to the east up both sides of what are now the Calle Mayor and the Calle de Segovia, with the Moors (who continued to live there until after the Christian reconquest of Spain was completed in 1492) jammed into the southwest corner, which is still called the Moreria. The whole of the city at this time was only 500 by 900 yards in area. Some of the street patterns of the pre-16th century city remain, but few buildings; one that still stands is the much-restored Casa de los Lujanes, where it is believed the French king Francis I was once held prisoner. Charles I enjoyed hunting near Madrid, and it is said that the widening of the city gates to accommodate his carriages opened the cramped streets to heavier traffic, a process that was to increase when the court was properly established in the city. By 1598 the population of Madrid had reached 60,000, and by the time of the first extant plan of Madrid (Pedro Teixeira's in 1656) it had grown to be an imposing city of 100,000 people and 11,000 buildings.

Christian takeover

Under the Habsburg monarchs, the Madrid de los Austrias expanded even more rapidly. The foreign ministry (1634), the Casa de Cisneros, and the Segovia Bridge date from this time, as does the church of San Isidro el Real. (It is also, incidentally, still Madrid's temporary cathedral—Nuestra Señora de la Almudena, begun rather tardily in 1883, is still under construction.) Architects such as Juan de Herrera (who died in 1597) and Francisco de Mora contributed to the monumental quality of the city. But the most striking contribution of this period is generally considered to be the Plaza Mayor, designed by Juan Gómez de Mora and built between 1617 and 1619; it was modified after the great fire of 1790. Graceful in concept, it is surrounded by five-story houses with balconies and topped with steeples. Nine archways open onto the plaza at oblique angles from surrounding streets, and the continuous arcade at street level contains shops and restaurants. Bullfights (in those days conducted by noblemen on horseback), fireworks displays, and plays all took place there, as did the grim ceremonies of the Inquisition. Until 1765 public garrotings were also carried out there. The last bullfight to take place in the Plaza Mayor was in honour of the wedding of Isabella II in 1846.

Habsburg rule

Development under the Bourbon kings. No less impressive a landmark dates from the next great phase of the city's growth, under the Bourbons, whose side Madrid took against the Habsburgs in the War of the Spanish Succession (1701–14), although the city was briefly occupied by pro-Habsburg troops. The Royal Palace was begun by Philip V after the disastrous fire that destroyed the Alcazâr on Christmas night, 1734. His grandiose plan, with 23 inner courts, was never realized, although the finished work did have 500 rooms. It was a fitting addition to the other major city features created under his patronage, the Royal Spanish Academy (Real Academia Española), the National Library, and the Royal Academy of History (Real Academia de Historia). The Royal Palace, with its elegant granite and limestone walls, contains a ceiling by Tiepolo in the throne room and, in the Armeria, one of the world's finest collections of armour, including the swords of the conquistadores Hernán Cortés and Francisco Pizarro. The last king of Spain actually to live there was Alfonso XIII, whose apartments have been preserved just as he left them when he abdicated in 1931. The royal family now resides in the more private and less ostentatious La Zarzuela Palace, set in its own grounds to the northwest of Madrid. A less important palace was also favoured by the Bourbon kings—the Palacio del Buen Retiro; its gardens, which were much admired for their French style, are still a much frequented open space.

Growth under the Bourbon kings

The greatest Bourbon builder was Charles III (1759–88), who is known as the mayor-king for his interest in the growth and development of the city and his many contributions to its skyline. This was the age of Enlightenment, and Charles tried to establish a harmony between leisure and science, culture and industry. His style was cosmopolitan, reflecting the tastes of the Europe of his time. With his concern for the appearance of the city, its gates, avenues, and trees, he anticipated the designs of modern city planners. He relied heavily in his schemes on the works of three Neoclassical architects, Francisco Sabatini, Ventura Rodríguez, and Juan de Villanueva. During this period the city continued to grow eastward to the present Plaza de la Independencia, which is the site of a monumental arch, the Puerta de Alcalá, built in 1778 and still a key landmark. Another famous landmark, the central post office, in the Puerta del Sol, also dates from this time. All distances in the country are still measured from the zero-kilometre stone beneath its wall. The square itself holds a firm place in people's affections; it is particularly popular on New Year's Eve when people go to hear the clock strike midnight and to eat the 12 grapes that supposedly ensure good luck in the following months. In earlier times the square was significant because of the stagecoaches that left from there to all parts of the peninsula. Its appeal made it the site of various innovations in urban amenities, from the first gas lamps in 1830 through the first mule

Charles III, the mayor-king

trams and first public urinal to the first electric streetlights and electric streetcars. Charles III would doubtless have approved of all these innovations. In his time he set up the Botanic Garden (which still exists), with a "physic" garden from which anybody could (and still can) collect medicinal herbs. He planned a natural history and science museum next to it but died before it could be completed.

Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was installed on the throne. On May 2, 1808, there was a mass uprising against Joseph, leading to what the Spaniards term "la Guerra de la Independencia." Ferdinand VII, on his return in 1814 from imprisonment by Napoleon, bestowed the title of "heroic" upon the city. In 1819 the building intended by Charles III to house a natural history and science museum was completed. Into it Ferdinand moved artworks of the royal collection, until then scattered among various palaces. This was the start of what was to become one of the world's major art galleries, the Prado. The Madrid of this period can still be studied in close detail, thanks to the remarkable model constructed by León Gil Palacios in 1830. It was during this period that the city expanded to the north, under the direction of Joaquín Vizcaíno, a nobleman who was also mayor (as was customary at the time). He is also known as the man who introduced such innovations as street numbers for buildings, street lighting, and municipal refuse collection. The Paseo del Prado was extended by a new boulevard, called the Paseo de los Recoletos. This area, for years presenting an almost rural atmosphere, being bordered by imposing town houses with great gardens, is now the site of tall office blocks, apartment buildings, luxury hotels, and embassies, as well as the National Library, the National Archaeological Museum, and the Spanish Museum of Contemporary Art. One of the area's mansions, now converted into a bank, belonged to the Marqués de Salamanca, who in 1872 also contributed to the drive northward by building 28 streets on a grid plan, starting from the Calle de Alcalá and parallel to the Paseo. Still bearing his name, this area remains one of the most elegant *barrios* in Madrid.

Modernization under the Plan Castro

Somewhat earlier, in 1860, the Plan Castro, also referred to as the Ensanche, or widening, had further expanded and modernized the city, adding convenience and meeting the economic and commercial needs of the time. It was the first comprehensive, forward-looking modern plan for Madrid. However, it was to be frustrated by population growth, land speculation, and the poor areas that sprang up outside the planned zones.

Modern Madrid. Other plans followed the Plan Castro: Arturo Soria's linear city of 1892, with a planned 31-mile axis stretching away from the city and served by a suburban railroad; and the satellite city idea of Núñez Granes (1910). Neither of these found support. It was in the latter year that a major new landmark appeared. The *barrio* of San Bernardo was bisected by a broad way running from the Calle de Alcalá downhill to the Plaza de España, which is where the city's first high-rise commercial buildings were erected. This, the Gran Vía, was designed to be the main street of the city, and it has a characteristic vitality, with cinemas, coffeehouses, shops, and banks. Following the Civil War it was renamed Avenida José Antonio after the founder of the Spanish Fascist party, the Falange Española. Many streets and squares acquired new names at this time, but since Franco's death they have reverted to their old ones.

With the advent of the Republic in 1931, radial and ring roads appeared, and the Paseo de la Castellana was extended even farther. The city suffered heavily in the Civil War, with two years of aerial and artillery bombardment and combat lines drawn up as close as University City. War damage to public buildings was repaired, and ambitious reconstruction plans were drawn up, but once again these were by and large unsuccessful in practice. The city spread outward, swallowing its own suburbs; between 1948 and 1951 Madrid burgeoned to cover a total of 205 square miles, an approximately 10-fold growth rate. Land speculation, uncontrolled urban sprawl, and industrial development led to the changes of the 1960s, which, in common with the experience of many European cities, caused often irreparable damage to aspects of the city's cultural heritage. A 1963 plan redirected growth without controlling it. Legislation has now been passed to protect many of the city's most prized structures, and there is emphasis on preserving older buildings, even if only elements of them can be saved.

Mid-20th century expansion

BIBLIOGRAPHY

General works: All Madrid, trans. from Spanish (1979); and ALASTAIR BOYD, *The Companion Guide to Madrid and Central Spain* (1974), are descriptive guides. JUAN ANTONIO CABEZAS, *Madrid,* 3rd ed. (1971), examines all districts and cultural institutions of the capital. ARCHIBALD LYALL, *Well Met in Madrid* (1960); and ELENA SAINZ, *Living in Madrid* (1981), are more personal accounts.

History: FEDERICO BRAVO MORATA, *Historia de Madrid,* 4 vol. (1966–78); and JOSÉ AMADOR DE LOS RIOS, *Historia de la villa y corte de Madrid,* 4 vol. (1860–64; reprinted 1978), give wide coverage of historical events. Modern history, and especially the Civil War, is presented in GEORGE HILLS, *Battle for Madrid* (1976); MATILDE VÁZQUEZ and JAVIER VALERO, *La guerra civil en Madrid* (1978); DAVID JATO MIRANDA, *Madrid, capital republicana* (1976); DAN KURZMAN, *Miracle of November: Madrid's Epic Stand, 1936* (1980); and JOSÉ MANUEL MARTÍNEZ BANDE, *Frente de Madrid* (1976).

Social and economic development: NINA EPTON, *Madrid* (1964), describes the inhabitants of the various quarters of the city; MICHAEL KENNY, *A Spanish Tapestry: Town and Country in Castile* (1969), compares life in Madrid to that in a rural parish; DAVID R. RINGROSE, *Madrid and the Spanish Economy, 1560–1850* (1983); JULIO VINUESA ANGULO, *El desarrollo metropolitano de Madrid* (1976); and SANTOS JULIA DIAZ, *Madrid, 1931–1934: de la fiesta popular a la lucha de classes* (1984), are scholarly studies of social developments and conflicts.

Art and monuments: HARRY B. WEHLE, *Great Paintings from the Prado Museum* (1963), includes a sketch of the museum's history; F.H. SANCHEZ CANTON, *The Prado,* trans. from the French, new rev. ed. (1966), also includes a historical essay; MANUEL LORENTE, *The Prado, Madrid,* 2 vol. (1965), is an impressive collection of reproductions; CONSUELO LUCA DE TENA and MANUELA MENA, *Guide to the Prado* (1980; originally published in Spanish, 1980), is a more recent survey. JONATHAN BROWN and J.H. ELLIOTT, *A Palace for a King: The Buen Retiro and the Court of Philip IV* (1980), is a well-researched description of the pleasure palace and its notable art collection. See also AUREA DE LA MORENA BARTOLOMÉ *et al., Catalogo monumental de Madrid* (1976).

(B.E./T.J.Co.)

Malta

An independent republic, Malta consists of a small but strategically important group of islands in the central Mediterranean Sea. Throughout a long and turbulent history, the little archipelago has played a vital role in the struggles of a succession of powers for domination of the Mediterranean and in the interplay between emerging Europe and the older cultures of Africa and the Near East. As a result, Maltese society was molded by many centuries of foreign rule, with influences ranging from Arab to Norman to English.

There are five islands—Malta (the largest), Gozo, Comino, and uninhabited Comminotto and Filfla—lying some 58 miles (93 kilometres) south of Sicily, 220 miles (350 kilometres) north of Libya, and about 180 miles (290 kilometres) east of Tunisia, at the eastern end of that constricted portion of the Mediterranean Sea separating Italy from the African coast. The islands cover a combined land area of about 124 square miles (320 square kilometres). Valletta is the capital, although Sliema is the largest city.

The article is divided into the following sections:

Physical and human geography 378
The land 378
Relief 378
Drainage and soils 378
Climate 378
Plant and animal life 378
Settlement patterns 378
The people 379
The economy 379
Industry and tourism 379
Agriculture and fishing 379
Finance 379
Transportation 379
Administrative and social conditions 379
Government 379
Social services 380
Education 380
Cultural life 380
History 380

Physical and human geography

THE LAND

Relief. The main physical characteristic of Malta Island is a well-defined escarpment that bisects it along the Victoria Lines Fault running along the whole breadth of the island from Ras ir-Raheb (east of Nadur Tower) to the coast northeast of Gharghur. The highest areas are coralline limestone uplands that comprise a triangular plateau to the west. The uplands are separated from the surrounding areas by blue clay slopes, while undercliff areas are found where the coralline plateau has fallen and forms a subordinate surface between the sea and the original shore.

Limestone and coral

To the north the escarpment is occasionally abrupt and broken by deep embayments. To the south, however, the plateaus gradually descend from a height of 600–800 feet (180–240 metres) into undulating areas of globigerina (derived from marine protozoa) limestone less than 400 feet high. On the west there are deeply incised valleys and undercliff areas, while on the east there are several valleys descending to the central plains.

The west coast of Malta presents a high, bold, and generally harbourless face. On the east, however, a tongue of high ground known as Mt. Sceberras separates the bays of Marsamxett and Grand Harbour. These deep-water harbours contribute to the strategic importance of Malta, because there are relatively few of them in the Mediterranean. The bays are associated with nine seasonal creeks that include those of Sliema, Lazaretto, Msida, and Marsa. The northern shore is again bare and craggy, characterized by its coves and hills, which are separated by fertile lowlands.

In Gozo, the landscape is characterized by a broken coralline plateau to the north and by low-lying globigerina limestone plains and hills to the south.

Drainage and soils. Malta possesses favourable conditions for the percolation and underground storage of water. The impermeable blue clays provide two distinct water tables between the limestone formations. The principal source for the public supply of water has been the main sea-level water table. In the absence of permanent streams or lakes and given a considerable loss of rainfall, however, water supply has often been a problem. There are two seawater desalination plants, at Grand Harbour and at Gozo.

Shortage of fresh water

Maltese soils are mainly young or immature and thin. By law, when soils are removed from construction sites, they must be taken to agricultural areas, and level stretches in quarries are often covered with carted soil. Organic refuse from the towns is also used. Consequently, the soils are unusual and are partly a manufactured medium.

Climate. The climate is typically Mediterranean, with hot, dry summers, warm and sporadically wet autumns, and short, cool winters with adequate rainfall. Nearly three-fourths of the precipitation falls between October and March; June, July, and August are normally quite dry.

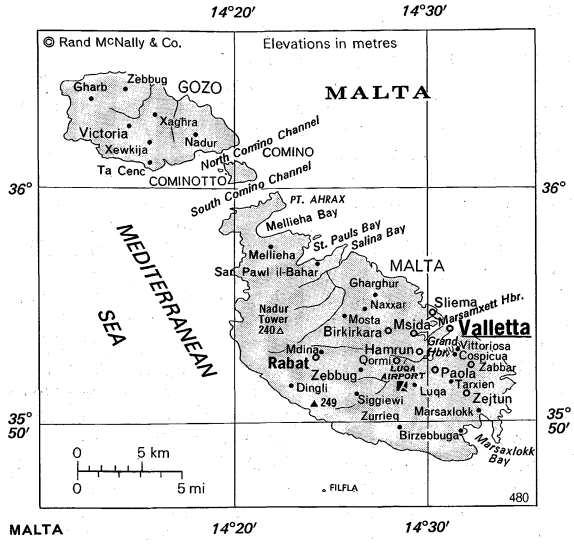
The temperature is very stable, the annual mean being 64° F (18° C) and the monthly averages ranging from 53° F (12° C) to 77° F (25° C). Winds are strong and frequent; the most common are the cool northwesterly (the Majjistral), the dry northeasterly (the Grigal), and the hot humid southeasterly (the Xlokk). The relative humidity is consistently high and rarely falls below 40 percent.

Plant and animal life. While wild vegetation is sparse, there is an abundance of cultivated potatoes, sulla (a leguminous fodder crop), onions, tomatoes, and vines. The great variety of trees includes the carob, fig, and chaste. The developed seashore vegetation includes golden and rock samphires, sea campion, spurge, saltwort, and marram grass.

Typical of the few native mammals are the hedgehog, the least weasel, the water and white-toothed shrews, and the pipistrel and other bats. Rats, mice, and some rabbits are also found. Resident birds include the spectacled and Sardinian warblers, the Manx and Cory's shearwaters, and the blue rock thrush. Linnets, tree and Spanish sparrows, buntings, rock doves, and several species of owl also form breeding populations, while birds of passage include ospreys, rollers, swallows, cuckoos, bee-eaters, and vultures. Common insects are beetles, grasshoppers, flies, mosquitoes, moths, bees, wasps, cockroaches, ants, and several species of butterflies. Ladybirds migrate from Sicily.

Settlement patterns. Until the mid-19th century, the Maltese lived mainly in the relative seclusion of clustered villages and hamlets; the fragmentation of farmholdings accentuated the individuality of the farming community. The *zuntier,* or church square, was the traditional focus of village life. With the growth of the Dockyard complex in the latter part of the 1800s, new settlements appeared around Grand Harbour, and the Sliema metropolitan region developed in the 20th century into the most fashionable part of Malta. The advent of industrial estates near major villages somewhat stemmed the exodus from the rural areas. Higher living standards and expatriate residents gave rise to residential developments in several hitherto forsaken areas such as Kappara (east of Zejtun) and The Gardens. By the end of 1979 one half of the island had become densely populated, threatening to devour the other half, and the need for town planning was recognized.

New towns



Gozo conserves its own rural character. The architecture of the new development at Ta' Cenc successfully blends with the island's natural beauty and is aesthetically stimulating.

THE PEOPLE

The islands' ethnic and linguistic composition reflects the heritage of many rulers. A European atmosphere predominates as a result of close association particularly with southern Europe. About 95 percent of the islanders are Maltese-born and the remainder includes mostly persons of English and Italian descent. During the 20th century, the increasing rate of Anglo-Maltese marriages added a new dimension to the ethnic structure of the population.

Maltese—the medium of daily conversation—is a distinct language that resulted from the interaction and fusion of North African Arabic and a Sicilian form of Italian. It is the only Semitic language that is officially written in Latin script; it became an official language of Malta in 1934. English is the other official language and is the medium of instruction in the schools. Italian was the language of church and government until 1934, and consequently it is still understood by a sizable sector of the islands' population.

Relations of state and church

Although Roman Catholicism is the state religion, there is full freedom for all religious beliefs. The islands are an independent province of the church, with two dioceses at Malta and Gozo, and two bishops serving the cathedrals at Valletta and Victoria. There are two Roman Catholic cathedrals at Mdina (Notabile) and Valletta, and an Anglican cathedral at Valletta. There is a mosque in Parxien.

THE ECONOMY

Malta's only mineral resource is the globigerina limestone that is used as building stone. Its other assets are its deep harbours, an adaptable, skilled labour force, and a strategic position as both a fuelling centre and, until March 1979, as a military and naval base. The economy, therefore, has been somewhat artificial and until 1979 determined by the vicissitudes of war and peace in the Mediterranean. In the 1950s Britain began the retrenchment of its naval and military forces on Malta, a move that necessitated a drastic diversification of the economy. A series of five- and seven-year plans were supported by government grants, loans, and other fiscal incentives, to encourage private investment mainly in the areas of manufacturing and tourism. These incentives were subsequently drastically altered in the hope of attracting large-scale industry from all over the world but particularly from Europe.

Industry and tourism. Economic plans professed to build on a tripod basis of industry, agriculture, and tourism. In fact, however, industrial growth lagged behind these plans, resulting in the successful establishment of only a few multinational corporations (mainly producing textiles). Many new enterprises failed to overcome the two main obstacles of marketing and management. From 1971, the government increasingly took over weak enterprises, sometimes closing them. New development was concentrated on manufactures, including the production of traditional consumer goods such as lace, silver filigree, pottery, glassware, and canework.

The influx of tourists and some immigration spurred the building of hotels and housing, but there are questions about the islands' capacity to cater to an annual total of tourists greater than the country's population. Besides transshipment services, the establishment of backup facilities for oil companies and of other outlets for traditional Maltese skills in cross-cultural dealings produced jobs and foreign currency earnings that boosted the Maltese currency.

Agriculture and fishing. Agricultural development is hampered by infertile soils and the lack of adequate water supplies, and large amounts of food and beverages are imported. Most farming is carried out on small terraced strips of land that preclude the introduction of large-scale mechanization. Consequently, agriculture continues to decline in terms of cultivated land. The farming labour force has grown increasingly older, female, and part-time, but agricultural production has risen gradually because of the improvement of techniques in the cultivation of some crops, especially horticultural ones. The major crops are grains, vegetables, fruit, and fodder for domestic consumption and early potatoes and onions for export. Flowers, seeds, plants, and cuttings also are exported. Cattle and goats are also important.

Fishing is seasonal and not fully developed. The fishing boats are small and are affected by weather conditions. The fish population off the island is quite sparse, and during the summer months small lamps are used by fishermen to attract the fish.

Finance. The Central Bank and the Malta Development Corporation were both founded in 1968. In that year also Malta joined the International Monetary Fund. In 1971 the island entered into a special trade relationship with the European Economic Community (EEC), a relationship that was subsequently revised with the aim of rendering it more favourable to Malta in its delicate stage of economic development.

Transportation. The road system connects all towns and villages and includes a coast road, a panoramic road, and a regional road, which connects Msida to the coast road via St. Andrews. Bus services radiating from Valletta provide cheap and frequent internal transportation. Taxis are abundant, and most families have a private automobile. Several daily sailings connect Malta and Gozo, and there is a regular ferry and car-ferry service between Malta and Sicily. Scheduled airlines, including the national airline, connect Malta with Britain and most of Europe and with North Africa and the Middle East.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The 1964 Independence Constitution, under which Malta was a constitutional monarchy and parliamentary state, was amended in 1974 to make Malta a republic within the Commonwealth. Its head of state is a president appointed by the Maltese Parliament, which, with a statutory life of five years, is basically derived from the British model. Local features include a single chamber with 65 members, while election is by proportional representation from 13 electoral divisions. The president acts on the advice of the Cabinet, which is composed of the prime minister and other ministers and is collectively responsible to Parliament. There is no municipal government in the islands.

The judicial system

Maltese law, which was codified mainly between the period of 1854 to 1873, is largely based on the Code Napoléon and Napoleonic law. Procedural common law and some commercial and maritime affairs are regulated by English principles, but judiciary precedent is not binding. Maltese is the language of the courts. Civil and criminal jurisdiction is almost exclusively vested in the Superior Court and the Court of Magistrates. The chief justice and judges of the Superior Court are appointed by the president on the advice of the prime minister, and their

duties are apportioned throughout the court system. The magistrates, who are appointed in the same way, sit in the lower courts.

Between 1964 and 1972 Malta's main defense dispositions were those contained in a 1964 Anglo–Maltese defense agreement with the United Kingdom guaranteeing mutual assistance. The island subsequently followed a policy of nonalignment and has its own regular armed force.

Social services. A scheme to integrate health services, essentially consisting of free hospital care and some domestic assistance, was introduced after 1971. Certain measures deemed to be necessary preliminaries to a wider-ranging national health scheme gave rise to conflict with the local medical association, and heavy reliance was then placed on foreign doctors.

In 1956, social insurance was introduced to cover employees more than 14 years of age and self-employed or unemployed persons between 19 years and pension age. A comprehensive contributory insurance scheme was introduced in 1971, integrating a variety of earlier legislation. The new plan included a pension amounting to two-thirds of an individual's salary at the time of retirement. Free hospitalization was introduced in 1979 as a step toward comprehensive national health insurance.

Education. The Labour government radically altered the education system, which was previously structured on British models and strongly influenced by the Roman Catholic Church. Compulsory education was extended to include all children from the ages of six to 16. An attempt at establishing an extreme form of the "comprehensive" system was abandoned; streaming (the selection of students to create groups that are homogeneous in age and intellectual ability) and examinations were at first discarded but later reintroduced; purely technical institutes were not compelled to follow the program. At the tertiary level, a student–worker scheme was introduced in 1978, students working for six months and studying for six months. Admission to institutions of higher learning was thereby linked to the availability of employment.

The University of Malta, founded as a Jesuit college in 1592 and reestablished as a state institution in 1769, is now known as the Old University and offers courses mainly in law. The former College of Arts, Science and Technology was reorganized as the New University in 1978 and took over most of the functions of the Old University. Both institutions were amalgamated by legislation into a single university to ensure the strict implementation of the government policy supporting utilitarian courses.

CULTURAL LIFE

Folk traditions and festivals

Folk traditions in Malta have evolved mainly around the festa to celebrate the patron saint of a village, marked by processions and fireworks. Good Friday is also celebrated with colourful processions in several villages. *Imnarja,* which takes place on June 29, is the principal folk festival; it is highlighted by folk-singing (*ghana*) contests and fried-rabbit picnics at Buskett. The annual Carnival is celebrated at Valletta with vigorous dancing displays that include the *Parata* and *Il-Maltija,* the Maltese national dance. Maltese radio and television stations are run by boards nominated by the government. A Libyan radio station is also in operation.

Apart from its unique Neolithic ruins, Malta contains important examples of its flourishing architectural school of the 17th and 18th centuries. It was essentially Classical with a balanced overlay of Baroque decorations. The Italian artists Caravaggio and Preti spent several years in Malta, the latter's most important paintings embellishing many of Malta's churches.

In the 20th century, a vernacular architecture was developed by Richard England and others. The composer Charles Camilleri introduced folk themes into his works, while Maltese literature was enriched by the poetry of the late national bard, Dun Karm. An interesting theatrical upsurge paralleled the emergence of Francis Ebejer as a brilliant playwright. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.*

History

The earliest archaeological remains date from about 3800 BC. Neolithic farmers lived in caves like those at Ghar Dalam (near Birzebbuga) or villages like Skorba (near Nadur Tower) and produced pottery that seems to be related to that of contemporary eastern Sicily. An elaborate cult of the dead of Stone Age or Copper Age culture evolved about 2400 BC. Initially centring around rock-cut collective tombs such as those at Ggantija (near Xaghra) and Hagar Qim (near Zurrieq), it culminated—probably through contacts with the cultures of the Cyclades and Mycenae—in the unique underground burial chamber (hypogeum) at Hal Saflieni (near Paola). This culture came to a sudden end at about 2000 BC, possibly as a result of invasions. The culture that replaced it, of southern Italian flavour, is evidenced today only by fragmentary remains. Bronze Age tools and weapons have been found at Borg in-Nadur (near Birzebbuga) and Tarxien Cemetery (near Paola), while Iron Age relics from about 1200 to 800 BC include cart ruts at Bingemma (near Nadur Tower).

Domination by Carthage and Rome

Between the 8th and 6th centuries BC contact was made with Semitic cultures. Evidence is scanty, however, and a few inscriptions found on Malta constitute the only indication of a Phoenician presence. There is more substantial proof of the Carthaginian presence in the 6th century BC; coins, inscriptions, and several Punic-type rock tombs have been found. It is certain that in 218 BC Malta came under Roman political control. Originally part of the Praetorship of Sicily, the islands were subsequently given the status of municipium and were thus allowed to coin their own money, send ambassadors to Rome, and control domestic affairs. According to tradition, St. Paul was shipwrecked in Malta in AD 60 and began to convert the inhabitants. The Maltese have been Christians uninterruptedly since that time.

With the division of the Roman Empire, in AD 395, Malta was given to the eastern portion dominated by Constantinople. Until the 15th century, it followed the more immediate fortunes of nearby Sicily, being ruled successively by Arabs (who left a strong effect on the language), Normans (who advanced the legal and governmental structures), and a succession of feudal lords. In 1530, however, it was ceded to the Order of the Hospital of St. John of Jerusalem (the Knights Hospitallers), a religious and military order of the Roman Catholic Church. Malta became a fortress and, under the Knights' grand master, Jean de La Valette, successfully withstood the Ottoman siege of 1565. The new capital city of Valletta became a town of splendid palaces and unparalleled fortifications. Growing in power and wealth—due mainly to their maritime adventures against the Turks—the Knights left the island an architectural and artistic legacy. Although there was little economic and social contact between them and the Maltese, they managed to imprint their cosmopolitan character on Malta and its inhabitants.

In 1798, Napoleon Bonaparte captured the island, but the French presence was short-lived, and the Treaty of Amiens returned the island to the Knights in 1802. The Maltese protested and acknowledged Great Britain's sovereignty, subject to certain conditions incorporated in a Declaration of Rights. The constitutional change was ratified by the Treaty of Paris in 1814.

British rule

Malta's political status under Britain underwent a series of vicissitudes in which constitutions were successively granted, suspended, and revoked. The economy became a function of British demands for Malta's military facilities, and the Dockyard developed into the colony's economic mainstay.

The island flourished during the Crimean War and was favourably affected by the opening of the Suez Canal in 1869. Self-government was granted in 1921 on a dyarchical basis whereby Britain shared power and responsibility with Maltese ministers who were elected by the legislature. But the battle over the relative roles of the English, Italian, and Maltese languages took its toll, and in 1936 the islands reverted to a strictly colonial regime in which full power rested in the hands of the governor. During World War II, the islands repelled the Axis powers against

severe odds, having been one of the most heavily bombed targets of that conflict. As a result, it was awarded the George Cross, Britain's highest civilian decoration. Self-government was granted in 1947, revoked in 1959, and then restored in 1962. Malta finally achieved independence within the Commonwealth on September 21, 1964. It became a republic on December 13, 1974.

The withdrawal of British military and naval personnel from its famous Dockyard—associated with the achievement of independence from the United Kingdom in 1964—created economic and political problems. From 1964 to 1971, Malta was governed by the Nationalist Party, whose attitude was firm alignment with the West. In 1971, however, when the Malta Labour Party came to power, its policy was nonalignment and special friendship with the People's Republic of China. In 1979, the total closure of the British base and the end of the British alliance were celebrated by the Maltese government as the arrival of "real" independence; but the existence of serious problems, especially with regard to the guarantee of full and productive employment and the deepening division between local political parties, was acknowledged. For current political history, see the annual issues of the *Britannica Book of the Year.*

BIBLIOGRAPHY

General works: *Malta: Isles of the Middle Sea* (1965), containing maps and photographs; SIR H.C.J. LUKE, *Malta: An Account and an Appreciation,* 2nd ed. rev. and enl. (1960, reprinted 1968); B. NEHRING, *Die maltesischen Inseln* (1966).

History: B. BLOUET, *A Short History of Malta* (British title, *The Story of Malta,* 1967); J.L. GODECHOT, *Histoire de Malte* (1952); E.D.S. BRADFORD, *The Great Siege* (1961); R. CAVALIERO, *The Last of the Crusaders: The Knights of St. John and Malta in the Eighteenth Century* (1960); J.J. CREMONA, *An Outline of the Constitutional Development of Malta Under British Rule* (1963); D. AUSTIN, *Malta and the End of Empire* (1971); H. FRENDO, *Party Politics in a Fortress Colony* (1978); PETER R. ELLIOT, *The Cross and the Ensign: A Naval History of Malta 1798–1978* (1980).

Landscape: B. BLOUET, *Malta: A Companion to the Landscape* (1963); R.H. BRYANS, *Malta and Gozo* (1966); "Malta: Past, Present and Future," special issue of the *Architectural Review,* vol. 146 (1969).

People: MALTA, CENTRAL OFFICE OF STATISTICS, *Annual Abstract of Statistics* (1947–); H. BOWEN-JONES, J.C. DEWDNEY, and W.B. FISHER, *Malta: Background for Development* (1962); J. BOISSEVAIN, *Hal-Farrug: A Village in Malta* (1969).

Economy: MALTA, CENTRAL OFFICE OF STATISTICS, *National Accounts of the Maltese Islands* (annual); CENTRAL BANK OF MALTA, *Annual Report and Statement of Accounts* (1968–); S. BUSUTTIL, *Fiscal Policy in Malta* (1965); O. DE LA GRANDVILLE, *Malte: Réalité géographique et perspectives économiques* (1968); W.F. STOLPER, R.E. HEILBERG, and S.O. CALLANDER, *Economic Adaptation and Development in Malta,* United Nations Commission for Technical Assistance, Dept. of Economic and Social Affairs (1964); M.M. METWALLY, *Structure and Performance of the Maltese Economy* (1977); J. GRECH, *Threads of Dependence* (1978).

Administration and social functions: P. CASSAR, *The Medical History of Malta* (1964); W.P. GULIA, *Local Government in Malta* (1967).

Cultural life and institutions: J. AQUILINA, *The Structure of Maltese: A Study in Mixed Grammar and Vocabulary* (1959); J.Q. HUGHES, *The Building of Malta During the Period of the Knights of St. John of Jerusalem, 1530–1795* (1956); U. ROLANDI, *Musica e musicisti in Malta* (1932); A.P. VELLA, *The University of Malta* (1969); R. ENGLAND, *Contemporary Art in Malta* (1974) and *Walls of Malta* (1976); D. MASSA (ed.), *Individual and Community in Commonwealth Literature* (1979).

Traveller's guides: THORNTON COX (ed.), *Malta* (1982); *Malta Travel Guide* (1982); INGE SEVERIN, *Malta and Gozo* (1983).

(S.Bu.)

Mammals

The class Mammalia is a group of vertebrate animals, collectively known as mammals, in which the young are nourished with milk from special secreting glands (mammae) of the mother. In addition to these characteristic milk glands, mammals are distinguished by several other unique features. Hair is a typical mammalian feature, although in many whales it has secondarily disappeared except in the fetal stage. The mammalian lower jaw is hinged directly to the skull, instead of through a separate bone (the quadrate) as in all other vertebrates. A chain of three tiny bones transmits sound waves across the middle ear. A muscular diaphragm separates the heart and the lungs from the abdominal cavity. Only the left aortic arch of the primitive fourth pair persists (in birds the right arch persists; in reptiles, amphibians, and fishes both arches are retained). Mature red blood cells in all mammals lack a nucleus; all other vertebrates have nucleated red blood cells.

Except for the monotremes (echidnas and duck-billed platypuses), which lay eggs, all mammals are viviparous (*i.e.*, bear live young). In the placental mammals (including man) the young are carried within the mother's womb, reaching a relatively advanced stage of development before being born. In the marsupials (kangaroos, opossums, and allies) the newborn, incompletely developed at birth, continue to develop outside the womb, attaching themselves to the female's body in the area of her mammary glands. Some marsupials have a pouchlike structure or fold, the marsupium, that shelters the suckling young.

The class Mammalia is worldwide in distribution. It has been said that mammals have a wider distribution and are more adaptable than any other single class of animals, with the exception of certain less complex forms such as the arachnids and insects. This versatility in exploiting the Earth is attributed in large part to the ability of mammals to regulate their body temperatures and internal environment both in excessive heat and aridity and in severe cold.

(Ed.)

The class Mammalia 382
- General features 382
  - Importance to man
  - Natural history
    - Reproduction
    - Behaviour
    - Ecology
    - Locomotion
    - Food habits
  - Form and function
    - The skin and hair
    - Dentition
    - Skeleton
    - Muscles
    - Viscera
- Evolution and classification 390
  - The evolution of the mammalian condition
  - Classification
    - Distinguishing taxonomic features
    - Annotated classification
    - Critical appraisal

Major mammal orders 396
- Monotremata (platypus, echidnas [spiny anteaters]) 396
- Marsupialia (kangaroos, bandicoots, phalangers, opossums, koala, wombats) 399
- Insectivora (shrews, moles, hedgehogs, tenrecs, solenodons) 407
- Chiroptera (bats) 413
- Primates (lemurs, lorises, tarsiers, monkeys, apes, and hominids, including man) 420
- Edentata (armadillos, sloths, anteaters) 435
- Lagomorpha (rabbits, hares, pikas) 441
- Rodentia (rats, mice, beavers, squirrels, guinea pigs, capybaras) 444
- Carnivora (cats, dogs, bears, skunks, seals, walruses) 455
- Cetacea (whales, dolphins, porpoises) 472
- Proboscidea (elephants) 477
- Sirenia (dugong, manatees) 481
- Perissodactyla (horses, asses, zebras, tapirs, rhinoceroses) 482
- Artiodactyla (pigs, goats, sheep, cattle, giraffes, deer, camels) 490

THE CLASS MAMMALIA

General features

The evolution of the Mammalia has produced tremendous diversity in form and habits. Living kinds range in size from tiny shrews, weighing but a few grams, to the largest of all animals that has ever lived, the blue or sulphur-bottom whale, which reaches a length of more than 30 metres (100 feet) and a weight of 136,000 kilograms (150 tons). Every major habitat has been invaded by mammals that swim, fly, run, burrow, glide, or climb.

There are approximately 4,000 species of living mammals, arranged in about 120 families and 20 orders. The rodents (order Rodentia) are the most numerous of existing mammals, both in number of species and number of individuals, and are one of the most diverse of living lineages. In contrast, the order Tubulidentata is represented by a single living species, the aardvark. The Proboscidea (elephants) and Perissodactyla (horses, rhinoceroses, and allies) are examples of orders in which far greater diversity occurred in mid- and late-Tertiary times (from about 3,000,000 to about 30,000,000 years ago) than today.

The tropical continental areas of the world around are the places of greatest present-day diversity of mammals, although members of the order occur on (or in seas adjacent to) all major land masses and on many oceanic islands (the latter principally, but by no means exclusively, reached by bats). Major regional faunas can be identified; these resulted in large part from evolution in comparative isolation of stocks of early mammals that reached these areas. South America (Neotropics), for example, was separated from North America (Nearctic) from Paleocene through much of Pliocene times (about 2,500,000 to 65,000,000 years ago) by inundation of the Panamanian Portal and adjacent areas in Middle America. Evolution of mammalian groups that had reached South America before the break between the continents, or some that "island-hopped" in after the break, proceeded quite independently from that of relatives that remained in North America. Some of the latter became extinct as the result of competition with more advanced groups, whereas those in South America flourished, some radiating to the extent that they have successfully competed with invaders since the rejoining of the two continents. Australia provides a parallel case of early isolation and adaptive radiation of mammals (monotremes, marsupials) thus isolated, but differs in that it was not later connected to any other land mass. The more advanced mammals of the infraclass Eutheria that reached Australia (rodents, bats) did

Radiation in South America

so, evidently by "island-hopping," long after the adaptive radiation of the early isolates.

In contrast, North America and Eurasia (Palearctic) are separate land masses but have closely related faunas, the result of connections several times in the Pleistocene and earlier across the Bering Strait. Their faunas frequently are thought of as representing not two distinct units, but one, related to a degree that a single name, Holarctic, is applied to it.

IMPORTANCE TO MAN

Wild and domesticated mammals are so interlocked with man's political and social history that it is impractical to attempt to assess the relationship in precise economic terms. Throughout his cultural evolution, for example, man has been dependent on other mammals for a significant portion of his food and clothing. Domestication of mammals helped to provide a source of protein for ever-increasing human populations and provided means of transportation and heavy work as well. Today, domesticated strains of the house mouse, European rabbit, guinea pig, hamster, gerbil and other species provide much-needed laboratory subjects for the study of human-related physiology, psychology, and a variety of diseases from dental caries to cancer. Recent emphasis on the study of nonhuman primates (monkeys and apes) has opened broad, new areas of research relevant to man's welfare. The care of domestic and captive mammals is, of course, the basis for the practice of veterinary medicine.

Some primitive peoples still depend on wild mammals as a major source of food, and many different kinds (from fruit bats and armadillos to whales) regularly are captured and eaten. On the other hand, the hunting, primarily for sport, of various rodents, lagomorphs, carnivores, and ungulates supports a multibillion-dollar enterprise. In the United States alone, for example, it is estimated that more than 2,000,000 deer are harvested annually by licensed hunters.

Influence on exploration

Geopolitically, the quest for marine mammals was responsible for the charting of a number of areas in both Arctic and Antarctic regions. The presence of terrestrial fur bearers, particularly beaver and several species of mustelid carnivores (*e.g.,* marten and fisher), was one of the principal motivations for the opening of the American West, Alaska, and the Siberian taiga. Sale of wild-taken furs still is an important industry in both the Old and New Worlds, but ranch-raised animals such as the mink, fox, and chinchilla now have become an important part of the fur industry, which directly and indirectly accounts for many millions of dollars in revenue each year in North America alone.

Aside from pelts and meat, special parts of some mammals regularly have been sought for their special attributes. The horns of rhinoceroses are used in concocting potions in the Orient; ivory from elephants and walruses is highly prized; and ambergris, a substance regurgitated by sperm whales, has been widely used as a base for perfumes (although now mostly replaced by synthetic substitutes).

Some mammals are directly detrimental to man and his activities. Murid rodents (house rats and mice of Old World origin) now occur in virtually all of the world's urban areas and each year cause substantial damage and economic loss. In some rural areas, herbivorous mammals eat or trample planted crops and compete with livestock for food, and native carnivores prey on domestic herds. These and other situations have led man to spend large sums annually to control populations of "undesirable" wild mammals, a practice long deplored by conservationists. Mammals are important reservoirs or agents of transmission of a variety of diseases that afflict man, such as plague, tularemia, yellow fever, rabies, leptospirosis, hemorrhagic fever, and Rocky Mountain spotted fever. The annual "economic debt" resulting from mammal-borne diseases of man and his domestic stock is incalculable.

Many large mammals that competed directly with man for food or space, or were specially sought by him for some reason, have been extirpated entirely or exist today only in parks and zoos; others are in danger of extinction, and their plight is receiving increased attention by a number of conservation agencies. Perhaps at least some can be saved. One of the most noteworthy cases of direct extirpation by man is the Steller's sea cow (*Hydrodamalis gigas*). These large (up to 12 metres, or 40 feet, long), inoffensive, marine mammals evidently lived in Recent times only along the coasts and shallow bays of the Commander Islands in the Bering Sea. Discovered in 1741, they were easily killed by Russian sealers and traders for food, their meat being highly prized, and the last known live individual was taken in 1768.

Extinction of Steller's sea cow

Of final note is the esthetic value of wild mammals and the relatively recent predilection of man to expend both considerable energy and resources to study and, if possible, conserve vanishing species, to set aside natural areas where native floral and faunal elements can exist in an otherwise highly agriculturalized or industrialized society, and to establish modern zoological parks and gardens. Such outdoor "laboratories" attract millions of visitors annually and will provide means by which present and future generations of humans can appreciate and study, in small measure at least, other kinds of mammals.

NATURAL HISTORY

The hallmarks of the mammalian level of organization are advanced reproduction and parental care, behavioral flexibility, and endothermy (the physiological maintenance of a relatively constant body temperature independent of that of the environment, allowing a high level of activity). Within the class, ecological diversity has resulted from adaptive specialization in food-getting, habitat preferences, and locomotion.

Throughout the past 70,000,000 years, the mammals have been the dominant animals in terrestrial ecosystems and important in nonterrestrial communities as well. The earliest mammals were small, active, predacious, and terrestrial or semiarboreal. From this primitive stock the mammals have radiated into a wide spectrum of adaptive modes against the background of the diverse environment of the Cenozoic (the last 65,000,000 years). Branches of the ancestral terrestrial stock early exploited the protection and productivity of the trees, whereas other lineages added further dimensions to the mammalian spectrum by adapting to life beneath the ground, in the air, and in marine and freshwater habitats.

Reproduction. In reproductively mature female mammals an interaction of hormones from the pituitary gland and the ovaries produces a phenomenon known as the estrous cycle. Estrus, or "heat," typically coincides with ovulation, and during this time the female is receptive to the male. Estrus is preceded by proestrus, during which ovarian follicles mature under the influence of a follicle-stimulating hormone from the anterior pituitary. The follicular cells produce estrogen, a hormone that stimulates proliferation of the uterine lining, or endometrium. Following ovulation, in late estrus, the ruptured ovarian follicle forms a temporary endocrine gland known as the corpus luteum. Another hormone, progesterone, secreted by the corpus luteum, causes the endometrium to become quiescent and ready for implantation of the developing egg (blastocyst), should fertilization occur. In members of the infraclass Eutheria (known as placental mammals) the placenta, as well as transmitting nourishment to the embryo, has an endocrine function, producing hormones that maintain the endometrium throughout gestation.

If fertilization and implantation do not occur, a phase termed metestrus ensues, in which the reproductive tract assumes its normal condition. Metestrus may be followed by anestrus, a nonreproductive period characterized by quiescence or involution of the reproductive tract. On the other hand, anestrus may be followed by a brief quiescent period (diestrus) and another preparatory proestrus phase. Mammals that breed only once a year are termed monestrous, and exhibit a long anestrus; those that breed more than once a year are termed polyestrous. In many polyestrous species the estrous cycle ceases during gestation and lactation (milk production), but some rodents have a postpartum estrus, mating immediately after giving birth.

The menstrual cycle of higher primates is derived from the estrous cycle, but differs from it in that when pro-

gesterone secretion from the corpus luteum ceases, in the absence of fertilization, the uterine lining is sloughed. In anthropoids other than man a distinct period of "heat" occurs around the time of ovulation.

Egg-laying by monotremes

Monotremes lay shelled eggs, but the ovarian cycle is similar to that of other mammals. The eggs are predominantly yolk (telolecithal), like those of reptiles and birds. Young monotremes are altricial (*i.e.,* in a relatively early stage of development and dependent upon the parent), reaching sexual maturity in about one year. (For additional information see below in *Monotremata,* subhead *Life cycle and reproduction.*)

The reproduction of marsupials differs from that of placentals in that the uterine wall is not specialized for the implantation of the embryos. The period of intrauterine development varies from about eight to 40 days. After this period the young migrate from the vagina to attach to the teats for further development. The pouch, or marsupium, is variously structured. Many species, such as kangaroos and opossums, have a single, well-developed pouch; in some phalangerids the pouch is compartmented, with a single teat in each compartment. The South American caenolestids or rat opossums have no marsupium. The young of most marsupials are dependent on maternal care for considerable periods, 13 to 14 weeks in *Didelphis.* Young koalas are carried in the pouch for nearly eight months, kangaroos to 10 months. (See below in *Marsupialia,* subheads *Life cycle* and *Reproductive adaptations.*)

Reproductive patterns in placental mammals are diverse, but in all cases a secretory phase is present in the uterine cycle and the endometrium is maintained by secretions of progesterone from the corpus luteum. The blastocyst emplants in the uterine wall. Villi are embedded in the lining of the uterus. The resulting complex of embryonic and maternal tissues is a true placenta. The uterine lining may be shed with the fetal membranes as "afterbirth" (a condition called deciduate) or may be resorbed by the female (nondeciduate). Placentae have been classified on the basis of the relationship between maternal and embryonic tissues. In the simplest nondeciduate placental arrangement the chorionic villi are in contact with uterine epithelium (the inner surface layer). In the "intimate deciduous" types, seen in primates, bats, insectivores, and rodents, the capillary endothelium (the layer containing minute blood vessels) of the uterine wall breaks down, and chorionic epithelium is in direct contact with maternal blood. In advanced stages of pregnancy in rabbits even the chorionic epithelium is eroded and the embryonic endothelium contacts the maternal blood supply. In no case, however, is there actual exchange of blood between mother and fetus; nutrients and gases must still pass through the walls of the fetal blood vessels.

The period of intrauterine development, or gestation, varies widely among eutherians, generally depending on the size of the animal, but influenced by the number of young per litter and the condition of young at birth. The gestation period of the domestic hamster (*Mesocricetus auratus*) is about two weeks, whereas that of the blue whale is 11 months and of the African elephant 21 to 22 months.

At birth the young may be well developed and able to move about at once (precocial) or they may be blind, hairless, and essentially helpless (altricial). In general, precocial young are born after a relatively long gestation period and in a small litter. Hares and many large grazing mammals bear precocial offspring. Rabbits, carnivores, and most rodents bear altricial young.

After birth young mammals are nourished by milk secreted by the mammary glands of the female. The development of milk-producing tissue in the female mammae is triggered by conception, and the stimulation of suckling the newborn prompts copious lactation. In therians (marsupials and placentals) the glands open through specialized nipples. The newborn young of marsupials are unable to suckle and milk is "pumped" to the young by the mother.

Composition of milk

Milk consists of fat, protein (especially casein), and lactose (milk sugar), as well as vitamins and salts. The actual composition of milk of mammals varies widely among species. The milk of whales and seals is some 12 times as rich in fats and four times as rich in protein as that of domestic cows but contains almost no sugar. Milk has meant an efficient energy source for the rapid growth of young mammals; the weight at birth of some marine mammals doubles in five days.

Behaviour. *Social behaviour.* The dependence of the young mammal on its mother for nourishment has made possible a period of training. Such training permits the nongenetic transfer of information between generations. The ability of young mammals to learn from the experience of their elders has allowed a behavioral plasticity unknown in any other group of organisms and has been a primary reason for the success of the mammals. The possibility of training is one of the factors that has made an increase in neural complexity selectively adaptive. Increased associational potential and memory extends the possibility of learning from experience, and the individual can make adaptive behavioral responses to environmental change. Individual response to short-term change is far more efficient than genetic response.

Some kinds of mammals are solitary, except for brief periods when the female is in estrus. In other species, however, social groups are present. Such groups may be reproductive, defensive, or may serve both functions. In those cases that have been studied in detail a more-or-less strict hierarchy of dominance prevails. Within the social group, maintenance of the hierarchy may depend on physical combat between individuals, but in many cases stereotyped patterns of behaviour evolve to displace actual combat, conserving energy, while maintaining the social structure.

Sexual dimorphism (a pronounced difference between sexes) frequently is extreme in social mammals. In large part this is because dominant males tend to be those that are largest or best armed. Dominant males also tend to have priority in mating, or may even have exclusive responsibility for mating within a "harem." Rapid evolution of secondary sexual characteristics can take place in a species with such a social structure even though the reproductive potential may be low.

A complex behaviour termed "play" frequently occurs between siblings, between members of an age class, or between parent and offspring. Play extends the period of maternal training and is especially important in social species, providing an opportunity to learn behaviour appropriate to the maintenance of dominance.

Territoriality. That area covered by an individual in his general activity is frequently termed the home range. A territory is a part of the home range defended against other members of the same species. As a generalization it may be said that territoriality is more important in the behaviour of birds than of mammals, but data for the latter are available primarily for diurnal species. The phenomenon of territoriality may be more widespread than is supposed. Frequently territories of mammals are "marked," either with urine or with secretions of specialized glands, as in lemurs, a form of territorial labelling less evident to humans than the singing or visual displays of birds. Many mammals that do not maintain territories per se nevertheless will not permit unlimited crowding and will fight to maintain individual distance. Such mechanisms result in more economical spacing of individuals over the available habitat.

Ecology. *Response to environmental cycles.* Mammals may react to environmental extremes by acclimatization, compensatory behaviour, or physiological specialization. Physiological responses to adverse conditions include torpidity, hibernation (in winter), and estivation (in summer). Torpidity may occur in the daily cycle or during unfavourable weather; short-term torpidity generally is economical only for small mammals that can cool and warm rapidly. The body temperature of most temperate-zone bats drops near that of the ambient air whenever the animal sleeps. The winter dormancy of bears at high latitudes is an analogous phenomenon and cannot be considered true hibernation.

Hibernation

True hibernation involves physiological regulation to minimize the expenditure of energy. The body temperature is lowered and breathing may be slowed to as low

as 1 percent of the rate in an active individual. There is a corresponding slowing of circulation and typically a reduction in the peripheral blood supply. When the body temperature nears the freezing point, spontaneous arousal occurs, although other kinds of stimuli generally elicit only a very slow response. In mammals that exhibit winter dormancy (such as bears, skunks, and raccoons), arousal may be quite rapid. Hibernation has evidently originated independently in a number of mammalian lines and the comparative physiology of this complex phenomenon is only now beginning to be understood (see BEHAVIOUR, ANIMAL: *Dormancy*).

Inactivity in response to adverse summer conditions (heat, drought, lack of food) is termed estivation. Estivation in some species is simply prolonged rest, usually in a favourable microhabitat; other estivating mammals regulate the metabolism, although the effects are typically not so pronounced as in hibernation.

Behavioral response to adverse conditions may involve the selection or construction of a suitable microhabitat (such as the cool, moist burrows of desert rodents). Migration is a second kind of behavioral response. The

Migrations

most obvious kind of mammalian migration is latitudinal. Many temperate-zone bats, for example, undertake extensive migrations, although other bat species hibernate near the summer foraging grounds in caves or other equable shelters during severe weather when insects are not available. Caribou (*Rangifer tarandus*) migrate from the tundra to the forest edge in search of suitable winter range and a number of cetaceans and pinnipeds undertake long migrations from polar waters to more temperate latitudes. Gray whales (*Eschrichtius robustus*), for example, migrate southward to calving grounds along the coasts of South Korea and Baja California from summer feeding grounds in the Okhotsk, Bering, and Chukchi seas. Of comparable extent is the dispersive feeding migration of the northern fur seal (*Callorhinus ursinus*).

Migrations of lesser extent include the altitudinal movements of some ungulates, the American elk or wapiti (*Cervus canadensis*) and bighorn sheep (*Ovis canadensis*), for example, and the local migrations of certain bats from summer roosts to hibernacula. Most migratory patterns of mammals are part of a recurrent annual cycle, but the irruptive emigrations of lemmings and snowshoe hares are largely acyclic responses to population pressure on food supplies.

Populations. A population consists of individuals of three "ecological ages"—prereproductive, reproductive, and postreproductive. The structure and dynamics of a population depend, among other things, on the relative lengths of these ages, the rate of recruitment of individuals (either by birth or by immigration), and the rate of emigration or death. The reproductive potential of some rodents (particularly muroids) is well known; some mice are reproductively mature at four weeks of age, have gestation periods of three weeks or less, and may experience post-partum estrus, with the result that pregnancy and lactation may overlap. Litter size, moreover, may average four or more, and breeding may occur throughout the year in favourable localities. The reproductive potential of a species is, of course, a theoretical maximum that is rarely met inasmuch as, among other reasons, a given female typically does not reproduce throughout the year. Growth of a population depends on the survival of individuals to reproductive age. The absolute age at sexual maturity ranges from less than four weeks in some rodents to some 15 years in the African elephant (*Loxodonta africana*).

Longevity

Postreproductive individuals are rare in most mammalian populations. Survival through more than a single reproductive season is probably uncommon in many small kinds, such as mice and shrews. Larger species typically have longer life spans than do smaller kinds, but some bats are known, on the basis of banding records, to live nearly 20 years. Many species show greater longevity in captivity than in the wild. Captive spiny anteaters (*Tachyglossus*) are reported to have lived more than 50 years. Horses are known to live more than 60 years, and the oldest known elephant lived to be nearly 70. Man has greater potential longevity than any other known species.

Locomotion. Specialization in habitat preference has been accompanied by locomotor adaptations. Convergent evolution within a given adaptive mode has contributed to the ecological similarity of regional mammalian faunas. Terrestrial mammals have a number of modes of progression. The primitive mammalian stock was doubtless ambulatory and plantigrade, walking with the digits, metacarpals, and metatarsals (bones of the midfoot), and parts of the ankle and wrist in contact with the ground. The limbs of ambulatory mammals are typically mobile, capable of considerable rotation.

Mammals modified for running are termed cursorial. The stance of cursorial species may be digitigrade (the complete digits contacting the ground, as in dogs) or unguligrade (only tips of digits contacting the ground, as in horses). In advanced groups limb movement consists of a single forward and backward direction.

Saltatory (leaping) locomotion, sometimes called "ricochetal," has arisen in several unrelated groups, and typically is found in mammals of open habitats. Jumping mammals (some marsupials, lagomorphs, and several independent lineages of rodents) typically have elongate, plantigrade hind feet, reduced forelimbs, and long tails.

Mammals of several orders have attained great size and have converged on specializations for "graviportal" (ponderous) locomotion. Such kinds have no digit reduction and deploy the digits in a circle around the axis of the limb for maximum support, like the pedestal of a column.

The bats are the only truly flying mammals. Only with active flight have the resources of the aerial habitat been successfully exploited. Mammals of several kinds (dermopterans, marsupials, rodents) are adapted for gliding. A gliding habit is frequently accompanied by scansorial (climbing) locomotion. Many nongliders, such as tree squirrels, are also scansorial.

Well-adapted arboreal mammals frequently are plantigrade, five-toed, and equipped with highly mobile limbs. Many New World monkeys have a prehensile tail, which is used as a "fifth hand." Gibbons have reduced the primitive, opposable anthropoid thumb as a specialization for brachiation, or "arm-walking," in which the animal hangs from branches and moves by a series of long swings. The highly arboreal tarsiers have expanded pads on the digits to improve the grasp, and many other arboreal kinds have claws (sloths) or well-developed nails.

Several groups of mammals have independently assumed aquatic habits. In some cases semi-aquatic or aquatic mammals are relatively unmodified representatives of otherwise terrestrial groups (for example, otters, muskrats, water shrews). Other kinds have undergone profound modification for natatorial locomotion and a pelagic habit. Pinniped carnivores (walruses and seals) give birth to their young on land, but cetaceans are completely helpless out of water, on which they depend for mechanical support and thermal insulation.

Food habits. The earliest mammals, like their reptilian forebears, were active predators. From such a basal stock there has been a complex radiation of trophic adaptations. Modern mammals occupy a wide spectrum of feeding niches. In most terrestrial and some aquatic communities, carnivorous mammals are at the top of the food pyramid. There are also mammalian primary consumers in most ecosystems. The voracious shrews, smallest of mammals, sometimes prey on vertebrates larger than themselves. They may eat twice their weight in food each day to maintain their active metabolism and compensate for heat loss caused by an unfavourable surface-to-volume ratio. The largest of vertebrates, the blue whale, feeds on krill, minute planktonic crustaceans.

Within a given lineage the adaptive radiation in food habits may be broad. Some of the Carnivora have become omnivorous (raccoons, bears) or largely herbivorous (giant panda). Marsupials exhibit a great variety of feeding types, and in Australia marsupials have radiated to fill ecological niches highly analogous to those of placental mammals elsewhere; there are marsupial "moles," "anteaters," "mice," "rats," "cats," and "wolves." Some bandicoots have ecological roles similar to those of rabbits, and wombats are semifossorial (*i.e.*, partially burrowing) herbivores

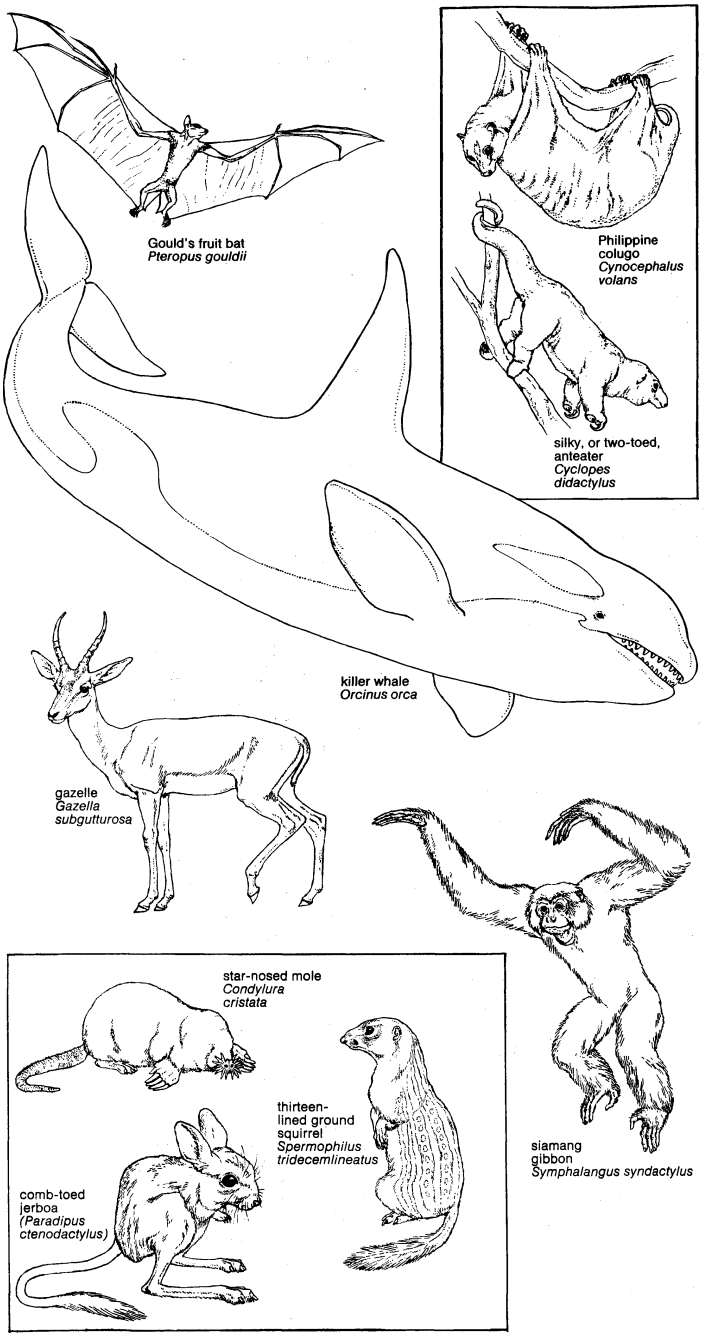


Figure 1: Range of body plans in mammals.

Drawing by R. Keane based on photographs courtesy of (Gould's fruit bat) J. Warham, (Philippine colugo) J.N. Hamlet, (silky, or two-toed anteater) New York Zoological Society, (comb-toed jerboa) Zeitschrift fur Saugetierkunde

analogous to marmots. In Australia the niche of large grazers and browsers is filled by a variety of kangaroos and wallabies.

Radiation of bats

Within the bats there has also been a remarkable adaptive radiation in food habits. Early in the history of the order there evidently was a divergence into insectivorous and frugivorous lines. The frugivorous line (Megachiroptera) has generally maintained its fruit-eating habit, although some have become rather specialized nectar-feeders. The Microchiroptera have been less conservative and have undergone considerable divergence in feeding habits. A majority of living microchiropterans are insectivorous, but members of two different families have become fish eaters. Within the large neotropical family Phyllostomatidae there are groups specialized to feed on fruit, nectar, insects, and small vertebrates (including other bats). Aberrant members of the family are the vampire bats (desmodontines) with a specialized dentition to aid the blood-lapping habit.

FORM AND FUNCTION

The skin and hair. The skin of mammals is constructed of two layers, a superficial nonvascular epidermis and an inner layer, the dermis or corium. The two layers interdigitate in dermal papillae (fingerlike projections), ridges of sensitive vascular dermis projecting into the epidermis. The outermost layers of the epidermis are cornified (*i.e.,* impregnated with various tough proteins) and enucleate (lacking cell nuclei). The epidermis is composed of flattened (squamous) cells in layers and is the interface between the individual and the environment. Its primary function is defensive, and it is cornified to resist abrasion. The surface of the skin is coated with lipids and organic salts, the so-called "acid mantle," thought to have antifungal and antibacterial properties. Deep in the epidermis is an electronegative layer, a further deterrent to foreign organic or ionic agents.

The dermis lies beneath the epidermis and nourishes it. The cutaneous circulation of the dermis is variously developed in mammals but it is typically extensive, out of proportion to the nutritional needs of the tissue. The major role of the cutaneous circulation is to moderate body temperature and blood pressure by forming a peripheral shunt, an alternate route for the blood. Also in the dermis are sensory nerve endings to alert the individual to pressure (touch), heat, cold, and pain. In general, skin bearing hairs has few or no specialized sensory endings, but hairless skin, such as the lips and fingertips of man, has specialized endings. The sensation of touch on hairy skin in man depends on stimulation of the nerve fibres of the hairs.

Hair is derived from an invagination (pocketing) of the epidermis termed a follicle. Collectively, the hair is called the pelage. The individual hair is a rod of keratinized cells that may be cylindrical or more-or-less flattened. Keratin is a protein also found in claws and nails. The inner medulla of the hair is hollow and contains air; in the outer cortex layer there are frequently pigment granules. Associated with the hair follicle are nerve endings and a muscle, the arrector pili. The latter allows the erection of individual hairs to alter the insulative qualities of the pelage. The follicle also gives rise to sebaceous glands that produce sebum, a substance that lubricates the hair.

Types of hair

Most mammals have three distinct kinds of hairs. Guard hairs protect the rest of the pelage from abrasion and frequently from moisture, and usually lend a characteristic colour pattern. The thicker underfur is primarily insulative and may differ in colour from the guard hairs. The third common hair-type is the vibrissa, or whisker, a stiff, typically elongate, hair that functions in tactile sensation. Hairs may be further modified to form rigid quills. The "horn" of the rhinoceros is composed of a fibrous keratin material derived from hair. Examples of keratinized derivatives of the integument other than hair are horns, hooves, nails, claws, and baleen.

Even though the primary function of the skin is defensive, it has been modified in mammals to serve such diverse functions as thermoregulation and nourishment of young. Secretions of sweat glands promote cooling due to evaporation at the surface of the body, and mammary glands are thought to be derived from sweat glands.

In certain groups (primates in particular) the skin of the face is under intricate muscular control and movements of the skin express and communicate "emotion." In many mammals the colour and pattern of the pelage is important in communicative behaviour. Patterns may be dymantic (startling, such as the mane of the male lion or hamadryas baboon), sematic (warning, such as the bold pattern of skunks), or cryptic (concealing), perhaps the most common adaptation of pelage colour.

Hair has been secondarily lost or considerably reduced in some kinds of mammals. In adult cetaceans insulation is provided by thick subcutaneous fat deposits, or blubber, with hair limited to a few stiff vibrissae about the mouth. The naked skin of whales is one of a number of features that contribute to the remarkably advanced hydrodynamics of locomotion in the group. Some fossorial mammals also tend toward reduction of the hair. This is shown most strikingly by the sand rats (*Heterocephalus*) of northeast

Africa, but considerable loss of hair has also occurred in some species of pocket gophers (Geomyidae). Hair may also be lost on restricted areas of the skin, as from the face in many monkeys, or the buttocks of mandrills, and may be sparse on elephants and such highly modified kinds as pangolins and armadillos.

Indeterminate (continuous) growth, as seen in the hair of the head in man, is rare among mammals. Hairs with determinate growth are subject to wear and must be replaced periodically—a process termed molt. The first coat of a young mammal is referred to as the juvenal pelage, which typically is of fine texture like the underfur of adults and is replaced by a postjuvenile molt. Juvenal pelage is succeeded by the subadult pelage, which in some species is not markedly distinct from that of the adult, or directly by adult pelage. Once this pelage is acquired molting continues to recur at intervals, often annually or semi-annually, sometimes more frequently. The pattern of molt typically is orderly, but varies widely between species. Some mammals apparently molt continuously, with a few hairs replaced at a time throughout the year.

Dentition. Specialization in food habits has led to profound dental changes. The primitive mammalian tooth had high, sharp cusps and served to tear flesh or crush chitinous material (primarily the exoskeletons of terrestrial arthropods, such as insects). Herbivores tend to have specialized cheek teeth with complex occlusal (contact) patterns and various ways of expanding the crown and circumventing the problem of wear. Omnivorous mammals, such as bears, pigs, and man, tend to have molars with low, rounded cusps, termed "bunodont."

Anteating

A prime example of convergence in conjunction with dietary specialization is seen in those mammals adapted to feeding on social insects (generally termed myrmecophagy, "anteating"). This habit has led to remarkably similar morphology in such diverse groups as the monotreme echidnas (Tachyglossidae), one dasyurid marsupial (Myrmecobiinae), some edentates (Myrmecophagidae) the aardvark, and pangolins (Pholidota). Trends frequently associated with myrmecophagy include: strong claws, an elongate, terete skull, a vermiform, extensible tongue, marked reduction in the mandible, and loss or extreme simplification of the dentition.

Specialized herbivores evolved early in mammalian history. The earliest (and with the longest evolutionary history) were the multituberculates. Evolutionary convergence in teeth (*i.e.,* resemblances not due to common ancestry) has occurred widely in herbivorous groups; most have incisors modified for nipping or gnawing, have lost teeth with the resultant development of a gap (diastema) in the tooth row, and exhibit some molarization (expansion and flattening) of premolars to expand the grinding surface of the cheek teeth. Rootless incisors or cheek teeth have evolved frequently, an open pulp cavity allowing continual growth throughout life. Herbivorous specializations have evolved independently in multituberculates, rodents, lagomorphs, primates, and in the wide radiation of ungulate and subungulate orders.

Skeleton. The mammalian skeletal system shows a number of advances over that of lower vertebrates. The mode of ossification of the long bones is characteristic. In lower vertebrates each long bone has a single centre of ossification, the diaphysis, and replacement of cartilage by bone proceeds from the centre toward the ends, which may remain cartilaginous, even in adults. In mammals secondary centres of ossification, the epiphyses, develop at the ends of the bones. Growth of bones occurs in zones of cartilage between diaphysis and epiphyses. Mammalian skeletal growth is termed determinate, for once the actively growing zone of cartilage is obliterated, growth in length ceases. As in all bony vertebrates, of course, there is continual renewal of bone throughout life. The advantage of epiphyseal ossification lies in the fact that the bones have strong articular surfaces before the skeleton is mature. In general, the skeleton of the adult mammal has less structural cartilage than does that of a reptile.

The skeletal system of mammals and other vertebrates is broadly divisible functionally into axial and appendicular portions. The axial skeleton consists of the braincase (cranium) and the backbone and ribs, and serves primarily to protect the central nervous system. The limbs and their girdles constitute the appendicular skeleton. In addition, there are skeletal elements derived from the gill arches of primitive vertebrates, collectively termed the visceral skeleton. Visceral elements in the mammalian skeleton include the jaws, the hyoid apparatus supporting the tongue, and the auditory ossicles of the middle ear. The postcranial axial skeleton in mammals generally has remained rather conservative during the course of evolution. The vast majority of mammals have seven cervical (neck) vertebrae; exceptions are sloths, with six or nine cervicals, and the Sirenia with six. The anterior two cervical vertebrae are differentiated as atlas and axis. Specialized articulations of these two bones allow complex movements of the head on the trunk. Thoracic vertebrae bear ribs and are variable in number. The anterior ribs converge toward the ventral midline to articulate with the sternum, or breastbone, forming a semirigid thoracic "basket" for the protection of heart and lungs. Posterior to the thoracic region are the lumbar vertebrae, ranging from two to 21 in number (most frequently four to seven). Mammals have no lumbar ribs. There are usually three to five sacral vertebrae, but some edentates have as many as 13. Sacral vertebrae fuse to form the sacrum, to which the pelvic girdle is attached. Caudal (tail) vertebrae range in number from five (fused elements of the coccyx of man) to 50.

The basic structure of the vertebral column is comparable throughout the Mammalia, although in many instances modifications have occurred in specialized locomotor modes to gain particular mechanical advantages. The vertebral column and associated muscles of many mammals is structurally analogous to a cantilever girder.

The skull

The skull is composite in origin and complex in function. Functionally the bones of the head are separable into the braincase and the jaws. In general, it is the head of the animal that meets the environment. The skull protects the brain and sense capsules, houses the teeth and tongue, and the entrance to the pharynx. Thus the head functions in sensory reception, food acquisition, defense, respiration, and (in higher groups) communication. To serve these functions, bony elements have been recruited from the visceral skeleton, the endochondral skeleton, and from the dermal skeleton of lower vertebrates.

The skull of mammals differs markedly from that of reptiles because of the great expansion of the brain. The sphenoid bones that form the reptilian braincase form only the floor of the braincase in mammals. The side is formed in part by the alisphenoid bone, derived from the epipterygoid, a part of the reptilian palate. Dermal elements, the frontals and parietals, have come to lie deep to (beneath) the muscles of the jaw to form the dorsum of the braincase. Reptilian dermal roofing bones, lying superficial to the muscles of the jaw, are represented in mammals only by the jugal bone of the zygomatic arch, which lies under the eye.

In mammals, a secondary palate is formed by processes of the maxillary bones and the palatines, with the pterygoid bones reduced in importance. The secondary palate separates the nasal passages from the oral cavity and allows continuous breathing while chewing or suckling.

Other specializations of the mammalian skull include paired occipital condyles (the articulating surfaces at the neck) and an expanded nasal chamber with complexly folded turbinal bones, providing a large area for detection of odours. Eutherians have evolved bony protection for the middle ear, the auditory bulla. The development of this structure varies, although an annular (ring-shaped) tympanic bone is always present.

Bones of the ear

The bones of the mammalian middle ear are a diagnostic feature of the class. The three auditory ossicles form a series of levers that serve mechanically to increase the amplitude of sound waves reaching the tympanic membrane, or eardrum, produced as disturbances of the air. The innermost bone is the stapes, or "stirrup bone." It rests against the oval window of the inner ear. The stapes is homologous with the entire stapedial structure of reptiles, which, in turn, was derived from the hyomandibular arch of primitive vertebrates. The incus, or "anvil," ar-

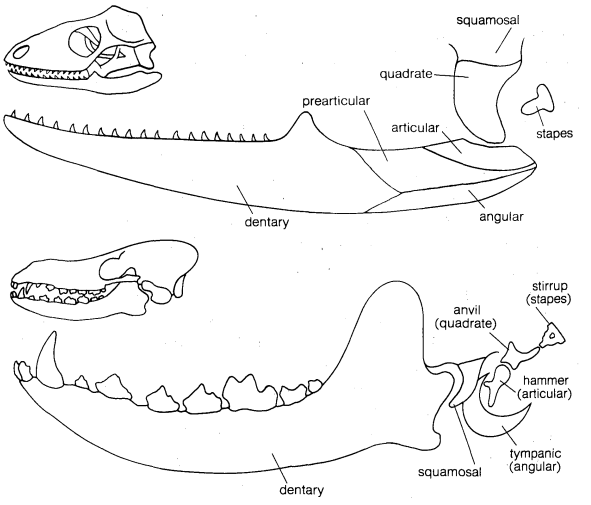


Figure 2: Comparison of lower jaw and ear region in the skull of (top) a reptile and (bottom) a mammal.

ticulates with the stapes. The incus was derived from the quadrate bone, which is involved in the jaw articulation in reptiles. The malleus, or "hammer," rests against the tympanic membrane and articulates with the incus. The malleus is the homologue of the reptilian articular bone. The mechanical efficiency of the middle ear has thus been increased by the incorporation of two bones of the reptilian jaw assemblage. In mammals the lower jaw is a single bone, the dentary, which articulates with the squamosal of the skull.

The limbs and girdles have been greatly modified with locomotor adaptations. The ancestral mammal had well-developed limbs and was five-toed. In each limb there were two distal (outer) elements (radius and ulna in the forelimb; tibia and fibula in the hindlimb) and a single proximal (inner or upper) element (humerus; femur). There were nine bones in the wrist, the carpals; and seven bones in the ankle, the tarsals. The phalangeal formula (the number of phalangeal bones in each digit, numbered from inside outward) is 2-3-3-3-3 in primitive mammals; in primitive reptiles it is 2-3-4-5-3. Modifications in mammalian limbs have involved reduction, loss, or fusion of bones. Loss of the clavicle from the shoulder girdle, reduction in the number of toes, and modifications of tarsal and carpal bones are typical correlates of cursorial locomotion. Scansorial and arboreal groups tend to maintain or emphasize the primitive divergence of the thumb and hallux (the inner toe on the hindfoot). Some details of limb modifications are included in the synopsis of the orders of mammals (see below *Annotated classification*).

Centres of ossification sometimes develop in nonbony connective tissue. Such bones are termed heterotopic or sesamoid elements. The kneecap (patella) is such a bone. Another important bone of this sort, found in many kinds of mammals, is the baculum, or os penis, which occurs as a stiffening rod in the penis of such groups as carnivores, many bats, rodents, some insectivores, and many primates. The os clitoridis is a homologous structure found in females.

Muscles. The muscular system of mammals is generally comparable to that of reptiles. With changes in locomotion, the proportions and specific functions of muscular elements have been altered, but the relationships of these muscles remain essentially the same. Exceptions to this generalization are the muscles of the skin and of the jaw.

The panniculus carnosus is a sheath of dermal (skin) muscle, developed in many mammals, which allows the movement of the skin independent of the movement of deeper muscle masses. Such movement functions in such mundane activities as the twitching of the skin to foil insect pests and in some species also is important in shivering, a characteristic heat-producing response to thermal stress. The dermal musculature of the facial region is particularly well developed in primates and carnivores, but occurs in other groups as well. Facial mobility allows expression that may be of importance in the behavioral maintenance of interspecific social structure.

Muscles of the jaw

The temporalis muscle is the major adductor (closer) of the reptilian jaw. In mammals the temporalis is divided into a deep temporalis proper and a more superficial masseter muscle. The temporalis attaches to the coronoid process of the mandible (lower jaw) and the temporal bone of the skull. The masseter passes from the angular process of the mandible to the zygomatic arch. The masseter allows an anteroposterior (forward-backward) movement of the jaw and is highly developed in mammals, such as rodents, in which grinding is the important function of the dentition.

Viscera. *Digestive system.* The alimentary canal is highly specialized in many kinds of mammals. In general, specializations of the gut accompany herbivorous habits. The intestines of herbivores are typically elongate and the stomach may also be specialized. Subdivision of the gut allows areas of differing physiological environments for the activities of different sorts of enzymes and symbiotic bacteria (which aid the animal by breaking down certain "indigestible" compounds). In ruminant artiodactyls the stomach has up to four chambers, each with a particular function in the processing of vegetable material. A cecum is common in many herbivores. The cecum is a blind sac at the distal end of the small intestine where complex compounds like cellulose are acted upon by symbiotic bacteria. The vermiform appendix is a diverticulum of the cecum. The appendix is rich in lymphoid tissue and in many mammals is concerned with defense against toxic bacterial products.

Hares and rabbits, the sewellel or "mountain beaver" (*Aplodontia rufa*), and some insectivores exhibit a phenomenon, known as reingestion, in which at intervals specialized fecal pellets are produced, which are taken in the mouth and passed through the alimentary canal a second time. Where known to be present, this pattern seems to be obligatory. Reingestion is poorly understood, but it is generally supposed that the process allows the animal to absorb in the upper gut vitamins produced by the microflora of the lower gut but not absorbable there.

Excretory system. The mammalian kidney is constructed of a large number of functional units called nephrons. Each nephron consists of a distal tubule, a medial section termed the loop of Henle, a proximal tubule, and a renal corpuscle. The renal corpuscle is a knot of capillaries, called a glomerulus, surrounded by a sheath, Bowman's capsule. The renal corpuscle is a pressure filter, relying on blood pressure to remove water, ions, and small organic molecules from the blood. Some of the material removed is waste, but some is of value to the organism. The filtrate is sorted by the tubules, and water and needed solutes are reabsorbed. Reabsorption is both passive (osmotic) and active (based on ion transport systems). The distal convoluted tubules drain into collecting tubules which, in turn, empty into the calyces, or branches, of the renal pelvis, the expanded end of the ureter. The pressure-pump nephron of mammals is so efficient that the renal portal system of lower vertebrates has been completely lost. Mammalian kidneys show considerable variety in structure, relative to the environmental demands on a given species. In particular, desert rodents have long loops of Henle and are able to reabsorb much water and to excrete a highly concentrated urine. Urea is the end product of protein metabolism in mammals, and excretion is therefore called ureotelic.

Action of the kidney

Reproductive system. The testes of mammals descend from the abdominal cavity to lie in a compartmented pouch termed the scrotum. In some species the testes are permanently scrotal and the scrotum is sealed off from the general abdominal cavity. In others the testes migrate to the scrotum only during the breeding season. It is thought that the temperature of the abdominal cavity is too high to allow spermatogenesis; the scrotum allows cooling of the testes.

The transport of spermatozoa is comparable to that in reptiles, relying on ducts derived from urinary ducts of earlier vertebrates. Mammalian specialities are the bulbourethral (or Cowper's) glands, the prostate gland, and

the seminal vesicle or vesicular gland. Each of these glands adds secretions to the spermatozoa to form semen, which passes from the body via a canal (urethra) in the highly vascular, erectile penis. The tip of the penis, the glans, may have a complex morphology and has been used as a taxonomic character in some groups. The penis may be retracted into a sheath along the abdomen or may be pendulous, as in bats and many primates.

Types of uterus

The structure of the female reproductive tract is variable. Four uterine types are generally recognized among placentals, based on the relationship of the uterine horns (branches). A duplex uterus characterizes rodents and rabbits; the uterine horns are completely separated and have separate cervices opening into the vagina. Carnivores have

From (A) A.S. Romer, *The Vertebrate Body*, 4th ed. (1970); W.B. Saunders Company, Philadelphia; (B, C) W.F. Walker, *Vertebrate Dissection*, 4th ed. (1970); W.B. Saunders Company, Philadelphia

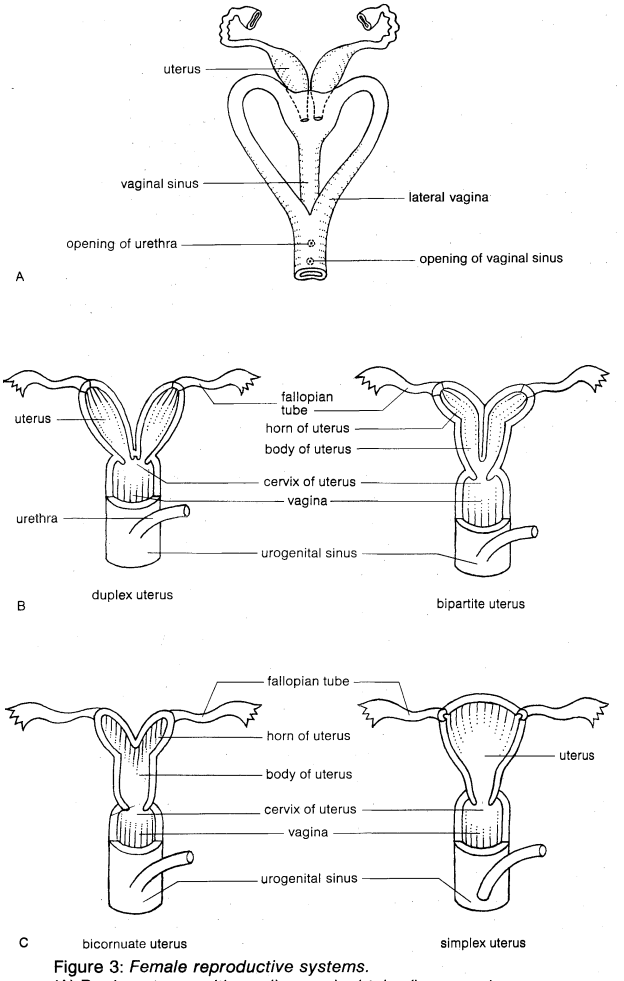


Figure 3: *Female reproductive systems.*
(A) Duplex uterus with median vaginal tube (kangaroo).
(B, C) Placental uteri.

a bipartite uterus, in which the horns are largely separate but enter the vagina by a single cervix. In the bicornate uterus, typical of many ungulates, the horns are distinct for less than half their length; the lower part of the uterus is a common chamber, the body. Higher primates have a simplex uterus in which all separation between the horns is lacking, forming a single chamber.

The female reproductive tract of marsupials is termed didelphous; the vagina is paired, as are oviducts and uteri. In primitive marsupials there are paired vaginae lateral to the ureters. In more advanced groups, such as kangaroos, the lateral vaginae persist and conduct the migration of spermatozoa, but a medial "pseud-vagina" functions as the birth canal.

Monotremes have paired uteri and oviducts, which empty into a urogenital sinus (cavity) as do fluid wastes. The sinus passes into the cloaca, a common receptacle for reproductive and excretory products.

Circulatory system. In mammals as in birds, right and left ventricles of the heart are completely separated, so that pulmonary (lung) and systemic (body) circulations are completely independent. Oxygenated blood arrives in the left atrium from the lungs and passes to the left ventricle, whence it is forced through the aorta to the systemic circulation. Deoxygenated blood from the tissues returns to the right atrium via a large vein, the vena cava, and is pumped to the pulmonary capillary bed through the pulmonary artery.

Among vertebrates, contraction of the heart is myogenic; rhythm is inherent in all cardiac muscle, but in myogenic hearts the pacemaker is derived from cardiac tissue. The pacemaker in mammals (and also in birds) is an oblong mass of specialized cells called the sinoatrial node, located in the right atrium near the junction with the venae cavae. A wave of excitation spreads from this node to the atrioventricular node, which is located in the right atrium near the base of the interatrial septum. From this point excitation is conducted along the atrioventricular bundle (bundle of His) and enters the main mass of cardiac tissue along fine branches, the Purkinje fibres. Homeostatic control of the heart by neuroendocrine or other agents is mediated through the intrinsic control network of the heart.

Blood leaves the left ventricle through the aorta. The mammalian aorta is an unpaired structure derived from the left fourth aortic arch of the primitive vertebrate. Birds, on the other hand, retain the right fourth arch.

Characteristics of the blood

The circulatory system forms a complex communication and distribution network to all physiologically active tissues of the body. A constant, copious supply of oxygen is required to sustain the active, endothermous physiology of the higher vertebrates. The efficiency of the four-chambered heart is important to this function. Oxygen is transported by specialized red blood cells, or erythrocytes, as in all vertebrates. Packaging the oxygen-bearing pigment hemoglobin in erythrocytes keeps the viscosity of the blood minimal and thereby allows efficient circulation while limiting the mechanical load on the heart. The mammalian erythrocyte is a highly evolved structure; its discoid, biconcave shape allows maximal surface area per unit volume. When mature and functional, mammalian red blood cells are enucleate.

Respiratory system. Closely coupled with the circulatory system is the ventilatory (breathing) apparatus, the lungs and associated structures. Ventilation in mammals is unique. The lungs themselves are less efficient than those of birds, for air movement consists of an ebb and flow, rather than a one-way circuit, so a residuum of air remains that cannot be expired. Ventilation in mammals is by means of a negative pressure pump made possible by the development of a definitive thoracic cavity with the evolution of the diaphragm.

The diaphragm is a unique, composite structure consisting of (1) the transverse septum (a wall that primitively separates the heart from the general viscera); (2) pleuroperitoneal folds from the body wall; (3) mesenteric folds; and (4) axial muscles inserting on a central tendon, or diaphragmatic aponeurosis.

The lungs lie in separate, airtight compartments called pleural cavities, separated by the mediastinum. As the size of the pleural cavity is increased the lung is expanded and air flows in passively. Enlargement of the pleural cavity is produced by contraction of the diaphragm or by elevation of the ribs. The relaxed diaphragm domes upward, but when contracted it stretches flat. Expiration is an active movement brought about by contraction of abdominal muscles against the viscera.

Air typically enters the respiratory passages through the nostrils where it may be warmed and moistened. It passes above the bony palate and the soft palate and enters the pharynx. In the pharynx the passages for air and food cross. Air enters the trachea, which divides at the level of the lungs into primary bronchi. A characteristic feature of the trachea of many mammals is the larynx. Vocal cords stretch across the larynx and are vibrated by forced expiration to produce sound. The laryngeal apparatus may be greatly modified for the production of complex vocalizations. In some groups, for example, howler monkeys

(*Alouatta*), the hyoid apparatus is incorporated into the sound-producing organ, as a resonating chamber.

Nervous and endocrine systems. The nervous system and the system of endocrine glands are closely related to one another in their function, for both serve to coordinate activity. The endocrine glands of mammals generally have more complex regulatory functions than do those of lower vertebrates. This is particularly true of the pituitary gland, which supplies hormones that regulate the reproductive cycle. Follicle stimulating hormone (FSH) initiates the maturation of the ovarian follicle. Luteinizing hormone (LH) mediates the formation of the corpus luteum from the follicle following ovulation. Prolactin, also a product of the anterior pituitary, stimulates the secretion of milk.

Control of the pituitary glands is partially by means of neurohumours from the hypothalamus, a part of the forebrain in contact with the pituitary gland by nervous and circulatory pathways. The hypothalamus is of the utmost importance in mammals, for it integrates stimuli from both internal and external environments, channelling signals to higher centres or into autonomic pathways.

The brain

The cerebellum of vertebrates is at the anterior end of the hindbrain. Its function is to coordinate motor activities and to maintain posture. In most mammals the cerebellum is highly developed and its surface may be convoluted to increase its area. The data with which the cerebellum works arrive from proprioceptors ("self-sensors") in the muscles and from the membranous labyrinth of the inner ear, the latter giving information on position and movements of the head.

In the vertebrate ancestors of the mammals the cerebral hemispheres were centres for the reception of olfactory stimuli. Vertebrate evolution has favoured an increasing importance of these lobes in the integration of stimuli. Their great development in mammals as centres of association is responsible for the "creative" behaviour of members of the class; *i.e.,* the ability to learn, to adapt as individuals to short-term environmental change through appropriate responses on the basis of previous experience. In vertebrate evolution the gray matter of the cerebrum has moved from a primitive internal position in the hemispheres to a superficial position. The superficial gray matter is termed the pallium. The paleopallium of amphibians has become the olfactory lobes of the higher vertebrates; the dorsolateral surface, or archipallium, the mammalian hippocampus. The great neural advance of the mammals lies in the elaboration of the neopallium, which makes up the bulk of the cerebrum. The neopallium is an association centre, the dominant centre of neural function, and is involved in so-called "intelligent" response. By contrast, the highest centre in the avian brain is the corpus striatum, an evolutionary product of the basal nuclei of the amphibian brain. The bulk of the complex behaviour of birds is instinctive. The surface of the neopallium tends in some mammals to be greatly expanded by convoluting, forming folds (gyri) between deep grooves (sulci).

Evolution and classification

THE EVOLUTION OF THE MAMMALIAN CONDITION

Mammals were derived in the Triassic Period from members of the reptilian order Therapsida. The therapsids, members of the subclass Synapsida (sometimes called the mammal-like reptiles), generally were unimpressive in relation to other reptiles of their time. Synapsids were present in the Carboniferous Period (about 280,000,000 to 345,000,000 years ago) and are one of the earliest known reptilian groups. They were the dominant reptiles of Permian times (about 225,000,000 to 280,000,000 years ago), and although they were primarily predacious in habit, the adaptive radiation included herbivorous species as well. In the Mesozoic (about 225,000,000 to 65,000,000 years ago), the importance of the synapsids was generally assumed by the archosaurs or "ruling reptiles," the therapsids, in general, being small, active carnivores. Therapsids tended to evolve a specialized heterodont dentition and to improve the mechanics of locomotion by bringing the plane of action of the limbs close to the trunk. A secondary palate was developed and the temporal musculature was expanded.

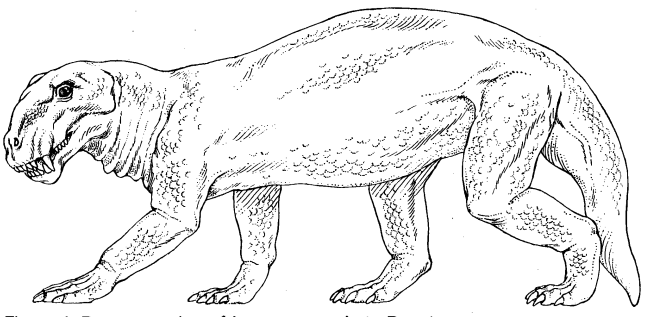


Figure 4: Reconstruction of *Lycaenops,* a Late Permian mammal-like reptile.

Drawing by R. Keane based on A.S. Romer, *The Vertebrate Story,* copyright 1959; the University of Chicago Press

The several features that separate modern reptiles from modern mammals doubtless evolved at different rates. Many attributes of mammals are correlated with their highly active habit; for example, efficient double circulation with a completely four-chambered heart, anucleate and biconcave erythrocytes, the diaphragm, and the secondary palate (which separates passages for food and air and allows breathing during mastication or suckling). Hair for insulation is a correlate of endothermy, the physiological maintenance of individual temperature independent of environmental temperature, and endothermy allows high levels of sustained activity. The unique characteristics of mammals thus would seem to have evolved as a complex interrelated system.

Mosaic evolution

Because the characteristics that separate reptiles and mammals evolved at different rates and in response to a variety of interrelated conditions, at any point in the period of transition from reptiles to mammals there were forms that combined various characteristics of both groups. Such a pattern of evolution is termed "mosaic" and is a common phenomenon in those transitions marking the origin of major new adaptive types. To simplify definitions and to allow the strict delimitation of the Mammalia, some authors have suggested basing the boundary on a single character, the articulation of the jaw between the dentary and squamosal bones and the attendant movement of accessory jaw bones to the middle ear as auditory ossicles. The use of a single osteological character allows the placement in a logical classification of numerous fossil species, other mammalian characters of which, such as the degree of endothermy and nursing of young and the condition of the internal organs, probably never will be evaluated. It must be recognized, however, that were the advanced therapsids alive, taxonomists would be hard-put to decide which to place in the Reptilia and which in the Mammalia.

CLASSIFICATION

Distinguishing taxonomic features. The higher classification of the Mammalia is based on consideration of a broad array of characters. Traditionally, evidence from comparative anatomy was of predominant importance, but more recently information from such disciplines as physiology, serology, and genetics has proved useful in considering relationships. Comparative study of living organisms is supplemented by the findings of paleontology. Study of the fossil record adds a historical dimension to knowledge of mammalian relationships. In some cases, the horses for example, the fossil record has been adequate to allow lineages to be traced in great detail.

Relative to that of other major vertebrate groups, the fossil record of mammals is good. Fossilization depends upon a great many factors, most important among which are the structure of the organism, its habitat, and conditions at the time of death. The most common remains of mammals are teeth and the associated bones of the jaw and skull. Enamel covering the typical mammalian tooth is composed of prismatic rods of crystalline apatite and is the hardest tissue in the mammalian body. It is highly resistant to chemical and physical weathering. Because of the abundance of teeth in deposits of fossil mammals, dental characteristics have been stressed in the interpretation of mammalian phylogeny and relationships. Dental

features are particularly well suited for this important role in classification because they reflect the broad radiation of mammalian feeding specializations from the primitive predaceous habit.

Annotated classification. The following classification has wide acceptance.

CLASS MAMMALIA

Vertebrate animals with the lower jaw formed by a single bone (dentary), articulating directly with the squamosal bone of the skull; dentition generally heterodont, thecodont, and diphyodont; body covered with hair derived from epidermis; young nourished by milk from mammary glands. For additional characteristics of the class, see *Form and function,* above. Groups preceded by a dagger mark (†) below are extinct.

Subclass Prototheria

Mammals in which principal cusps of molars form an anteroposterior row and much of the side of the braincase is formed by the periotic bone rather than by the alisphenoid. Primitive characters include uncoiled cochlea of the inner ear and absence of spine on the scapula (shoulder blade).

†*Infraclass Eotheria*

Fossil only; Upper Triassic to Lower Cretaceous of Eurasia and North America. Upper molars with internal and external cingula (shelves at base of crown), lower molars with internal cingulum only.

†*Order Triconodonta* (triconodonts). Fossil only; Northern Hemisphere; Upper Triassic to Lower Cretaceous. Probably carnivorous, the teeth well differentiated; molars typically with 3 prominent cusps; dentary bone lacking angular process.

†*Order Docodonta* (docodonts). Known only from fossil teeth and jaws from Upper Jurassic of Europe. Molars expanded, with transverse median furrow; angular process of mandible directed ventrally rather than posteriorly ("pseudangular").

Infraclass Ornithodelphia

Order Monotremata (monotremes). Pleistocene to Recent of Australia and New Guinea; possible record in Miocene of Australia family (Ectopodontidae). Reproduction by shelled eggs, incubated by female; no organized nipple, the milk flowing from mammary ducts onto the fur of the abdomen; cloaca present; adults toothless, rostrum covered with horny beak. Total length to 80 cm (31.5 in.). Examples: duck-billed platypus, and echidnas or spiny anteaters.

†*Infraclass Allotheria*

†*Order Multituberculata* (multituberculates). Fossil only; Upper Jurassic to lower Eocene of North America and Eurasia. The earliest herbivorous mammals; specializations of teeth and jaws superficially rodent-like; molars elongate, with two or three longitudinal rows of cusps; lower premolars tending to become serrated shearing blades. Size larger than typical Mesozoic mammals, to about that of a woodchuck.

Subclass Theria

†*Infraclass Tuberculata*

†*Order Symmetrodonta* (symmetrodonts). Fossil only; Upper Jurassic to Lower Cretaceous of Eurasia and North America. Molars with 3 prominent cusps arranged in symmetrical triangles; lower molars without posterior talonid basin ("heel").

†*Order Pantotheria* (pantotheres). Fossil only; Middle Jurassic to Lower Cretaceous of North America and Europe. Molars with 3 prominent anterior cusps (trigonid), asymmetrical; posterior basin (talonid) of lower molars developed, as in higher mammals.

Infraclass Metatheria

Order Marsupialia (marsupials). Upper Cretaceous to Recent of North and South America, Europe, Australia. Placenta formed from yolk sac (simple chorio-allantoic placenta in bandicoots), the young born in a relatively undeveloped state after brief gestation, developing further attached to mammae in marsupium (pouch) or along marsupial fold; female reproductive tract bifid, uterus and vagina double, penis bifurcate, posterior to scrotum. Angular process of mandible inflected; auditory bullae (when present) formed from alisphenoid bone and thus not homologous with bullae of eutherians. Epipubis ("marsupial bone") present; length of head and body varying from about 10 cm (4 in.; small opossums) to more than 160 cm (63 in.; kangaroos). Examples: opossums, kangaroos, wallabies, bandicoots, wombats, phalangers, koala.

Infraclass Eutheria (placental mammals)

Order Insectivora (insectivores). Upper Cretaceous to Recent, worldwide except Australia, polar regions, and many oceanic islands. A diverse group of mammals sharing certain primitive characteristics but highly specialized in several adaptive modes; dentition generally heterodont, rooted, variously modified; molars usually tuberculosectorial. Stance plantigrade to semiplantigrade; habit fossorial, terrestrial, arboreal, semiaquatic; insectivorous to omnivorous. Braincase low, rising but little above plane of rostrum; snout generally long, pointed, modified for olfaction and touch; zygomatic arch complete or incomplete. Length of head and body, 5 cm (2 in.; smallest shrews) to 45 cm (18 in.; largest hedgehogs). Shrews of the genera *Microsorex, Sorex,* and *Suncus* are the smallest of living mammals. Examples: shrews, moles, hedgehogs, tenrecs, solenodon.

Order Macroscelidea (elephant shrews). Lower Oligocene to Recent of Africa. Anteromost incisors widely spaced (incisor series forming a straight, anteroposterior row); posterior incisors caniniform; upper canine premolariform, bicuspid; posterior premolars molariform, last premolar largest tooth in dental series; molars quadrate, reduced in size or absent posteriorly; auditory bullae well developed. Stance plantigrade, hind feet elongate, modified for hopping gait; tail elongate. Length of head and body of living species, 12 to 20 cm (5–8 in.); tail 12 to 18 cm (5–7 in.; treated in section on *Insectivora*).

Order Dermoptera (colugos or "flying lemurs"). Upper Paleocene to lower Eocene of North America (Plagiomenidae), Recent of Southeast Asia (Cynocephalidae). In living species, upper incisors reduced, lower incisors procumbent, spatulate, pectinate (comb-like), molars low-crowned, used for shearing the vegetable diet. The most specialized and efficient of mammalian gliders; the patagium (gliding membrane) extends from the neck to phalanges of hand, to phalanges of foot, to tip of tail; total length of head and body, 34–45 cm (13–18 in.).

†*Order Tillodontia* (tillodonts). Fossil only; upper Paleocene to middle Eocene of Northern Hemisphere. Omnivorous or herbivorous; incisors rootless, enlarged, molars low-crowned; perhaps superficially bearlike in appearance; total length to approximately 120 cm (47 in.).

†*Order Taeniodonta* (taeniodonts). Fossil only; lower Paleocene to upper Eocene of North America, middle Eocene of Asia. Large, browsing herbivores; outer incisors rodent-like, inner incisors lost; canines well developed; cheek teeth simple, rootless pegs; length of skull to about 35 cm (14 in.).

Order Chiroptera (bats). Middle Eocene to Recent in tropical and temperate regions; volant, the wings consisting of a fold of skin (patagium) supported by elongated bones of the second through fifth digits; thumb clawed in suborder Microchiroptera, thumb and index finger clawed in Megachiroptera; hind limbs variously specialized but generally weak. Incisors generally reduced, canines prominent, cheek teeth rather primitive, with prominent ectoloph. Length of head and body from about 3 to 40 cm (1.2–16 in.); wing span to about 120 cm (47 in.).

Order Primates (primates). Late Cretaceous to Recent of Eurasia, Africa, North and South America. A diverse group of generalized, generally omnivorous mammals, modified for arboreality (some secondarily terrestrial). Dentition little specialized; a pair of incisors lost above and below, canines prominent, sexually dimorphic in some species; premolars usually bicuspid, sometimes caniniform; molars typically low-crowned, tuberculosectorial to bunodont. Limbs and girdles adapted for flexibility with little loss or fusion of elements; clavicle present; digits I (hallux and pollex) divergent, tending to become opposable; stance plantigrade. Olfaction de-emphasized in favour of vision, and snout reduced; eyes rotated forward allowing trend toward stereoscopic vision; in more advanced groups cerebrum expanded to cover cerebellum, braincase expanded. Length of head and body (including extended hind limbs), 15 cm (6 in.; pygmy marmosets, tarsiers) to more than 200 cm (79 in.; *Homo*). Examples: lemurs, lorises, Old World and New World monkeys, marmosets, great apes, man.

Order Edentata (edentates). Lower Eocene to Recent of North and South America. Anterior teeth lost and cheek teeth reduced to simple rootless pegs without enamel, or (in anteaters) lost altogether. Secondary articulations (xenarthrous processes) present between vertebrae in addition to zygapophyses (the usual articulating surfaces); cervical vertebrae 6 to 9, in some cases fused; acromion and coracoid processes of scapula enlarged, often joining to form bony ring; claws generally stout. Several groups walk on the sides of feet and structure may be modified accordingly. Brain small, skull low, zygomatic arch usually incomplete. Total length of fossil edentates to 300 cm (118 in.); living species to 185 cm (73 in.; giant anteaters). Examples: sloths, armadillos, anteaters.

Order Pholidota (pangolins). Oligocene to Recent of Eurasia and Africa; possible North America representatives (Paleanodontidae) upper Paleocene to lower Oligocene. Body covered with heavy, overlapping scales with hairs only at bases of scales; tail prehensile in some species. Myrmecophagous (anteating); skull elongate, braincase low; lower jaw reduced to thin shaft

of bone; feet 5-toed but only 3 toes of forefeet strongly clawed. Total length 75 to 170 cm (30–67 in.).

Order Tubulidentata (aardvark). Miocene to Recent of Africa, Pliocene of Eurasia; possible representatives in Eocene and Oligocene of Europe. Teeth reduced to 4 or 5 simple pegs in each jaw of adults, without enamel, the dentine intersected by numerous tubules (hence the ordinal name). Four toes on forefeet, 5 behind, each bearing a strong, stout, hooflike "nail." Skull elongate and tubular; tongue vermiform; mandible not so reduced as in other "anteaters." Legs powerful, skeleton generally primitive but tibia and fibula fused. One species; total length to 200 cm (79 in.).

Order Lagomorpha (lagomorphs). Upper Paleocene to Recent of Eurasia, North America, Africa, northern South America. Dentition superficially rodent-like, but with 2 pairs of upper incisors (a small pair situated behind the larger functional pair); canine and anterior premolars lost, forming prominent diastema (gap); cheek teeth high-crowned, rootless, 6 above and 5 below (more than in any rodent), and not occluding directly, so that mastication is a lateral movement; maxillary bone fenestrate (*i.e.,* with holes). Fibula articulating with calcaneum; infraorbital canal small. Scrotum anterior to penis. Total length, 15 cm (6 in.; pikas) to 70 cm (28 in.; hares). Examples: hares, rabbits, pikas.

Order Rodentia (rodents). Lower Eocene to Recent, worldwide except Antarctica and some oceanic islands. Single pair of rootless incisors above and below; canines and anterior premolars lost forming prominent diastema; never more than 2 premolars above and 1 below; cheek teeth frequently rootless, hypsodont; cusp patterns variable, cheek teeth ranging from pegs to massive teeth with highly complex occlusal patterns. Powerful jaw musculature developed in connection with grinding dentition; angular process of mandible well developed for muscle attachment. Postorbital bar absent; infraorbital foramen variously modified, in advanced groups passing portions of masseter muscle; postcranial skeleton relatively unspecialized. Forelimbs flexible, digits clawed, little reduction in number on forefeet, some specialization of hindfeet and reduction of digits associated with saltation; stance plantigrade or semiplantigrade. Total length of living species, 10 cm (4 in.; pygmy mice, some pocket mice) to 125 cm (49 in.; capybara). Examples: mice and rats, squirrels, porcupines, beaver.

†*Order Creodonta* (creodonts). Fossil only; Middle Cretaceous to upper Oligocene of North America, Eurasia, Africa. Primitive carnivorous mammals; canines well developed; cheek teeth specialized for shearing, but carnassial pair farther back in tooth row than in Carnivora. Auditory bullae unossified. Distal phalanges channelled for insertion of claws; no fusion of scaphoid, lunar, and centrale bones of wrist. More than 50 genera, in about 5 families. Size to that of bear or larger.

Order Carnivora (carnivores). Middle Paleocene to Recent on all continents (probably accompanied early man to Australia). Teeth variously modified, sectorial to bunodont, but the carnassial (shearing) pair of teeth always consisting of the last upper premolar over the first lower molar; dentition simplified in advanced pinnipeds (seals), nearly homodont, with no carnassials. Auditory bullae ossified in advanced forms. Terminal phalanges not channelled for insertion of claws; scaphoid, lunar, centrale bones of wrist fused. Carnivorous, omnivorous, herbivorous. Total head and body length 15 to 370 cm (6–146 in.). Examples: cats, dogs, bears, weasels, civets, hyenas, raccoons, seals, sea lions, walruses.

†*Order Archaeoceti* (archaeocetes, zeuglodonts). Fossil only; lower Eocene to middle Miocene of North Africa, Europe, North America. Primitive whale-like marine mammals, dentition not highly modified; anterior teeth simple, peglike; cheek teeth sectorial as in early carnivores; teeth not exceeding basic placental number (44); nostrils placed well back on elongate snout; skull not "telescoped" as in modern whales; hindlimbs vestigial by upper Eocene; body elongate, some serpentine; total length to more than 2,000 cm (65 ft).

Order Odontoceti (toothed whales). Upper Eocene to Recent in all oceans and occasionally in freshwater. Dentition generally simple, undifferentiated (homodont), exceeding the primitive placental number (to 300 teeth in some porpoises) or secondarily reduced (to a single tooth in *Monodon,* the narwhal); nostrils located on top of head, above and between eyes, forming single "blow-hole"; skull asymmetrical, markedly "telescoped," with reduction of frontals and parietals and concomitant extensive posterior expansion of maxillaries and premaxillaries. Relative to Archaeoceti, body shorter, stouter, total length 150 cm (59 in.) (small porpoises) to about 1,800 cm (59 ft) (male sperm whales). Examples: dolphins, porpoises, narwhal, sperm whales.

Order Mysticeti (baleen or "whalebone" whales). Upper Eocene to Recent in all oceans. Teeth absent; "whalebone" (baleen) present, consisting of ridged "curtains" of keratin extending from roof of mouth, used to strain krill (zooplankton); mouth cavernous, esophagus constricted to maximum diameter of about 20 cm; mandibular symphysis lacking; skull symmetrical, markedly "telescoped," with anterior expansion of occipitals; nostrils moved posteriorly, "blow-hole" double, anterior to orbits (eyes). Total length from 600 cm (236 in.) (pygmy right whale) to more than 3,000 cm (98 ft) (blue or sulfur-bottom whale); the blue whale, weighing to 136,000 kilograms (150 tons), is the largest of all known vertebrates, living or extinct. Examples: rorquals (including blue whale), gray whales, bow-heads, right whales.

†*Order Condylarthra* (condylarths). Fossil only; Upper Cretaceous to lower Oligocene of Eurasia, North and South America. Probably the basal stock from which most later ungulate groups evolved; directly descended from primitive placentals, and characteristics diverse. Tendency toward rounded molar cusps of equal height (bunodonty), formation of diastema, reduction of canine teeth, assumption of digitigrade stance and "experimentation" with hooves (ungules). About 90 genera and 7 families: some species at least as large as modern bears.

†*Order Pyrotheria* (pyrotheres). Fossil only; Eocene and Oligocene of South America. Incisors enlarged, tusklike; diastema prominent; cheek teeth low-crowned and with transverse crests; nostrils on top of heavy skull, suggesting development of fleshy proboscis. About 4 genera known; as in many other ungulate groups, trend to large size (to that of elephants); length of skull to 60 cm (24 in.).

†*Order Xenungulata* (xenungulates). Known only from a few fragmentary fossils from upper Paleocene of South America. Generally poorly known, but molars with tendency toward converging transverse crests as in Dinocerata of Northern Hemisphere; postcranial remains scanty but build obviously massive. Size to that of a modern rhinoceros.

†*Order Pantodonta* (pantodonts). Fossil only; Paleocene to Oligocene of North America and Asia. A varied group of herbivores, prominent in the Paleocene. Cheek teeth low-crowned, with transverse crests variously arranged, canines small or greatly developed. About a dozen genera; total length approximately 100 to 450 cm (39–177 in.).

†*Order Dinocerata* (uintatheres). Fossil only; Paleocene and Eocene of North America and Eocene of Asia. Skull low, later representatives with bony projections on nasal, maxillary, and frontal bones. Upper incisors absent, lower incisors reduced, saber-like upper canines (evidently only in males); upper molars unique, with 2 transverse crests forming a "V." Limbs graviportal (massive); about 8 genera; size to that of large rhinoceros.

Order Proboscidea (elephants and allies). Upper Eocene to Recent in Eurasia, Africa, North and South America. Earliest known representatives characterized by low-crown molars with two transverse lophs, diastema resulting from loss of outer lower incisor, canine and first premolar, and formation of upper and lower tusks by modification of incisors. Nasal opening toward top of skull, indicating that a fleshy proboscis ("trunk") was present, but perhaps of modest size; early proboscideans relatively long-bodied and short-limbed, superficially tapir-like. Middle and later Cenozoic times saw wide radiation from basal stock; about 2 dozen fossil genera known. Two living genera (*Elephas,* the Indian elephant; *Loxodonta,* the African elephant) with high and truncate skull; massive tusks; short mandible; peculiar cheek tooth arrangement in which a single, high-crowned, lophodont tooth is exposed in each half of the jaw at any one time, and replacement occurs continually from behind; long proboscis; graviportal limbs and girdles; length of head and body to 640 cm (252 in.).

Order Sirenia ("sea cows," dugongs, and manatees). Middle Eocene to Recent of Africa, Asia, North and South America. In living species: hindlimbs lacking, pelvis vestigial; front limbs modified into flippers; tail flattened, expanded; body fusiform, neck indistinct, anterior teeth reduced, replaced by horny plates used in cropping estuarine vegetation; molars of living dugongs bunodont, premolars lost; cheek teeth of manatees bilophodont, increased to 20 or more in each jaw, with 6 in each jaw at a given time and worn teeth replaced from behind as in modern proboscideans. Total length to 450 cm (177 in.).

†*Order Desmostylia* (desmostylians). Known only as fossils from lower Miocene to lower Pliocene in shallow marine deposits of North Pacific Ocean. Skull massive, low, elongate; anterior teeth modified as tusks, 1 pair (canines) above, 2 pairs (canines and incisors) below; cheek teeth peculiar, cusps forming cylinders of enamel and replaced in a continual forward movement as in elephants and manatees; limbs and girdles massive; appearance perhaps superficially hippopotamus-like; total length to about 250 cm (98 in.).

†*Order Embrithopoda* (embrithopods). Known only from fos-

sils from lower Oligocene of North Africa. Limbs 5-toed, progression graviportal; superficially rhinoceros-like, large paired horns on nasal bones and smaller horns on frontals (*i.e.,* behind the larger pair). Incisors not enlarged as in other subungulates; cheek teeth high-crowned; total length to about 330 cm (130 in.).

Order Hyracoidea (hyracoids). Lower Miocene to Recent of Africa and southwestern Asia. Most generalized of subungulates, although postorbital bar present and permanent dentition reduced; last upper incisors and canine lost, forming diastema; remaining incisors rootless; molars high-crowned. Superficially rabbit- or rodent-like; herbivorous, living on rock slopes or semiarboreal. Total length of fossil species to about 200 cm (79 in.); living members 30 to 65 cm (12–26 in.). Living examples: hyrax species.

†*Order Notoungulata* (notoungulates). Known only as fossils; upper Paleocene to Pleistocene of South America and Paleocene and Eocene of North America and Asia. Varied assemblage, united by dental peculiarities; accessory cusps in central valley of upper molars; lower molars with crescentric ridges, the entoconid lying within the arc of the larger posterior crescent; usually no diastema or reduction in dentition; cheek teeth tending to become rootless, high-crowned, lophodont. Tendency for reduction of toes from 5 to 3; total length approximately 30 to 300 cm (12–118 in.).

†*Order Astrapotheria* (astrapotheres). Known only as fossils; upper Paleocene to upper Miocene of South America. Upper incisors lost, canine enlarged, rootless; lower incisors and canine enlarged; diastema prominent; premolars reduced; molars greatly enlarged. Forelimbs stout, feet digitigrade, hind-limbs relatively weakly developed, semiplantigrade. Skull truncate; position and character of nasal opening suggests presence of fleshy proboscis; absence of upper incisors suggests that heavy lower incisors may have opposed a modified upper lip to allow cropping of vegetation. Total length to about 280 cm (110 in.).

†*Order Litopterna* (litopterns). Known only as fossils; upper Paleocene to Pleistocene of South America. Hoofed herbivores with tendency toward reduced toes; from 5 to 3 or even 1. Dentition usually complete (or with some loss of incisors); no diastema; cheek teeth low-crowned, selenodont; tendency for molarization of premolars. Size to that of small horse.

Order Perissodactyla (perissodactyls). Lower Eocene to Recent; North America, Eurasia, Africa. Incisors variously modified, canines lost or specialized, diastema nearly always present, with loss of first premolars above and then below; molars primitively bunodont, becoming progressively more lophodont, hypsodont; molarization of premolars common at later stages. Hornlike projections on skull not uncommon. Trends toward lengthening of limbs, with limitation of motion to forward and backward. Feet mesaxonic (*i.e.,* with axis through toe III), reduction typically to 3 digits and later to a single digit; digits mostly hooved, but those of chalicotheres clawed. Range in body size from that of small dog (Eocene horses) to that of largest of known land mammals, *Baluchitherium,* an Oligocene rhinocerotid of Asia, nearly 600 cm (236 in.) in height at the shoulder. Examples: titanotheres, chalicotheres, tapirs, rhinoceroses, horses.

Order Artiodactyla (artiodactyls). Lower Eocene to Recent; worldwide except Australia. Dentition complete, bonodont (as in pigs) or variously reduced; incisors often reduced and upper incisors lost, vegetation cropped with lower incisors opposing gums of premaxilla; canine and first premolars frequently lost, forming prominent diastema; remaining premolars simple, not completely molariform; cheek teeth primitively low-crowned, bunodont, becoming hypsodont, selenodont (cusps expanding into crescents) in more advanced forms. Unguligrade, foot paraxonic (*i.e.,* the axis between digits III and IV); digit I lost first; digits II and V frequently rudimentary (dew claws), but more prominent in pigs and hippopotamuses; in advanced types metapodials III and IV fused to form cannon bone, adding extra functional joint to limb; astragalus distinctive, the articular surface a double pulley allowing efficient flexion and extension of hindlimb but precluding lateral movement. Parietal bones of skull reduced, frontal bones expanded, frequently bearing antlers or horns. Height at shoulder, 25 cm (10 in.) in small antelope to 370 cm (146 in.) giraffe. Examples: pigs, cattle, sheep, goats, antelope, pronghorn, giraffe.

Critical appraisal. Three subclasses of Mammalia generally have been recognized for many years: Prototheria, the monotremes; Allotheria, the extinct multituberculates; and Theria, including marsupials and placentals as well as the extinct symmetrodonts and pantotheres. However, a number of Mesozoic groups could not be placed satisfactorily in any of these subclasses as defined. Evidence on the formation of the braincase in nontherian mammals has suggested that the primitive taxa may in fact constitute a more coherent group than was once supposed. All nontherian mammals are thus placed in a single subclass, Prototheria; infra-ordinal status is proposed for monotremes, triconodonts and symmetrodonts, and multituberculates.

Monotremes and multituberculates. Monotremes and multituberculates represent highly modified offshoots of the earliest mammalian stock. The fossil record of undoubted monotremes is limited to the Pleistocene. Surely the group is an ancient one and, although living representatives are highly specialized, represents the grade of evolution of the reptilian-mammalian transition. Their primitive organization is underscored in the relatively unspecialized brain, retention of the oviparous habit and a cloaca, incomplete homeothermy, and reptile-like frontal and pterygoid bones.

The multituberculates also were derived from the earliest mammalian stock. They were adapted early to a herbivorous diet and developed gnawing incisors and elongate, grinding cheek teeth. The fossil record of the multituberculates, from Upper Jurassic to lower Eocene, is longer than that for any other mammalian order. The extinction of multituberculates in the Early Cenozoic may have been due to competition from placental herbivores, in particular the early rodents.

Marsupials. The split between marsupials and placentals has yet to be identified with any certainty. *Pappotherium* of the Lower Cretaceous of Texas was thought for a time to represent a common ancestor of the therian infra-classes, but more recently it has been discovered that recognizable didelphoid marsupials were contemporaneous with the pappotheriids.

The earliest records of marsupials are from North America but the great adaptive radiation of this group occurred in South America and Australia. Mostly free of competition from placentals on Cenozoic island continents, marsupials converged in many cases with the specializations of placentals elsewhere. Australia has been isolated from other major land masses throughout the Cenozoic and South America was an island from Paleocene to Pliocene. Invasions by placentals in later Cenozoic times reduced the once flourishing South America marsupial fauna to the remnant extant today. The South American and Australian radiations were independent, although both assemblages had a common source in generalized didelphoid (opossum-like) groups and evolved in parallel fashion in a number of instances. In general, higher categories in the classification of marsupials are not coordinate with those of the Eutheria. Some families of marsupials, for example, encompass a range of diversity comparable to that in eutherian orders.

Insectivores and colugos. Modern insectivores are divisible into several subgroups. The lipotyphlans include modern shrews, moles, and hedgehogs. The molars of hedgehogs are quadritubercular; those of shrews and moles are complicated by the development of a W-shaped external ridge or ectoloph. Lipotyphlans maintain the primitive insectivorous habits of their Mesozoic forebears. Zalambdodonts are represented by the peculiar selenodonts of the Greater Antilles, the tenrecs of Madagascar and West Africa, and the chrysochlorids—Cape golden moles—of southern Africa. The zalambdodonts may or may not represent a natural group. Their early history is obscure. All have V-shaped upper molars.

Tree shrews (tupaiids) of Southeast Asia are the sole survivors of a third insectivore group, sometimes called proteutherians, which includes a great variety of Late Cretaceous and Early Cenozoic mammals of uncertain relationships. Living tupaiids have frequently been classed as Primates, for they have remarkably advanced brains and versatile limbs. In such features as dentition, however, they remain quite primitive.

The macroscelids, or elephant shrews, are here treated as a separate order. They have frequently been grouped with tupaiids as a suborder of the Insectivora, or as a distinct Order Menotyphla. Colugos, the so-called flying lemurs, here maintained as a distinct order, Dermoptera, likewise have been considered by some authorities as insectivores.



Figure 5: Reconstruction of *Melanodon*, a shrew-sized insectivorous pantothere from the Late Jurassic of North America.
Drawing by R. Keane

Tillodonts and taeniodonts. The relationships of tillodonts and taeniodonts are obscure. Although apparently quite independent, both groups were evidently derived from primitive placental stocks, and both evolved rootless, rodent-like incisors. Neither group was ancestral to later, more successful mammalian types.

Bats. Bats are an ancient group; except for extensive skeletal and sensory modifications associated with flight they share many primitive characteristics with their insectivore ancestors. Bats first appear in the fossil record in the Eocene as highly adapted fliers. Were well-preserved postcranial (body and limb) remains of the earliest Cenozoic insectivores better known, some might well be found to represent intermediate stages in the evolution of flight; to date no such intermediate stages have been found. Many have thought that the Dermoptera might represent a grade through which the lineage of the bats must have passed in evolving truly volant habits; surely dermopterans and bats are not closely related, save in their retention of primitive characters of early placentals. In recent years research has brought to light important differences between the two suborders of bats, the Megachiroptera and the Microchiroptera, and some workers have suggested that the two groups be treated as separate orders, maintaining that the similarities between them are due to parallelism and convergence in adapting to the stringent demands of flight.

Primates. The Primates form a relatively coherent group, although the distinction between them and their insectivore ancestors is a difficult one and some groups, such as the tree shrews, can be placed in either order with nearly equal justification. Disagreement on classification within the order centres on the validity of the suborder Prosimii as a natural unit that includes such divergent kinds as tarsiers, lemurs, lorises, and galagos, and on the relationship between the Old World and New World monkeys. The two monkey groups may have evolved independently from prosimian ancestors; considerable convergence and parallelism no doubt are involved in the similarities between the two as both have radiated widely in response to available arboreal habitats.

Edentates and pangolins. The edentates have undergone their greatest radiation in South America. They are a highly aberrant group, of unknown ancestry, and apparently were isolated early in South America, as were the archaic South American ungulates. Edentates have commonly been subdivided into Loricata (armadillos and extinct glyptodonts), Pilosa (tree sloths and extinct ground sloths) and Vermilingua (anteaters). Each of these subgroups has had a long and independent history, but they share the unique character of xenarthrous vertebrae. The palaeanodonts of the Eocene and Oligocene of North America have been considered as nonxenarthrous edentates (a view presented in the section on *Edentata*); they are here allocated tentatively to the order Pholidota. Little is known of the relationships of pholidotes (pangolins), once considered relatives of the Edentata; the groups are now generally conceded to be convergent.

Aardvarks. Aardvarks (Tubulidentata) also were thought at one time to be related to true edentates, but are now considered an independent lineage. Although they are known from Miocene deposits, no direct evidence exists of relationships with other groups. In some respects the skeleton resembles that of certain early condylarths, and it seems probable that tubulidentates descended from an early ungulate stock.

Lagomorphs. The Lagomorpha long were included (as a suborder Duplicidentata) in the Rodentia on the basis of dental similarities. Although ordinal rank for the Lagomorpha has been firmly established, some recent classifications still include rodents and lagomorphs together in a cohort Glires, again based on dental similarities, which almost certainly represent convergence. Various other mammalian lineages of diverse affinities (multituberculates, taeniodonts, tillodonts, and certain primates, for example) have "experimented" with gnawing incisors. The earliest known fossil lagomorphs clearly represent the group and yield no clues as to origin. On the basis of serology and some dental characters, a relationship to ungulates through the condylarths has been suggested. More than likely, however, those characters held in common are primitive and lagomorphs arose directly from a primitive insectivore stock.

Rodents. In terms of numbers of individuals and numbers and diversity of species, the rodents are the most successful of living mammals and have evidently been so throughout much of the middle and later Cenozoic. The earliest known rodents (paramyids), of Paleocene age, were already highly specialized mammals. Fossil evidence indicating relationships with more primitive groups is lacking; rodents may have been derived directly from a basal eutherian stock.

The higher classification of rodents has been in a state of flux in recent years. Three suborders (Sciuromorpha, Myomorpha, and Hystricomorpha) were recognized, based on the relationship of the zygomasseteric musculature to the infraorbital foramen. The sciuromorphs, or Protrogomorpha, included the earliest known rodents and also modern squirrels, beavers, and sewellels or "mountain beavers." In these taxa the masseter muscle originates from the anterior portion of the zygomatic arch or from the side of the skull in front of the orbit. The infraorbital canal is small, carrying nerves and blood vessels as in primitive mammals. In the Myomorpha, the rats and mice and their allies, the deep portion of the masseter muscle passes through the enlarged infraorbital canal. The so-called hystricomorphs, including Old World porcupines and a great variety of South American rodents, have a greatly enlarged and round infraorbital foramen, rather than a slitlike one like that of myomorphs.

Recent authors have doubted the validity of this arrangement and the utility of the infraorbital canal and zygomasseteric structure as a diagnostic feature. Many would admit the general validity of the Sciuromorpha and Myomorpha although some groups traditionally included would be reallocated. The "Hystricomorpha," on the other hand, is almost unanimously considered to be polyphyletic, and is now commonly subdivided into Caviomorpha for the complex South American "hystricomorph" radiation, and Hystricomorpha for Old World porcupines and their allies. The tendency has generally been to de-emphasize the use of subordinal classification and instead to bring together related families into a number of superfamilies. Perfection of the higher classification of rodents is difficult because fossil evidence is meager for some groups, particularly those of small or modest size, and because the successful radiation of the Rodentia has obviously led to perplexing convergence in numerous lines.

Creodonts. Creodonts were derived from the basic insectivore radiation as early carnivorous mammals, largely in response to the ecological opportunity presented by the evolution of herbivorous types.

Carnivores. True carnivores evidently are not descendants of the creodonts, but direct derivatives of Paleocene insectivores. The diversification of fissiped (terrestrial) carnivores from the basal stock (family Miacidae) occurred in the upper Eocene and lower Miocene, early dividing into arctoid or canoid (canids, ursids, procyonids, mustelids) and aeluroid or feloid (felids, hyaenids, viverrids) lines. The Pinnipedia (seals, sea lions, walruses) are here considered a suborder of the Carnivora. Appearing in the fossil record in the Miocene, the modern pinniped families already were distinct. The previous history of the group

is obscure, but pinnipeds apparently were derived from arctoid fissipeds, although some have thought them to have been derived from creodonts. Some authorities have treated pinnipeds as a distinct order, in which case they represent the most recently evolved mammalian order.

Whales. The whales and their allies have traditionally been treated as a single order, Cetacea, a reasonable arrangement on the basis of certain criteria. It is obvious, however, that cetaceans diverged early into archaeocetes, odontocetes, and mysticetes, if, indeed, the groups had a common ancestry. The earliest known fossils are perfectly good whales and clearly are assignable to one of the three groups; thus there is no concrete evidence linking the three nor is there evidence as to their ancestry. The morphological characters that whales share can be interpreted as due to parallelism resulting from stringent selection in similar habitats, or to the retention of primitive placental characteristics. Until evidence is available to establish the monophyly of cetaceans it seems preferable to treat them as representing three separate orders.

Ungulates. The term "ungulate" denotes a broad, loose association of orders including most large herbivorous mammals. The name of the group implies that its members are hoofed, but some of the ungulates (chalicotheres, for example) had clawed digits and others (such as sirenians) show considerable reduction of digits and limbs. The ungulates may be of common ancestry or may be a product of convergence from several sources among primitive placental stocks. Certain supposed early ungulates (condylarths) and early carnivorous mammals (creodonts) have many characters in common; their supposed descendants frequently have been united in a common cohort, Ferungulata.

The principal characteristics shared by ungulates are those of the dentition and limbs, related to modification of the primitive placental into an efficient herbivore.

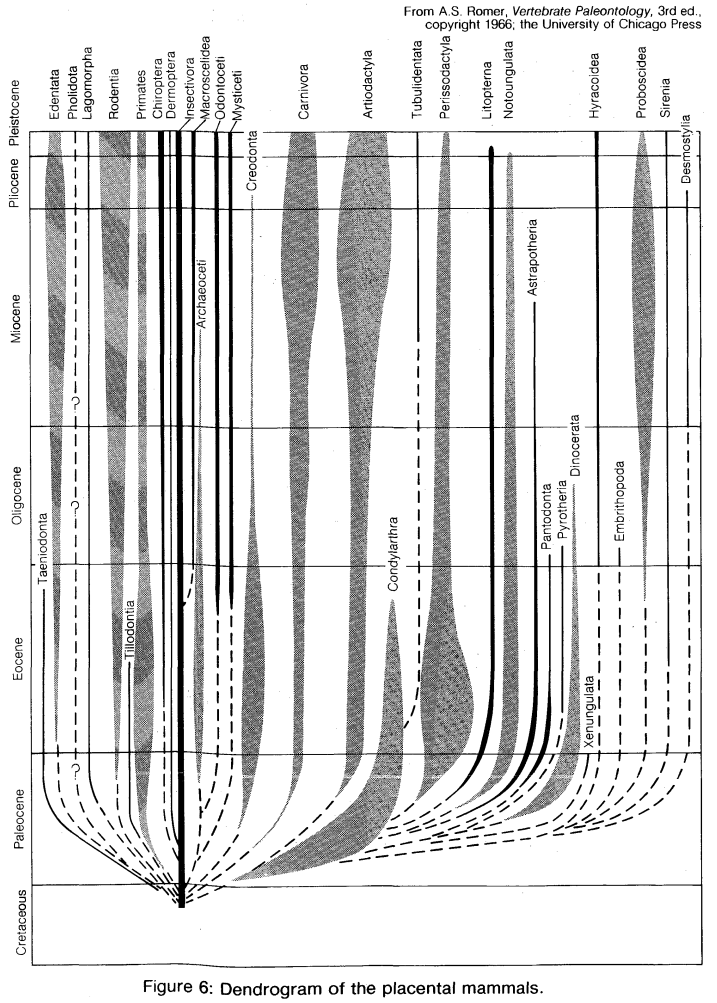


Figure 6: Dendrogram of the placental mammals.

There has been a tendency toward elongation of individual molar cusps to form crescentric ridges (selenodonty) or to form ridges between adjacent cusps (lophodonty). A concomitant trend has been an increase in the height of molars from low-crowned (brachyodont) to high-crowned (hypsodont), which has allowed the efficient utilization of abrasive forage such as grass. The extent of the grinding surface has been increased by increasing the length of the molars or by "molarization" of premolars.

The radiation of large herbivores was accompanied closely by a radiation of large carnivores. The adaptive response of ungulates to predation has been varied: modified dentition (tusks), skeletal elements (horns, antlers), and behaviour have been turned to defensive purposes. Extremely large size or swift, cursorial locomotion were at a premium. Defensive gregariousness is common in extant species.

The general trend in mammalian locomotion has been plantigrade to digitigrade to unguligrade, frequently with reduction of digits. One mode of digit reduction has been mesaxonic, the axis of the foot passing through digit III (as in horses); other groups have paraxonic limbs, with the axis lying between digits III and IV. Elongation of metapodials has occurred commonly, forming a third functional segment of the limb, allowing fast cursorial gaits.

Exceptions to these trends in limb modifications occur in lineages with early trends toward extremely large size. In such groups, columnar, graviportal limbs have developed with proximal segments (femur, humerus) emphasized and metapodials de-emphasized; all digits usually are retained.

The ungulate radiation has been complex. Various subdivisions have been proposed. The condylarths may or may not be basal to the entire radiation, but they evidently gave rise to the perissodactyls.

Perissodactyls. This is a compact order, consisting of five subdivisions: horses, chalicotheres, titanotheres, rhinoceroses, and tapiroids. The history of the perissodactyls is perhaps better known than that of any other mammalian group. *Hyracotherium* ("*Eohippus*"), usually considered the base of the horse lineage, was probably near the source of the entire order.

Artiodactyls. The Artiodactyla first appear in the early Eocene. Dental characters point to an origin for the group among hypsodontid condylarths but the earliest known artiodactyls already had the unique and specialized "double-pulley" astragalus. Knowledge of transitional stages between condylarthran and artiodactyl tarsi would be of considerable interest and importance.

Archaic ungulates. A number of kinds of archaic ungulates, which generally evolved in the Western Hemisphere, have been lumped in the past in a single order, Amblypoda. Each of these groups—pantodonts, uintatheres (Dinocerata), pyrotheres, xenungulates—is here accorded ordinal rank. The relationships of these groups—if any—are obscure and their origin is unknown.

Litopterns. The Litopterna of South America were early derivatives of the condylarths, and some workers would include both in a single order. One family (Prototheriidae) closely paralleled the contemporaneous radiation of perissodactyls on the northern continents, but the macrauchenids converged toward the camels.

Notoungulates. The order Notoungulata was an early and varied assemblage, apparently restricted to South America in the Eocene, perhaps by competition with early members of more advanced ungulate groups. The greatest development of the order was in the Oligocene, after which there was a gradual decline, hastened by the entry into South America of placental carnivores and modern artiodactyls. The origin and affinities of the notoungulates are unknown, but the order may well have descended from the condylarths.

Astrapotheres. Astrapotheres were once considered to have been notoungulates. It is now generally conceded that their peculiarities recommend recognition as a separate order. This group may have had its origin among primitive notoungulates, or it may have arisen independently.

Hyracoids. The products of the early Cenozoic radiation of ungulates in Africa are commonly grouped as "subungulates." The Hyracoidea is the most generalized

of these orders but the true relationships of the hyracoids to other ungulates are unknown. In common with other subungulate orders, hyracoids were more diverse in the Middle Cenozoic than they are today.

Embrithopods. The extinct Order Embrithopoda contains a single known genus, *Arsinoitherium.* Relationship with hyracoids has been suggested but concrete evidence is lacking. Aside from logical association with other subungulate groups, little can be said of the position or affinities of the Embrithopoda.

Elephants and allies. The best known of the subungulates are the Proboscidea. The later record of the order is relatively good and indicates a remarkably successful and widespread radiation in the latter part of the Cenozoic—a radiation characterized by the recurrence of several trends in separate lines. Among these trends were an emphasis on the development of the upper tusk with loss of the lower tusk in most groups, increase in the size of the skull and shortening of the neck, increase in the length of the proboscis, increase in overall size, with perfection of graviportal limbs and girdles, lengthening of the lower jaw (secondarily shortened in some groups), and reduction and specialization of cheek teeth.

Sirenians. The origin of the Sirenia is obscure. The earliest known sirenians were completely aquatic although the pelvic girdle was not so much reduced as in later forms. Dental and cranial peculiarities suggest that proboscideans and sirenians may have arisen from a common swamp-dwelling or amphibious ancestor. Concrete evidence is generally lacking regarding this and other relationships among subungulates due to a paucity of fossiliferous continental sediments of early Cenozoic age in Africa.

Desmostylians. Because of dental peculiarities and their association with shallow-water marine sediments, fossil remains of the order Desmostylia were long thought to represent aberrant sirenians; recent discoveries of massive limb bones of these mammals indicate, however, that they had quite different adaptations. The nature of the cheek teeth and their mode of replacement suggests relationships with the largely African subungulates, although desmostylians are known only from fossil deposits along the shores of the North Pacific. (J.K.J./D.M.A.)

MAJOR MAMMAL ORDERS

The following sections consist of a review of living mammals, beginning with the most primitive—monotremes and marsupials—and continuing with the major orders of placental mammals. Groups are identified at the order level of biological classification and arranged in accordance with the *Annotated classification* above.

Extinct groups and species (designated in the annotated classification by a dagger [†]) are treated throughout this article under the subheadings *Evolution and paleontology;* for additional information about extinct mammals, see GEOCHRONOLOGY: *Fossil record,* which also treats the subject of mammalian evolution. (Ed.)

Monotremata (platypus, echidnas [spiny anteaters])

The Monotremata (monotremes) are a distinctive order of primitive mammals, the only surviving members of the subclass Prototheria.

Monotremes lay eggs and are reptilelike in many other ways but possess such essentially mammalian characters as mammary glands, hair, a large brain, and a complete diaphragm.

GENERAL FEATURES

Only two monotreme types, the platypus (*Ornithorhynchus anatinus,* family Ornithorhynchidae) and the echidnas, or spiny anteaters (*Tachyglossus aculeatus* and *Zaglossus bruijni,* family Tachyglossidae), are known. The earliest monotreme fossils known come from the Australian Pleistocene (2,000,000 or more years ago), and they are essentially the same as the living forms, which range in body length from about 30 to 80 centimetres (12 to 32 inches) and weigh from about one to 10 kilograms (two to 22 pounds).

The platypus, found in eastern Australia from north Queensland to Tasmania, is one of the most remarkable of all mammals. It has a ducklike "bill," webbed feet, and a flattened beaver-like tail. When the first stuffed specimens reached England around the end of the 18th century, they were thought to be fakes, made by sticking together bits of different animals.

The echidnas also rank high among the world's most interesting mammals and are readily kept in captivity for physiological studies. Superficially, they are quite unlike the platypus. They are recognized by their sharp-pointed spines and tubelike noses. *Tachyglossus* is found throughout Australia, including Tasmania, and part of New Guinea. *Zaglossus* has an extensive distribution in New Guinea.

Neither the platypus nor the echidnas are readily seen by the casual visitor, but they may be fairly common in some areas. Platypuses are not kept in zoos outside Australia.

The monotremes do not appear to be carriers of diseases harmful to man. Both monotreme types are now wholly protected by law in Australia.

NATURAL HISTORY

Ecology. The platypus is amphibious, with a habitat ranging from alpine waters at an altitude of 1,500 metres (about 4,900 feet) in southeastern Australia to the sluggish subtropical Queensland coastal streams. It digs winding burrows with side branches in the banks of the streams and lakes that it inhabits. These burrows are usually five to 10 metres (15 to 30 feet) long but may be as long as 30 metres (100 feet). The entrance is normally above the water level, however during floods it may be submerged.

Enemies of the platypus

The platypus has few natural enemies. It was formerly hunted extensively for its valuable velvet-like fur but for many years has been completely protected by law, and as a result it has increased in numbers. It appears to be relatively free of competition from other species in the water, where it feeds mainly on insect larvae, small crayfish, tadpoles, and other aquatic animals. Some authorities have suggested that the introduced rabbit (*Oryctolagus cuniculus*), which burrows in the banks of streams where the platypus once dwelt in quiet seclusion, is responsible

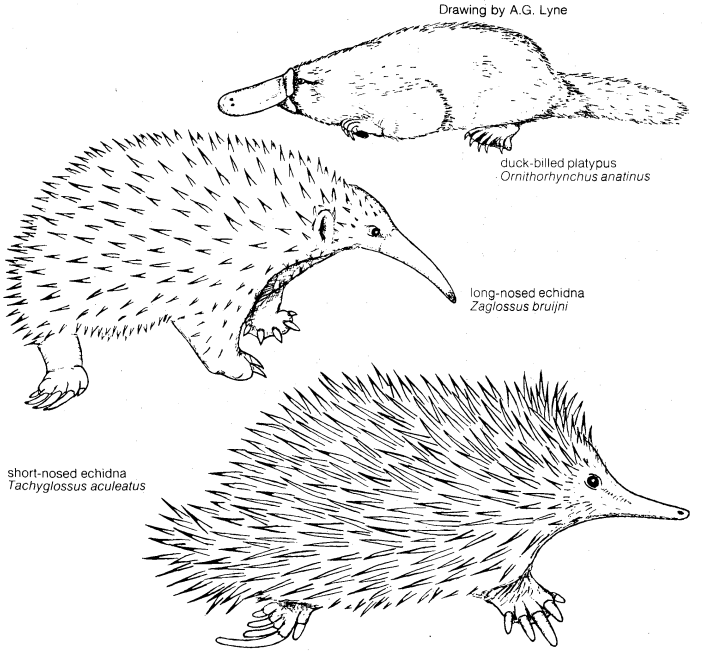


Figure 7: The three extant monotreme species.

for the disappearance of the platypus from many settled regions.

The echidnas are terrestrial animals, living in forests, scrubland, and deserts. Although usually encountered only by accident, *Tachyglossus* is one of the most widely distributed mammals in Australia. In New Guinea the habitats of *Zaglossus* include the humid forests that are almost continually blanketed by cloud cover. The animal has a known altitudinal range in these forests from about 1,100 to about 2,900 metres (3,600 to 9,500 feet).

Tachyglossus lives largely on ants and termites. The one element common in all the habitats occupied by this echidna is the presence of ants; termites are present in only some of the localities occupied.

Unfortunately, little is known of many aspects of the ecology of the monotremes, such as population density and social behaviour.

Life cycle and reproduction. The platypus breeds once annually, from July to August in the northern part of its range and to October in the southern part. Mating is believed to occur in the water, and it has been suggested that the spurs on the hind legs of the male may be used to hold the female. No spur grip was noted in an observation of copulation recorded by Australian zoologist D.H. Fleay in the course of an extensive study. An unusual courtship precedes mating, in which, among other manoeuvres, the male grabs the female's tail and the pair swim in circles.

In the breeding season, the female platypus digs a special burrow that twists and turns and may be very long. She then builds a nest of grass, leaves, and other plant material in a chamber at the end of the breeding burrow. This special burrow is blocked with earth in one or more places by the female during the period when she is laying and incubating the eggs. Only the left ovary is functional, and the eggs are always found in the left uterus. One to three, usually two, leathery-shelled eggs are laid. When there are two or three eggs they adhere together, but the sticky fluid that covers the eggs when they are laid soon dries and loses its stickiness. The eggs vary from 16 to 18 millimetres (about 0.60 to 0.70 inches) in length and from 14 to 15 millimetres in breadth.

The platypus has been bred in captivity on one occasion only, by Fleay in 1943. His observations suggest that the period from mating to egg laying is probably 12 to 14 days and the incubation period an additional 10 to 12 days. The young are not more than about 18 millimetres long at hatching and do not leave the breeding burrow until they are about 17 weeks old. At this age they are fully haired and about 34 centimetres (13 inches) long. Because the female platypus has no teats, the milk is apparently sucked from the skin as it exudes from the ducts of the mammary glands. The life-span of the platypus is said to be 10 to 15 or more years; at least one captive animal has lived more than 10 years.

More research has been done on the echidna *Tachyglossus* than on the platypus, mainly because it is less difficult to maintain in captivity. In southeastern Australia, echidnas breed once a year from late June to late September. The period between copulation and egg laying is at least 16 days and may be as long as 28 days. Both ovaries are functional.

Egg laying in the echidna

In the breeding season the female echidna develops an abdominal pouch. Egg laying has not been observed, but there is circumstantial evidence that the egg is laid directly into the pouch by the female's curving her body during this process. One egg (rarely two or three) is laid at a time; it is creamy yellow and the shell has a rubbery texture. The eggs of echidnas are slightly smaller than those of the platypus, being from 14 to 17 millimetres (0.55–0.67 inch) long and from 13 to 15 millimetres wide. Within the pouch the egg is not stuck to the hairs but is quite loose and from time to time is located in different positions, apparently retained by the apposition of the lips of the pouch. The incubation period of the egg in the pouch is 10 to 10.5 days, and the newly hatched echidna is about 15 millimetres (0.6 inch) long and weighs about 0.38 gram (about 0.01 ounce). After hatching, the young lives in the pouch and imbibes milk by sucking two special areas of the skin—the mammary areolae—where the ducts



(Left) Echidna (*Tachyglossus aculeatus*) held head downward to show egg in pouch. (Right) Echidna hatching.
M.E. Griffiths

from the mammary glands open to the exterior on either side of the midline of the pouch. As in the platypus, no teats are present.

Authorities disagree on whether the young echidna takes milk by sucking or by licking. After detailed observations, an Australian zoologist, Mervyn Griffiths, stated that advanced pouch young take milk by sucking and not by licking it off the hairs in the mammary region, but he considered it possible that newly hatched and early stages of pouch young get their milk by licking it up as it is ejected onto the surface of the skin.

Information on how long the young stay in the pouch and what happens to them when they are cast out is meagre. One of the echidnas studied by Griffiths increased in body weight from 240 grams (about 0.5 pound) to 850 grams (about 1.9 pounds) in 43 days, a mean daily increase in weight of about 14 grams per day. The young was cast out of the pouch when it reached a body weight of 400 grams (about 0.9 pound). Thereafter, the mother's pouch regressed quickly, but she continued to suckle the young for some time. At the time when the young was dropped from the pouch, its eyes were open and the sharp spines projected from the skin. In wild echidnas it has been reported that the mother deposits the young in a burrow when its spines emerge and it is too large to be carried in the pouch.

Tachyglossus has been bred in captivity on only three occasions, once in Berlin and twice in Basel in Switzerland. In captivity *Tachyglossus* has lived up to 50 years, and *Zaglossus* up to about 30 years.

Behaviour and locomotion. Both the platypus and echidnas exhibit extreme specializations for feeding and locomotion.

The platypus emerges from its burrow for feeding usually in the early morning or late afternoon, but is sometimes in the water at other times of the day. When it is burrowing or walking on land the web of the front feet, which extends beyond the claws, is folded under the palms. The muzzle and the claws of the front feet are used in burrowing. The platypus swims gracefully and expertly, using the front feet. The hind feet, in conjunction with the flattened tail, are used mainly for stablization. The tail also assists the platypus in diving. The animal often swims along the surface with only the upper part of the muzzle and a small part of the head and body above the water. When submerged, the eyes and ears are closed by a fold of skin. The sensitive muzzle is an important organ for guiding the platypus while it swims blind, often close to the bottom. It may stir up the mud or move stones to locate its food. It surfaces often to breathe and to chew the prey, which it stores in cheek pouches while it is submerged. It can rest near the bottom for a few minutes by wedging itself under stones or logs. In captivity, the platypus has an enormous appetite, devouring each day the equivalent of half or more of its own body weight.

Swimming by the platypus

Echidnas are generally solitary. They appear to be more uncommon than they are because of the inaccessible nature of the places they choose to hide in and because of their remarkable ability to dig vertically into the ground and cover themselves with dirt when disturbed.

Observations on *Tachyglossus* in various parts of Australia indicate that its activity is closely related to the ambient temperature. In north Queensland, they are active primarily at night; but in the more temperate parts of eastern Australia, they are active both day and night. In desert regions they retire into crevices in rock outcrops during the day, where the humidity is higher and they are protected from intense sunlight.

Although echidnas are well adapted for digging, they do not excavate burrows, and they dig primarily to obtain food and escape enemies. They also retreat into hollow logs or under rocks or roots and wedge themselves into crevices and even into slight depressions in such a way that it is practically impossible to dislodge them. They can withdraw their appendages and erect their spines, as does a hedgehog. Echidnas walk with the legs fully extended so that the undersurface of the body is relatively high off the ground and with the hind toes directed outward and backward. They can run and climb well. *Tachyglossus* is extremely powerful and can tear apart rotten logs in search of termites or dig down to ant nests in rocky terrain.

FORM AND FUNCTION

General characteristics. One of the most striking features of the monotremes is that they lay eggs from which the young are subsequently hatched. These eggs are like those of reptiles in that they have large yolks and rubbery shells. Other interesting reptilian features are seen in the structure of the urogenital organs and in the skeleton. In both sexes of the monotremes, as in all reptiles, the intestine, bladder, and reproductive organs all open into a common chamber, the cloaca, with only one external opening. The penis of monotremes is used solely for delivering sperm and not also for urination as in most other mammals. In the skeleton, the shoulder gridle not only retains well-developed coracoid bones, which are reduced to vestiges in other mammals, but also an interclavicle. The pelvic girdle carries epipubic bones, which, as in some reptiles and marsupials, act as supporting structures for the ventral (abdominal) body wall.

Similarities to reptiles

The skull of monotremes has a smooth, rounded cranial portion, terminating in a long rostrum, or snout. Both monotreme types are toothless as adults, and the echidnas are toothless throughout life. The platypus has teeth when young, but they are shed before the animal becomes adult and are replaced by horny pads. The testes are abdominal and there are no teats and no vibrissae (whiskers). The limbs are modified for digging or swimming. The posture of the platypus resembles that of reptiles, especially lizards. In the echidnas, however, the limbs support the body well off the ground, even when the animals are stationary or walking slowly.

An interesting feature of the digestive system of monotremes is that the whole inner surface of the stomach is lined with a cornified (horny) epithelium. There are no glands of any kind. Particularly interesting is the lack of those that produce hydrochloric acid and peptic enzymes, which together initiate the digestion of protein in other mammals. Some other insectivorous mammals also have cornified epithelium covering most of the stomach lining, but there is a small glandular area. It has been suggested that the breakdown of food in the stomach of *Tachyglossus* is assisted by the grinding action of ingested dirt.

Platypus. The platypus has a flattened body covered with dense, short, fine hairs and coarse guard hairs, usually dark brown on the back and much paler on the belly. There is practically no neck. The tail is broad and flat like that of a beaver. It is covered with coarse hair, which is usually worn off on the under surface. The adult male is 50 to 60 centimetres (20 to 24 inches) in total length and the adult female is 40 to 50 centimetres (16 to 20 inches). An average-sized male weighs nearly two kilograms (4.4 pounds). One of the most conspicuous features of the platypus is its elongated and flattened muzzle, which bears a superficial resemblance to a duck's bill. This muzzle is covered with a soft, rubbery, naked, and sensitive skin, which is a blue-gray colour on its upper surface and pale pinkish or mottled below. Dorsal (toward the back) and ventral flaps, or shields, of naked skin extend backward from the muzzle over the adjacent hair. The nostrils are located on the upper surface of the muzzle near its anterior end. The eyes are small and there are no external ears. Both the eyes and ears open into a facial furrow that is closed when the platypus is below the surface of the water. There are cheek pouches in which food is stored until it can be chewed. The limbs are short and the feet webbed. The webbing on the hind feet reaches to the base of the claws, and on the front feet beyond the claws, making extra large paddles for swimming.

Venom apparatus of the platypus

The male platypus has a sharp, movable, horny poison spur on the inner side of each hind limb near the heel. Each spur is about 15 millimetres (0.6 inch) long and connected by a duct to a venom-secreting gland, called the crural, or poison, gland, situated on the dorsal aspect of the upper part of the hind limb. The function of the venom apparatus is unknown. An Australian zoologist, J.H. Calaby, stated that "If, as is generally believed, the crural gland is actively secreting only during and near the breeding season it seems very likely that the chief use of the venom apparatus would be combat between males for territory or females." The poison is not fatal to man but causes intense pain.

Echidnas. Superficially, the echidnas are quite unlike the platypus. They have a rounded body covered on the back and sides with stout, sharp-pointed spines and some hairs. The spines of *Tachyglossus* may measure up to six centimetres (2.5 inches) in length, those of *Zaglossus* up to about three centimetres. The undersurface is covered with coarse hair and is usually without spines. The colour of the spines and hair ranges from light brown to almost black, depending on species and individual variation. Adult males are larger than adult females. There is a long, naked, and sensitive muzzle, a very short tail, and no neck. In some echidnas a definite external ear is present, but in others it is inconspicuous. The nostrils open near the tip of the muzzle. The mouth is small, only sufficient for the protrusion of the long, sticky tongue. The limbs are short and powerful and have strong claws for digging. In *Tachyglossus* one of the claws of the hind foot is particularly elongated and is used for scratching the skin between the spines.

The echidnas resemble the platypus in that there is a venom apparatus on each hind leg, but the spur and crural gland are smaller. In *Tachyglossus,* the spur (about five to 10 millimetres [0.2–0.4 inch] long), which is situated on the inside of the ankle, is present in all males and some females. No observations on the use of the venom apparatus by the echidnas have been published.

Temperature regulation and hibernation. Monotremes are better able to control their body temperature than are reptiles. Except during hibernation, or torpidity of some species, the eutherian, or placental, mammals maintain a high and virtually constant body temperature usually between 36° and 40° C (97° and 104° F). The body temperature of most reptiles varies with that of the surroundings. The body temperature of the monotremes, however, does not quite reach that of the eutherian mammals, being about 31° to 32° C (88° to 90° F). The monotremes are usually described as physiologically primitive in regard to temperature regulation. It has been shown, however, that *Tachyglossus* is an excellent temperature regulator in some situations but less capable in others. It is probably no poorer in this respect than many eutherian mammals (*e.g.,* anteaters, armadillos, and sloths). In cold surroundings the echidna maintains a well-regulated body temperature. Studies of temperature regulation and torpor in *Tachyglossus* have shown that this echidna appears to be a true mammalian hibernator, with the ability to arouse from torpor without the aid of heat from external sources. This ability is lost after repeated periods of torpor in the laboratory at 5° C (41° F). The echidna resembles other hibernators in that torpor is not continuous but is interrupted by frequent arousals. Continuous periods of torpor

lasting 9.5 days, with intervening periods of wakefulness ranging from 30 hours to 11 days, have been recorded for the echidna.

The platypus has periods of hibernation or semihibernation, at least in southeastern Australia. In captive animals, Fleay found that the hibernation periods were short and irregular during the cooler months, the longest single period being 6.5 days.

Laboratory studies have shown that the platypus is slightly superior to *Tachyglossus* in temperature regulation. The platypus has sweat glands on the muzzle and over the body surface, whereas *Tachyglossus* has them only in the pouch area. The heat tolerance of *Zaglossus,* which has many well-developed sweat glands distributed over the body, has not been studied. In its natural environment it is unlikely that the platypus will be exposed to high heat loads. The highly insulating coat of the platypus would protect it against heat loss in water and allow it to exploit cold environments.

EVOLUTION, PALEONTOLOGY, AND CLASSIFICATION

The monotremes are without living relatives outside Australia and New Guinea, and, apart from the presence of fossil monotremes in the Australian deposits of the relatively modern Pleistocene Epoch (the last 2,500,000 years or so), little is known of their ancestry. The fossils so far discovered differ from the existing forms only at the species level. It is interesting to note, however, that large echidnas of the genus *Zaglossus,* which now live only in New Guinea, were common and widespread in Australia during the Pleistocene.

Although the existent monotremes are highly specialized in many of their features, their possession of such typically mammalian characters as mammary glands, hair, and a large brain carries the suggestion that such characters are primitive features of early mammals.

Origin of the order

Most authorities regard the order Monotremata as the sole group of the subclass Prototheria and believe that it originated from a line of mammal-like reptiles different from that which gave rise to the other mammals. It has been suggested, because of the similarities between marsupials and monotremes, that they should be grouped together in a subclass of the Mammalia, the Marsupionta, but the idea has not gained wide acceptance. It does not take into account the many similarities between marsupials and eutherian mammals. It has also been postulated that a group of early prototherian mammals, of which the monotremes are specialized derivatives, gave rise to the marsupials and the eutherian mammals, but this view is not accepted by most biologists.

The ducklike muzzle, or "bill," alone distinguishes the platypus from all other mammals. Although platypuses have been divided into several subspecies, based mainly on differences in size, all authorities agree that the family Ornithorhynchidae has only a single genus and species, *Ornithorhynchus anatinus.*

Two genera of echidnas are recognized in the family Tachyglossidae; in *Tachyglossus* (short-nosed echidnas) the elongated muzzle is straight and not greatly prolonged, and in *Zaglossus* (long-nosed echidnas) the elongated muzzle is greatly prolonged and curved downward.

Although Griffiths recognized *Tachyglossus aculeatus* as the only species of the genus, he distinguished six subspecies, some of which he considered of doubtful validity. Some authorities consider that the relatively hairy Tasmanian echidna is a distinct species (*T. setosus*). Some mammalogists who have studied *Zaglossus* in detail have concluded that only one variable species, *Z. bruijni,* should be recognized. It can be seen that much more work on the biology of the monotremes needs to be done to clarify the place these extraordinary animals occupy in nature.

(A.G.Ly.)

Marsupialia (kangaroos, bandicoots, phalangers, opossums, koala, wombats)

The Marsupialia is an order (or superorder, depending on the authority) of mammals characterized by premature birth and continued development of the newborn while attached to the nipples on the lower belly of the mother. The pouch, or marsupium, from which the group takes its name, is a flap of skin covering the nipples. Although prominent in many species, it is not a universal feature among marsupials; in some species, for example, the nipples are in a well-defined area but are fully exposed or are bounded by mere remnants of a pouch. The young remain firmly attached to the milk-giving teats for a period corresponding roughly to the latter part of development of the fetus in the womb of a eutherian, or placental, mammal.

The largest and most varied assortment of marsupials—more than 100 species—is found in Australia alone: kangaroos, wallabies, wombats, the koala, and a bewildering assemblage of smaller rodent-like forms. About 70 more species are distributed more widely, in Australia (including Tasmania), New Guinea, and a cluster of nearby islands. The wide array of Australian marsupials is reflected in the extensive popular vocabulary of names, many of which are derived from descriptive Aboriginal words. Only two families of marsupials—totalling more than 70 species—are found in the Americas, vestiges of a larger group that originated there as long ago as the Cretaceous Period (from 136,000,000 to 65,000,000 years ago). The family Didelphidae comprises about 65 species of South and Central American opossums, one of which ranges as far north as southern Canada. The family Caenolestidae consists of seven species of ratlike marsupials confined to South America.

GENERAL FEATURES

Scientific and economic importance. Marsupials are of interest to zoologists for several reasons: their current geographical distribution, their remarkable evolutionary expansion in Australasia, their similarity in many respects to placental mammals, and, most obviously, their reproductive adaptations and often bizarre structure.

The question of origin and spread

Fossil evidence indicates clearly that the marsupials originated in the New World, and although the oldest fossils referable to marsupials are found in North America, it is believed that South America is equally likely as the origin of these animals. The current concentration of diverse types of marsupials in Australia and its nearby islands is thought to have occurred as a result of passage over presumed land connections with South America during a mild early geological period before the rise of the placental mammals. Later, toward the beginning of the Tertiary Period (starting 65,000,000 years ago), Australia was isolated from all other continental masses, and the marsupials were free to evolve unrestricted by competition from other groups of mammals. Elsewhere, however, the marsupials did badly; they faced strong competition from and were supplanted by the more advanced placental mammals.

Australian marsupials provided products—meat and hides particularly—for the Aboriginal people, whose primitive hunting methods posed no threat to the continued success of the animals as a group. With the appearance of modern methods of hunting and trapping, however, several species of the kangaroo family (Macropodidae) were soon rendered extinct, and many others were brought close to that fate. For a time during the first half of the 20th century, many of the larger marsupials were slaughtered in great numbers for their pelts, which were an item of export in the fur trade; for their hides, which were made into shoes; and for their flesh, which was processed into dog and cat food. The gray kangaroo (*Macropus giganteus*), for example, is a casualty of such wholesale killing. Fortunately, in Australia most marsupials have since that time been accorded protection under law, and efforts are proceeding to conserve many species that were brought to the brink of extermination. The Tasmanian wolf, tiger, or thylacine (*Thylacinus cynocephalus*), thought to have become extinct in the 1930s, may still survive. In 1961 an expedition claimed to have found thylacine tracks in a rain forest on the west coast of Tasmania. The Tasmanian devil (*Sarcophilus harrisii*), once dangerously close to extinction, is now thriving in Tasmania. In the Americas marsupials are also hunted locally for food and other products, but, since numerous and equally desirable placental mammals abound, hunting pressure and the resulting threat of species death are not so strongly felt.

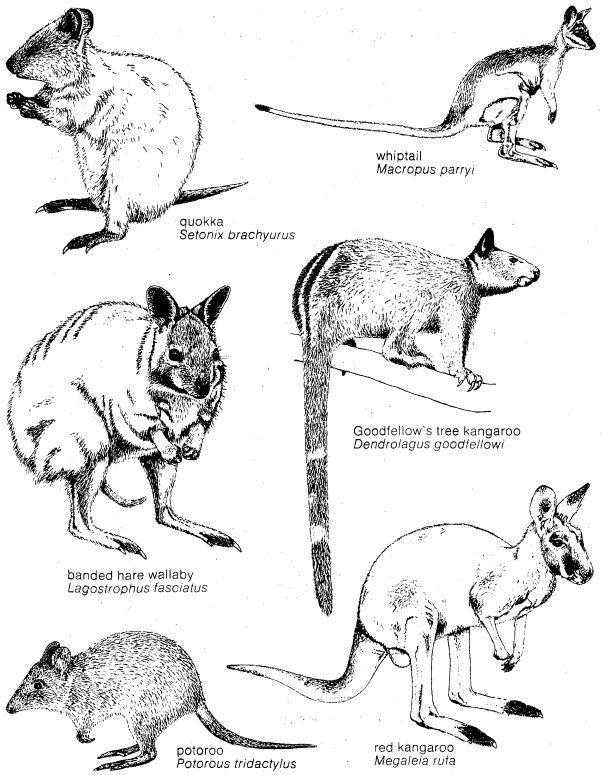


Figure 8: Representative species of the family Macropodidae.
Drawing by A.G. Lyne

Not only man but also his introduced animals have played a part in the extinction of certain marsupials. Foxes, dogs, and cats continue to prey upon many species of marsupials, as also do the dingoes, canines introduced by the Aboriginals long ago and since gone wild. Brush and forest fires as well as climatic change and the encroachment of civilization are also factors tending to reduce the numbers of marsupial species.

Size range and diversity. No extant marsupial even approaches the tremendous size of certain extinct forms, which apparently rivalled the mastodons in bulk. Among the largest living marsupial is *Macropus giganteus,* some individuals of which reach about two metres (six feet) in height, three metres (10 feet) from muzzle to tail tip, and weigh up to 90 kilograms (about 200 pounds); red kangaroos (*Megaleia rufa*) grow to about the same size. The smallest is the planigale (*Planigale ingrami*), a species of marsupial mice that is barely 12 centimetres (4.7 inches) in total length. The vast majority of marsupials lie in the range from the size of a squirrel to that of a medium-sized dog (Figures 8–10).

Parallels with placental mammals

Structural and behavioral parallels with placental mammals are in some cases quite striking. Such resemblances are examples of convergent evolution, a tendency for organisms to adapt in similar ways to similar habitats. Thus, there are marsupials that look remarkably like moles, shrews, squirrels, and mice. Others, less in structure than in habits, are the ecological counterparts of cats, small bears, and rabbits. Even the larger grazing marsupials, which resemble no placental mammal at all, can be thought of as filling the same ecological role as deer and antelope found elsewhere.

Distribution and abundance. There are about 175 species of marsupials native to the Australasian area, primarily in Australia (including Tasmania), New Guinea, Timor, and Celebes. Through man's agency, however, marsupials have been introduced to nearby islands of Australia and especially to New Zealand. The more than 70 native American species occur in South and Central America, with the range extended into North America by the common, or Virginian, opossum (*Didelphis marsupialis*). A number of other species extend into the southern portion of North America.

In terms of numbers of species the didelphids, or New World opossums, rank highest with 65. In sheer abundance, there are more vegetarian than carnivorous marsupials and perhaps more pouched mice than any other kind of marsupial.

The brush-tailed possums (*Trichosurus vulpecula*) are examples of marsupials that have readily adapted to changing conditions brought about by man. As recently as 1932, more than 1,000,000 pelts of this species were sold under such names as Adelaide chinchilla. Protected for part of the year, the brush tails have now become plentiful over the vast interior of Australia and even pestiferous in urban centres. Their adaptability to different locales is attributed to their tolerance for a variety of food, including household refuse.

Koalas (*Phascolarctos cinereus*), reduced to a few thousand in the 1930s as a result of disease and logging and trapping for fur, have recovered their numbers sufficiently to be reintroduced into forested regions.

Although the red kangaroos (*Megaleia rufa*) are being killed in large numbers for commercial products and by farmers and ranchers whose grazing lands they sometimes visit during periods of drought, they are numerous in many parts of inland Australia. The total range of this

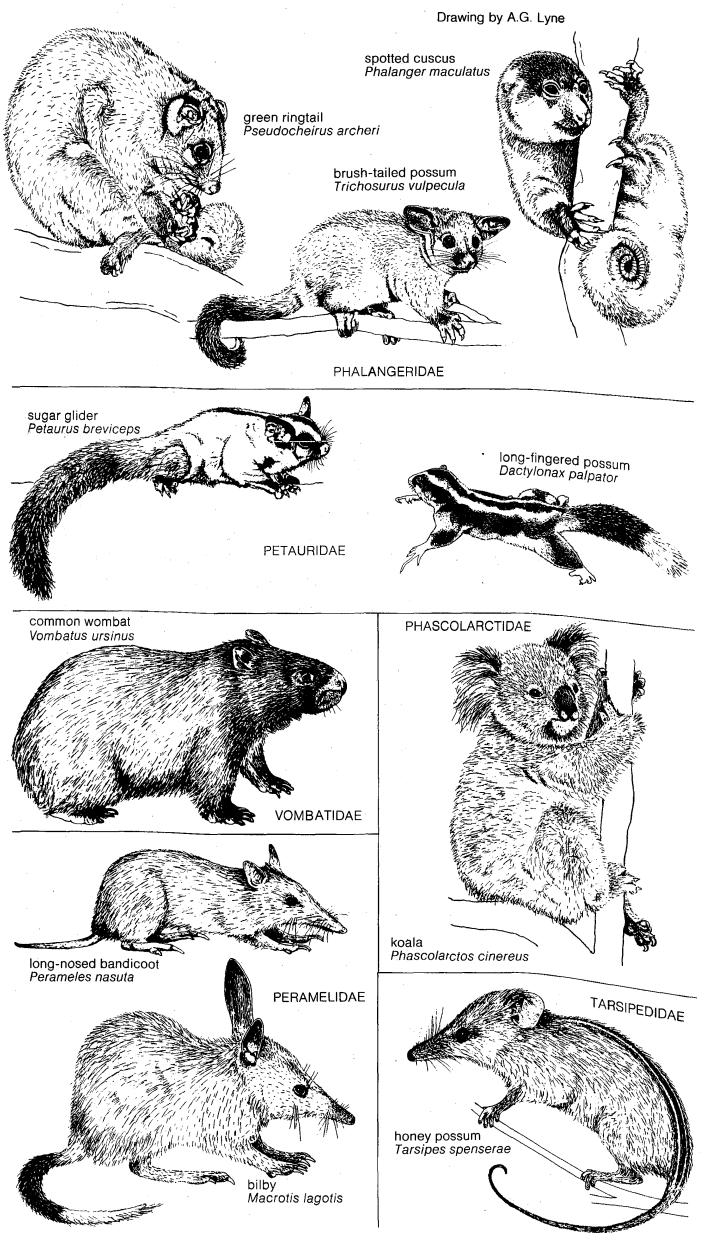


Figure 9: Representative species of the families Phalangeridae, Petauridae, Vombatidae, Phascolarctidae, Peramelidae, and Tarsipedidae.

species is about 2,000,000 square miles. Changes in vegetation resulting from domestic stock have improved the habitat for it in many parts of its range, and the species has increased in abundance.

NATURAL HISTORY

Life cycle. The life cycle of marsupials exhibits the peculiarities of a mammalian group advanced over the egg-laying monotremes but primitive to the placental mammals. The uterine cycle of the female marsupial has no secretory phase, and the uterine wall is not specialized for the implantation of the embryos. The period of intrauterine development in marsupials ranges from about eight days in the native cat *Dasyurus viverrinus* to 40 days in the red-necked wallaby (*Wallabia rufogrisea*). The young, born in a vulnerable embryonic condition, make their own way to the shelter, warmth, and nourishment of the pouch; in pouchless marsupials the young simply cling to the teats. Those fortunate enough to survive this arduous journey may succeed in attaching themselves to the mother's nipples, which then swell and become firmly fastened—almost physically fused—to the mouth tissues of the young. In this condition the young continue their development for weeks or months, after which they are weaned and begin to look after themselves. Frequently the partially developed young outnumber the available teats, and the excess individuals perish. In *Didelphis marsupialis,* litters of as many as 25 young have been reported, but the pouch usually contains only 13 teats.

The primitive reproductive cycle of *Didelphis marsupialis*

Didelphis marsupialis exemplifies one of the more primitive cycles and serves to illustrate the sequence of events in greater detail. It may have two litters yearly throughout most of its range (possibly only one in the northern reaches). The female experiences estrus (heat) and becomes receptive to the male about every 28 days during the breeding season. Heat lasts one or two days, during which an average of 22 ova are shed. Usually only 10 young are born in a litter.

During the first five or six days embryos form very slowly, but thereafter the rate of development speeds up. The gestation period is short—about 13 days. Just before birth, the mother thoroughly cleans the pouch and her belly fur, licking a smooth path from the birth canal to the pouch. When expelled from the birth canal, the newborn are no larger than honeybees. Blind and grublike, except for their well-developed, clawed forelimbs, they emerge and immediately begin grasping and swimming-like movements along the mother's dampened fur. The strong ones reach the pouch in less than a minute, an extraordinary feat for so small and incomplete a creature. Only about 60 percent of the newborn reach the pouch; these attach themselves immediately to the teats, at which they remain for four or five weeks (Figure 11). They then begin to leave the pouch for short intervals, scampering back to it whenever danger threatens. The young stay with the mother for 90 to 100 days—toward the last not in the pouch but clinging to her fur. Shortly afterward they are weaned and begin their independent lives.

Behaviour. The marsupials are notably less intelligent than placental mammals, a fact that is attributable in part to a simpler brain (see below *Form and function*). It is not surprising, therefore, to find a repertory of behavior that differs somewhat from that of the more advanced placentals. One peculiarity that may stem from this underdevelopment is restricted vocal ability. Although marsupials are not entirely silent, few of them emit loud sounds of excitement or distress; apparently, none utter grunts of contentment or even cries of hunger when young. What vocalizing they do is more limited and less variable than that of placentals.

There seems to be little detectable social organization among marsupials beyond the short-lived pair bonds during mating. The reproductive cycle in many marsupials is seasonally controlled, and estrus occurs in a predictable rhythm. In the quokka (*Setonix brachyurus*) and many other kangaroos and wallabies, however, the females may bear young at any time of the year. Only one young, called a joey, can be carried by the mother at a time. After suckling for about six months in the pouch, the joey begins to venture out but remains very close to the mother, plunging headfirst into the pouch when alarmed. The males have no interest in the young and will protect neither them nor the females against danger.

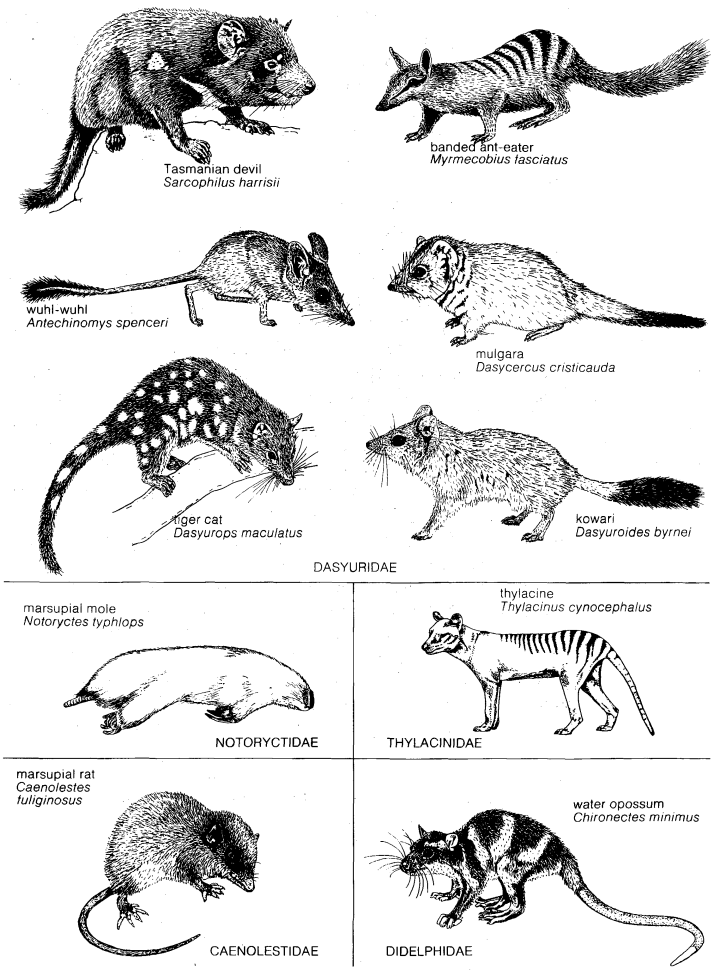


Figure 10: Representative species of the families Dasyuridae, Notoryctidae, Thylacinidae, Caenolestidae, and Didelphidae.
Drawing by A.G. Lyne

Many of the grazing marsupials, such as the kangaroos and wallabies, move in feeding groups called mobs, but these associations do not merit distinction as true herds since the individual members appear to move at liberty, without attention to any leaders or elders. One member can send the mob into a wild rout—individuals bounding off in all directions—by thumping its tail on the ground in a signal of alarm.

The larger marsupials, especially, embody nervous alertness and dull-witted consternation at the same time. A typical response of a red kangaroo when startled is to bound off in full flight—up to 65 kilometres per hour (about 30 miles per hour)—for a relatively short distance, stop short to reconnoitre the disturbance, and, in what appears to be sheer panic, double back toward its adversary and jump over it, sometimes clearing more than three metres (10 feet) in height and a distance of more than eight metres (25 feet) in a single leap.

Defensive behaviour of kangaroos

Although well equipped for escape, a boomer, or large male kangaroo, can stand fast against many of its foes. With agile arms it can spar vigorously, behaviour learned in playful bouts as a joey. But, much more effectively, it can use the forepaws to grip its enemy, while it rocks back on its tail and then swiftly drops its huge clawed hindfeet, an action that has been known to disembowel dogs and men. Despite such fearsome ability, the kangaroos are usually gentle, as are most of the vegetarian marsupials. Whatever violence they commit is usually in response to a distinct threat to themselves.

Such marsupials as the possums (families Burramyidae,

Petauridae, and Phalangeridae), wombats (Vombatidae), and koala (Phascolarctidae) usually sleep by day and move about toward dusk in search of food. Except for the agile possums—notably the gliders (*e.g., Acrobates, Schoinobates*)—most of these animals are slow-moving and, consequently, easy prey for carnivorous animals. The bush-tailed possum, however, is not easy prey for carnivorous animals.

The most mild-mannered marsupial is perhaps the numbat, or banded anteater (*Myrmecobius fasciatus*), a small, slow forest prowler who searches out termites, its favourite food. When disturbed, the numbat hisses; if caught it neither struggles nor bites but simply vents a few grunts in protest. Equally inoffensive and similarly vulnerable to predation is the teddy-bear-like koala, which rummages about at night among eucalyptus branches munching endlessly on the only food it eats—eucalyptus leaves. Its name, an Aboriginal dialect word meaning "no drink," refers to the koala's curious and apparently lifelong abstinence from water. Koalas provide a eucalyptus "soup" for their young at the time of weaning; it is lapped up directly from the parent's anus. The densely furred cuscuses (*Phalanger*) move among the trees with the same deliberation of the koalas, eating leaves as they go.

Bandicoots—whose apt name means "pig rat"—are especially fond of earthworms and insects and are only occasionally herbivorous. Although generally shy and retiring, bandicoots can at times display a surprising belligerence: with their sharp-clawed feet they can literally scratch their victims to death. Certain marsupial mice (*e.g., Sminthopsis* species) are so hyperactive—like shrews—that to supply their high energy needs they must devour their own weight in food, chiefly insects, each day.

A curious and effective defensive behaviour has been developed to a high degree in *Didelphis marsupialis* of North America: when harassed it hisses and departs, but when attacked and not able to flee, it feigns death (plays possum) so well that the aggressor may lose interest and leave.

The carnivorous marsupials are equipped with other behavioural traits. They are the fierce hunters of flesh. Although as a rule not as fast as comparable placental carnivores, they are persistent in the hunt and swift to kill. The Tasmanian devil is a stout fighter, likened in its bloodlust to the American wolverine. Yellow-footed marsupial mice (*Antechinus flavipes*) cling so tenaciously to their prey that they can be caught like fish on bait. The thylacine, which has been known to rise up on its hindlegs kangaroo fashion when pressed, makes up in perseverance what it lacks in speed and cunning. In sharp, foxlike snaps, thylacines have been known to dispatch an antagonist as formidable as a hunting dog.

Ecology. The niches, or ecological roles, that marsupials fill are closely associated with structure. The burrowing species, such as the marsupial mole and the wombats, have powerful foreclaws with which they can tunnel into the ground for food and for shelter. Terrestrial forms, such as the kangaroos and wallabies, possess well-developed hindlimbs that serve both as formidable weapons and catapults by which they can bound over the plains. The greater glider (*Schoinobates volans*) and other Australian "flying" possums have a membrane along either flank, attached to the forelegs and hindlegs, that enables these arboreal animals to glide down from a high perch. A few marsupials spend most of their lives in trees; koalas, for example, are so thoroughly arboreal that they seem comfortable only when they are among the branches of trees. One marsupial is semi-aquatic, the water opossum—yapok or yapó (*Chironectes minimus*)—of Central and South America. Its thick, oily fur, partially webbed feet, and constrictable pouch opening suit it admirably for swimming and diving in search of food.

Lifestyles

The diet of marsupials is as varied as the niches. Many live chiefly on insects and other small animals. Such are the marsupial mice and many of the smaller native cats (family Dasyuridae). The tiger cat and the Tasmanian devil feed largely on birds and small mammals. The numbat uses its remarkable wormlike tongue to lap up termites and ants. Many Australian possums, bandicoots, and American opossums have a mixed diet of plant matter and insects. Certain other marsupials are strictly vegetarian. The wombats feed on grasses, roots, and fungi, which they dig up with their clawed forepaws. The small honey possum (*Tarsipes spenserae*) is specialized to feed on the nectar of flowers.

The entire family of macropodids—the "big-footed" ones—are primarily grass eaters. They include the kangaroos and walleroo, the wallabies and pademelons, and the rat kangaroos and musk kangaroo. The walleroo, or euro (*Macropus robustus*), in particular, has a diet almost as restrictive as that of the koala: it can subsist in large part on spiky spinifex grass. Macropodids digest their vegetable food in the same manner as do sheep, cattle, and other ruminants, by relying on intestinal bacteria and protozoans to break down plant matter.

In their turn, marsupials are food for other animals. Foremost among the predators of didelphids and small dasyurids are owls and other birds of prey, snakes, and many carnivorous mammals, including dogs and cats. Some of the larger species of American opossums are eaten locally by man. The enemies of the larger marsupials, the macropodids, include the dingo, fox, eagle, pythons, goannas, and especially man.

Charles Philip Fox



Figure 11: Development of the common opossum. (Top left) In the pouch at three weeks; (bottom left) at seven weeks; (right) mother carrying 10-week-old young.

FORM AND FUNCTION

Marsupials share with other mammals the presence of hair and mammary glands. With minor differences the various systems of the body, such as the muscular and skeletal systems, are those of the placentals generally. The skull and the brain, however, differ considerably from those of placentals. Differences also exist in the dentition and in the arrangement of digits of the feet.

Bodily adaptations. In gross structure, many marsupials have hindlimbs that are larger and more powerful than the forelimbs; such an arrangement is most obvious in the large terrestrial grazing marsupials but is of wide occurrence in some degree throughout the Marsupialia. Arboreal and burrowing marsupials have forelimbs almost as well developed as the hindlimbs, understandable adaptations to their modes of living. Claws, often of considerable size, are invariably present in all marsupials and assist in climbing or digging. The tail, of tremendous size in kangaroos and wallabies, serves as an organ of stability and balance; in smaller and arboreal species it may also be prehensile (adapted for grasping), allowing the animal to hang by it.

Brain peculiarities

Compared with placentals, marsupials differ markedly in both the structure and bulk of the brain. Most notably they lack a corpus callosum, that part of the placental brain that connects the two cerebral halves. In addition, the marsupial brain is smaller relative to skull size: a marsupial cat has about half as much brain tissue as a placental cat of similar skull size. Such insufficiency may account in part for the relative backwardness of marsupials when contrasted with placental mammals.

A feature peculiar to marsupials is the presence of a pair of bones associated with the pelvic girdle. Called the epipubic, or marsupial, bones, they were earlier thought to be a supportive element for the pouch.

The teeth in marsupials are numerous—usually 40 to 50, but as few as 22 in the honey possum and as many as 52 in the numbat. The relationship of milk teeth to adult teeth is not clear; some authorities claim that a single set of teeth lasts throughout life. Typically, there are seven cheek teeth on each side of the jaw, top and bottom, which are divided into three premolars and four molars; this contrasts with the typical placental condition in which there are four premolars and three molars. Also unlike the placental condition, the number of front, or incisor, teeth differs in upper and lower jaws.

In earlier studies of marsupials much was made of the number and type of incisors (Figure 12). Marsupials with more than three incisors in the upper jaw were said to be polyprotodont and of a presumed more primitive type since the teeth were more or less unspecialized. Marsupials with three or fewer upper incisors and a corresponding set of modifications of the primitive condition were said to be diprotodont. Subsequent investigations disclosed a gamut of intermediate dentitions, thereby discrediting the distinction.

Another equally unsatisfactory distinction was based on the curious but minor condition in which the second and third toes of the hindfeet are covered in a common sheath of skin. Species showing that condition were syndactylous, and the remainder were didactylous. Classifications were founded on such matters, and the groups were named appropriately Polyprotodontia and Diprotodontia in one scheme, and Syndactyla and Didactyla in another (see below *Classification*).

Reproductive adaptations. The most extraordinary anatomical features of the marsupials are the specializations associated with the reproductive system. The most striking of these is the protective pouch around the nipples of the female in many species. The pouch is well developed in kangaroos, wallabies, and wombats; in the tuan (*Phascogale tapoatafa*) and some dasyurids it is poorly developed, represented by lateral abdominal folds. In several South American didelphids, in the rat opossums (family Caenolestidae), and in the numbat, the pouch is lacking entirely. Even among species that have pouches, the degree of development depends upon the breeding condition of the animal, being most obvious when the female is lactating. The pouch in most forms opens forward, especially in upright and arboreal species; in the water opossum and many quadrupedal species the pouch is directed backward. There may be as few as two nipples (as in the koala and wombats) or as many as 27 (as in *Monodelphis,* the short-tailed opossum). The number of teats does not indicate the number of young that are nursed, however; in some cases more young are born than can be suckled at one time, but in most cases fewer young than the number of nipples are brought to full term.

Differences in pouches

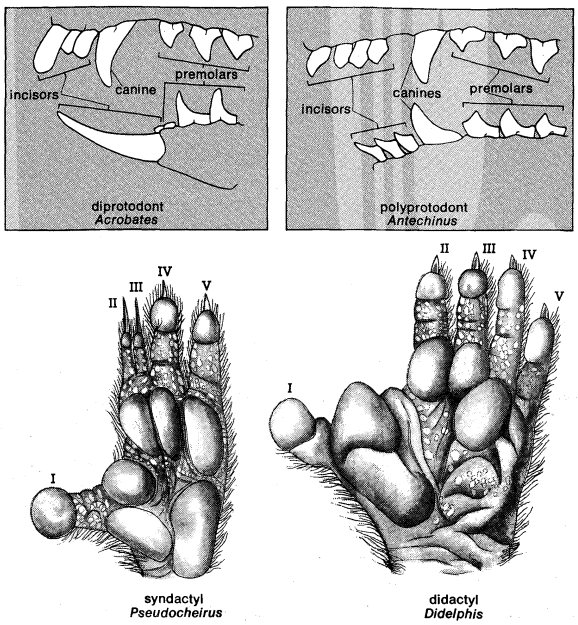


Figure 12: Dentition and fusion of the digits of the hindfoot of marsupials. (Top left) Diprotodont in the pygmy gliders and (top right) polyprotodont in the broad-footed pouched mice. (Bottom left) Digits II and III syndactyl, digit I opposable, in the ring-tailed opossums and (bottom right) digits II and III didactyl, digit I opposable, in the common opossum.

By courtesy of Owen J. Poe

The female reproductive tract is double for most of its length, becoming joined at the posterior ends of the two lateral vaginae. A birth canal, or median vagina, forms at the time of birth; copulation is accomplished by the insertion of the male's penis into the urogenital sinus of the female. The end of the penis is forked in some species. In addition to the peculiar penis, the male also has a scrotum in which the testes hang in front of the penis rather than behind; the testes are abdominal in the pouched male (*Notoryctes*). Most marsupials specialize in brief gestation, as indicated earlier. The developing embryo obtains its food chiefly from its own yolk sac, an embryonic organ provided by the mother. In the majority of marsupials, embryonic gaseous exchange is also effected by means of the highly vascular yolk sac. Typically, there is no intimate mother–fetus connection by means of a true (chorioallantoic) placenta, as in placental mammals. The bandicoots are unique among living marsupials in having a rudimentary placenta, but it lacks the extensive ramifying projections (villi) that provide the close connection of mother and fetus tissue in the placentals.

EVOLUTION AND PALEONTOLOGY

Geologically, no marsupials have been discovered in Asia, and they are definitely not known from Africa. On the whole they are considered as primitive mammals, but for some 100,000,000 years they have shown a remarkable parallel and convergent evolution with placental mammals in habits, physiological processes, and structures.

Among the oldest marsupials are those represented by undisputed fossils from the Late Cretaceous strata of North America. These specimens are represented mostly by tiny isolated teeth, but parts of jaws have also been found. Until recently, all of the fragments were classified in the American opossum family, Didelphidae, but it has been observed that these early marsupials were already specialized in different directions. *Alphadon,* a genus of primarily

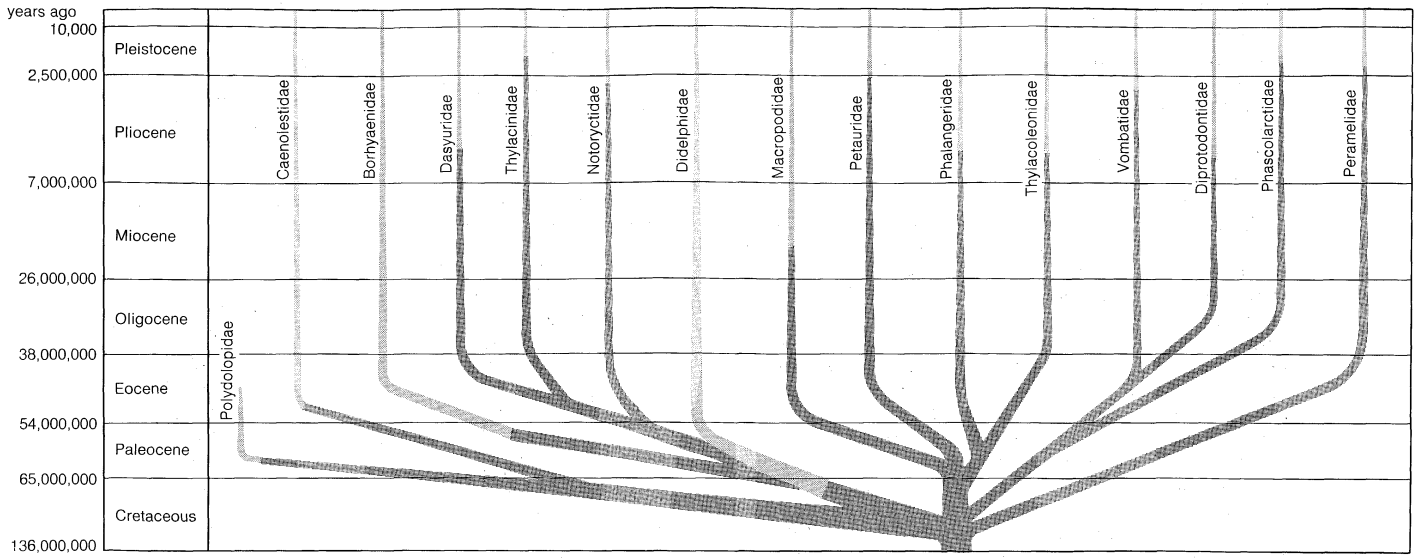


Figure 13: Dendrogram of the marsupials.

tiny animals, is most likely a didelphid. On the other hand the group of small to medium-sized creatures, represented by the genus *Pediomys,* seem rather clearly not didelphids and are now placed in their own family, the Pediomyidae. The widely known genus *Eodelphis* is almost certainly not didelphid. Among its more conspicuous differences from the Didelphidae are the single- (not double-) rooted, much reduced first lower premolar and the presence of three, not four, lower incisors, of which the middle incisor is greatly enlarged. It is likely that this marsupial is not ancestral to any of the known later genera. Another specimen, known as *Thlaeodon,* is the largest known Cretaceous mammal. Its bulbous premolars and large canines are indicative of an adaptation to a mixed (omnivorous) diet. It apparently is referable to another family. *Alphadon,* earlier thought to be nearer than any of the other Cretaceous genera to an ideal ancestral position for the marsupials as a whole, has been superseded by later finds. Fossil fragments of even earlier date (the Middle Cretaceous of Texas, about 100,000,000 years ago) have been tentatively referred to as *Holoclemensia* and assigned as a new genus in the family Didelphidae.

Evidence suggests that marsupials became widespread in the Tertiary Period, beginning 65,000,000 years ago, but they never succeeded as well on other continents as they did in South America and Australasia.

Primitive marsupial characteristics

American marsupials. *Didelphoids.* Primitive marsupial characteristics in the superfamily Didelphoidea have long been recognized by mammalogists. Some of the features in the dentition, skull, and body skeleton may be reasonably expected to be remarkably like those in the Mesozoic ancestors of all later marsupials. Characters in the molars indicate that didelphoids and the South American borhyaenoids of the Paleocene Epoch (beginning 65,000,000 years ago) and Eocene Epoch (beginning 54,000,000 years ago) are closely related, but the lack of enough fossils makes it impossible to trace the lineage of all marsupial superfamilies back to the known North American Cretaceous fossils (possibly the more direct ancestors of some of the other superfamilies existed at that time). Better knowledge of the skeletons and dentitions of fossil marsupials from the Cretaceous Period and even the Early Cenozoic Era (beginning 65,000,000 years ago) may eventually alter considerably the present interpretation of their superfamily relationships.

Borhyaenoids. The oldest known South American marsupials are of the later Paleocene Epoch (about 55,000,000 years old). At that time three superfamilies, Didelphoidea, Borhyaenoidea, and Caenolestoidea, were already in existence. Patterns of the known cheek teeth of the Paleocene and early Eocene borhyaenoids and didelphoids indicate that these marsupials may have descended from a common ancestor no earlier than the Late Cretaceous.

The fossil record of the family Borhyaenidae starts in the late Paleocene and continues into the Pliocene (ending 2,500,000 years ago). No genera are known to have occurred outside South America. Borhyaenids were the flesh-eating mammals of the southern continent during nearly all the Tertiary Period. (True carnivores presumably did not reach that region until late in the Pliocene and Pleistocene.) The family name is derived from the genus *Borhyaena,* hyena-like specimens found in the early Miocene strata (about 26,000,000 years ago) of Argentina. These carnivorous marsupials had massive skulls with thick crushing teeth. Some genera—*e.g., Proborhyaena*—were shorter faced and more robust.

All borhyaenids were not hyenoid, however. There were the wolflike *Prothalycinus* and *Lycopsis* and the fox- to marten-sized *Cladosictis* and *Amphiproviverra.* One of the most specialized of all was the Pliocene sabre-toothed borhyaenid *Thylacosmilus;* it was about the size of a puma and even more specialized as a stabbing mammal than most sabre-toothed cats. In all, more than 20 borhyaenid genera have been described, revealing a long evolutionary history in South America from a Cretaceous didelphid-like mammal.

Necrolestoids. One fossil with a molelike adaptation, called *Necrolestes,* is known from the Miocene of Patagonia. It is the only known genus that can be referred to the family Necrolestidae (superfamily Necrolestoidea) and is probably basically related to the Borhyaenoidea or Didelphoidea.

The early rat opossums

Caenolestoids. The South American rat opossums (superfamily Caenolestoidea) are also undoubtedly descendants of a Cretaceous group, but the oldest known caenolestoids are already too specialized to reveal affinities with any of the known Mesozoic fossils. Any resemblances they have to Australian groups apparently result from convergent evolutionary trends in certain structures.

The family Caenolestidae has been recorded from early Eocene deposits, but the genera were more numerous later in the Oligocene and Miocene periods. These little marsupials were long known from fragmentary fossils, but in 1895 this supposedly extinct family was discovered as represented in the living Andean mammalian fauna of Ecuador and Bolivia by an animal resembling a small rat in size and appearance and named *Caenolestes.* The diprotodont incisor teeth are suggestive of those in some of the Australian phalangeroids, but the hindfeet show no trace of the phalangeroid syndactylism of the second and third toes. Other genera with living species are *Orolestes* in Peru and *Rhyncholestes* in Chile.

The Polydolopidae is an Early Cenozoic family of small specialized marsupials related to the Caenolestidae. Their fossil remains are abundant in the South American strata of that time. They had transversely compressed and laterally ridged premolars, with serrated crests, and rodent-like incisors. The polydolopids, probably in a large measure,

occupied the ecological niches that were later occupied by rodents.

Australasian marsupials. Although pre-Pleistocene fossils of Australian marsupials are relatively rare, representatives of the families Dasyuridae, Perameli-dae, Phalangeridae, Thylacoleonidae, Macropodidae, and Diprotodontidae appear as fossils in strata as early as the late Oligocene Epoch (26,000,000 years ago). During the Miocene and Pliocene epochs that followed, fossils of the family Vombatidae were laid down. These findings suggest that the Australian families of marsupials originated in the Early Cenozoic.

Dasyuroids. The oldest dasyuroid lived during the Middle Tertiary of South Australia; about the size of the present-day *Dasyurus quoll,* it was probably near the stem from which the thylacine evolved. *Glaucodon ballaratensis,* one of the most important fossil dasyurids, is possibly of Pliocene age; it bore a resemblance to both *Dasyurus* and a Pleistocene member of *Sarcophilus.* The thylacine (*Thylacinus cynocephalus*), which constitutes the family Thylacinidae, is the largest and most widely known of the dasyuroid marsupials. Although its distribution is usually recorded as being restricted to Tasmania, it was represented on the mainland of Australia less than 10,000 years ago and also lived on New Guinea. The last certainly known thylacine was captured in Tasmania in 1930; however, sightings have been recorded of thylacines and their tracks as late as 1961 along a wild stretch of the west coast of Tasmania.

The doubtfully extinct thylacine

The family relationships of the thylacine are still not fully understood. (Its classification with the extinct Borhyaenidae of South America has almost conclusively been disproved.) Its characters, however, display a remarkable convergent evolution with the American borhyaenids. It seems reasonable to assume that the thylacine and the dasyures arose from a common ancestry. Many authorities place *Thylacinus* in the family Dasyuridae, but until its ancestry is adequately known it seems best to treat it as forming a separate family.

Without a fossil record it is not possible to conclude how closely *Notoryctes,* constituting the family Notoryctidae, is related to the basal stocks of the dasyuroids and to that of the bandicoots, but the absence of syndactyly in the hindfeet suggests they are closer to the dasyuroids. *Myrmecobius,* also lacking a fossil record, is grouped with the family Dasyuridae because of similarities.

Perameloids. Members of the superfamily Perameloidea display one of the most primitive characters seen in any Australian marsupials: the retention of an incisor formula of $\frac{5}{3}$ (polyprotodont) in most of the genera. They have syndactylism in the hindfoot: the proximal and median phalanges (bones) of the second and third digits are enclosed in the same sheath of skin.

Among the perameloids, both *Macrotis* and *Perameles* are represented in the late Pleistocene faunas of Australia. An extinct genus, *Ischnodon,* has been found in the Pliocene strata of South Australia. The most peculiar of all peramelids is the probably extinct (it has not been seen since about 1926) pig-footed bandicoot (*Chaeropus ecaudatus*). The common name is derived from the construction of the front foot, in which the second and third digits are of equal length and closely united. The nails, of equal size and length, give the appearance of cloven hooves.

Phalangeroids. Only a very hypothetical phylogeny, indicating possible ancestral relationships of the superfamily Phalangeroidea to the other superfamilies, can be outlined until an adequate fossil record from the Late Mesozoic and Early Cenozoic has been compiled. The Phalangeroidea were probably derived from an ancestral stock that also gave rise to the Perameloidea. Evidently, in this early stock (in contrast to the Dasyuroidea) natural selection was toward syndactylism. Subsequently, one syndactyl stock led to the ancestors of the rat kangaroos, kangaroos, diprotodontids, wombats, koalas, marsupial lions, and phalangers; the other stock gave rise to the bandicoots.

Of the phalangers, *Eudromicia* is the most primitive of all the living genera. It comprises in part the dormouse possums. The retention of the fourth molars and the double-rooted first and second upper premolars distinguish *Eudromicia* from the other dormouse possums, *Cercartetus,* which they notably resemble.

The remains of a tiny marsupial called *Burramys* has been found in New South Wales, Australia. The shape of the skull and the pattern of the molars resemble those in *Cercartetus* (including *Eudromicia*), but the premolars are high and serrated. This genus is the basis for the family Burramyidae. Once thought to be extinct, several specimens of *Burramys* have been found since 1966 in northeastern Victoria.

One of the oldest fossil marsupials of Australasia is part of the skeleton of a brushtail-like possum, Wynyardia *bassiana,* found in late Eocene or Oligocene deposits near Wynyard, Tasmania. It constitutes the family Wynyardidae. Fossils more closely related to later species occur in the Pleistocene of Australia; others much older occur in the Middle Tertiary of South Australia.

The oldest Australasian fossil

Among the most bizarre of all marsupials are the woolly possums, or cuscuses (*Phalanger*), from which the family and superfamily names are derived. *Wyulda,* the scaly-tailed possum, which is terrestrial and lives among rocks in northwestern Australia, is somewhat intermediate between the brush-tailed possums and the cuscuses. There has been much discussion about the relationships of the so-called marsupial lion (*Thylacoleo*). Most authorities agree that it belongs in a separate family, the Thylacoleonidae. Specimens of the cranium and lower jaws are well-known, and parts of the limb bones and body skeleton have been found. *Thylacoleo* was about the size of an African lion. Its teeth were even more specialized for meat shearing than those in the true cats; for example, the large posterior premolars were as long as 5.7 centimetres (2.24 inches). All the remaining teeth were greatly reduced, except the upper and lower median incisors, which apparently were utilized in capturing other animals. These great carnivorous marsupials became extinct in Australia less than 26,000 years ago.

Although the genera referred to as the rat kangaroo family (Potoroidae) are closely related to the kangaroo family (Macropodidae) and are classified by some as a subfamily of that group, they apparently have represented a distinct lineage since Eocene time. They may be distinguished from kangaroos by, among other features, several peculiarities of dentition and skull features. The female urogenital system is more specialized than in the Macropodidae. A Pleistocene genus, *Propleopus,* is known, and a species of *Bettongia* and an undescribed genus occur in the Middle Tertiary of South Australia.

The most widely known marsupials are the kangaroos and their relatives (family Macropodidae). The gigantic forms during the Pleistocene were *Procoptodon, Sthenurus, Protemnodon,* and two species of *Macropus. Prionotemnus* is from the Pliocene, but the oldest kangaroos come from rocks probably as old as Miocene.

The macropodids are adapted for jumping. The principal digit in the hindfoot is the fourth, the fifth being somewhat reduced; the syndactyl second and third are reduced to splinters but still bear tiny claws. *Procoptodon* and *Sthenurus* differ from the other kangaroos in having lost all but the fourth digit.

Vombatoids. The superfamily Vombatoidea includes the diprotodonts, the koala, and the wombats.

The extinct family Diprotodontidae included the largest of all marsupials, *Diprotodon optatum.* It was built something like a huge ground sloth and attained the size of a large rhinoceros. Special dentition included the large, chisel-shaped median incisors and molar teeth, somewhat like those in certain kangaroos or tapirs, composed of two cross crests. *Diprotodon* was probably abundant in Australia in the late Pleistocene and subrecent time. Its remains have been found in levels contemporaneous with man. At least two other genera, *Nototherium* and *Euowenia,* representing much smaller diprotodontids, also occur in Australian Pleistocene faunas. *Palorchestes,* also of the Pleistocene, previously considered as the largest of all kangaroos, is now known to be a diprotodontid.

The massive diprotodonts

This family, like the Macropodidae, had a long Cenozoic history in Australia. *Meniscolophus* is known from the Pliocene, and remains of a rather primitive undescribed

genus and species are thought to be as old as the Oligocene Epoch. It appears that all known diprotodontids were herbivorous. The family may have been an early offshoot of the phalangeroid stock and perhaps was distantly related to kangaroos and wombats.

One of the most famous of all Australian marsupials is the koala, family Phascolarctidae. Only one species is known, although a rather distantly related genus, *Perikoala,* from the Tertiary of South Australia, has been described. The koala, with its short tail and rather stocky body, so resembles the wombat that the two are considered by some authorities to have arisen from a common ancestor, perhaps not very far back in Cenozoic time. Others believe, however, that the koala is much closer to the ringtails and greater gliders. The wombats (family Vombatidae), which resemble koalas in having no tails, are represented by two Pleistocene fossil genera: the giant wombat, *Phascolonus,* which was as large as a black bear, and *Ramsayia.* (Ed.)

CLASSIFICATION

Distinguishing taxonomic features. Two earlier classifications of marsupials were based on the number and arrangement of the front teeth and on the appression or separateness of certain toes. Marsupials having at least four upper incisors were named Polyprotodontia; those having three or (usually) fewer upper incisors and no lower canines, Diprotodontia. Marsupials having entirely separate toes on the hindfeet were called Didactyla; those having the second and third toes of the hindfeet enclosed in a common envelopment of skin, Syndactyla. These terms no longer have taxonomic validity, but their adjectival forms are still used to describe the dentition and the character of the hindfeet.

Additional aspects, such as the pattern of hair tracts, blood chemistry, details of reproductive anatomy and physiology, chromosomal morphology, sperm morphology, and behaviour, are currently being employed to structure a new taxonomic system representing more natural units that reflect what are presumed to be fundamental phylogenetic relationships.

Annotated classification. The time-honoured position of the Marsupialia as an order of mammals, generally unacceptable for a variety of reasons (see below *Critical appraisal*), is gradually giving way to a system first presented in 1964. As further modified, that arrangement raised the Marsupialia to the rank of a superorder and encompassed four orders, 10 superfamilies, and 23 families, as given below. Wholly extinct groups are preceded by a dagger (†).

SUPERORDER MARSUPIALIA (Metatherian, or pouched mammals)

Origin during the Cretaceous, expansion at beginning of the Tertiary, and decline with the rise of placentals thereafter. Currently established in Australia and nearby islands (introduced on New Zealand) and in the Americas. Young born after brief gestation in an embryonic condition; development completed while young are attached to teats, usually in a pouch, or marsupium. Skull elements differ from placentals; braincase small; brain relatively small and simple compared with placentals. Epipubic bones associated with the pelvic girdle. Fewer incisor teeth in lower than in upper jaws. Female reproductive tract doubled in large part (paired vaginae and uteri); testes usually in a scrotal sac and anterior to the forked penis (abdominal testes in *Notoryctes*). Comprises about 80 genera and about 240 species in 4 orders.

Order Marsupicarnivora

Comprises 3 extinct superfamilies and 2 extant superfamilies of primarily carnivorous marsupials.

†*Superfamily Argyrolagoidea*

Comprises a single family, Argyrolagidae, represented by *Argyrolagus* from the upper Pliocene to the lower Pleistocene in South America.

†*Superfamily Borhyaenoidea*

Comprises a single extinct family, Borhyaenidae, flesh-eating marsupials of South America during the late Paleocene and into the Pliocene Epoch (about 5,000,000 years ago). Many species had massive skulls and heavy crushing teeth. Named after hyaena-like specimens from Argentinian fossils. Examples: *Prothalycinus, Borhyaena, Lycopsis,* and the sabre-toothed marsupial *Thylacosmilus.*

Superfamily Dasyuroidea

Primarily terrestrial carnivores that resemble the didelphids in certain anatomical features. Incisors small and unspecialized (polyprotodont). Digits in hindfeet never joined (didactylous). Marsupium not present in all species; when present it is poorly developed and opens backward. Three families ranging from Australia to nearby islands.

Family Dasyuridae. A group widespread over Australasia and represented by marsupial mice and rats (*e.g., Sminthopsis, Antechinus, Antechinomys, Planigale, Murexia,* and *Phascogale*), native cats (*Dasyurus, Dasyurops*), Tasmanian devil (*Sarcophilus*), numbat (*Myrmecobius*). Includes the smallest known extant marsupial, *Planigale ingrami* (about 12 centimetres [4.7 inches] in total length). About 18 genera and 48 species.

Family Notoryctidae (marsupial moles). Mole-sized marsupials, of the deserts of central and western Australia, remarkably convergent with placental moles of other continents. Pouch opens backward, and there are epipubic bones. One genus, *Notoryctes,* and 2 species.

Family Thylacinidae (thylacines). German-shepherd-sized wolflike marsupials. Represented on the mainland of Australia less than 10,000 years ago and in New Guinea, the thylacine was last captured in Tasmania in 1930. A single species, *Thylacinus cynocephalus,* constitutes the family.

Superfamily Didelphoidea

Primarily terrestrial American marsupials. Earliest fossils from Cretaceous strata. One extant family, Didelphidae, and 2 fossil families.

Family Didelphidae (American opossums). Central and South America boasts many species of unusual adaptations, notably *Chironectes* (water opossum), *Philander* (4-eyed opossums), *Metachirus,* and *Lutreolina. Didelphis* ranges as far north as southern Canada. Includes 12 genera and about 65 species.

†*Family Pediomyidae.* One genus, *Pediomys,* from the Upper Cretaceous in North America.

†*Family Stagodontidae.*

†*Superfamily Necrolestoidea*

Comprises 1 family, Necrolestidae. Based upon molelike fossils from the Miocene (starting 26,000,000 years ago) of Patagonia. One genus, *Necrolestes.*

Order Paucituberculata

Comprises a single superfamily of marsupials.

Superfamily Caenolestoidea

South American terrestrial marsupials. Earliest fossils from the early Eocene Epoch (about 50,000,000 years ago). One extant family, Caenolestidae, and 2 extinct families.

Family Caenolestidae. Ratlike in size and appearance. Diprotodont incisors suggest relationships to Australian phalangeroids, but other evidence disputes such affinities. Includes *Caenolestes* (Ecuador and Bolivia), *Orolestes* (Peru) and *Rhyncholestes* (Chile). Comprises 3 genera and 7 species.

†*Family Groeberiidae.* One genus, *Groeberia,* from the late Oligocene.

†*Family Polydolopidae.* Known from several genera including *Polydolops* and *Amphidolops* from upper Paleocene to late Eocene in South America.

Order Peramelina

Comprises 1 extant superfamily with 1 family of primarily carnivorous marsupials. Specimens of modern species are common in Pleistocene cave deposits in various parts of Australia.

Superfamily Perameloidea (bandicoots)

Comprises 1 family, Peramelidae. Terrestrial Australian marsupials that resemble rodents, rat-sized to hare-sized. Most genera have primitive polyprotodont dentition: 5 incisors above and 3 below. Syndactylism in hindfeet. Pouch is directed backward. Eight genera and 22 species, including: *Peroryctes, Microperoryctes,* and *Rhynchomeles,* primitive genera restricted to New Guinea and adjacent islands; *Perameles* and *Isoodon,* of wide distribution; *Macrotis* (bilbies) and *Chaeropus* (pig-footed bandicoot), restricted to Australia. Largest of the family (weight to 7 kg, or 15 lb.) is *Peroryctes broadbenti.*

Order Diprotodonta

Comprises 3 extant superfamilies of primarily herbivorous marsupials.

Superfamily Phalangeroidea

Australasian marsupials ranging from squirrel-sized arboreal species to large terrestrial bounders. Syndactylism of the second and third digits in the hindfeet. Four extant families; 2 extinct families.

Family Burramyidae. Primarily arboreal mouse- to squirrel-sized marsupials. Includes *Acrobates* (feathertail gliders), *Bur-*

ramys, and *Cercartetus* (pigmy possums). Comprises 3 genera and 6 species.

Family Macropodidae. Primarily terrestrial medium- to large-sized Australian marsupials. Adapted for jumping, with long hindlegs and long tail for balance. Forelimbs have sharp claws, and thumbs are not opposable. Main digit in hindfoot is the fourth. Extinct giant forms occurred during the Pleistocene, but oldest forms can be traced to the Miocene. Extant macropodids include *Macropus* (gray kangaroos and wallaroos), *Megaleia* (red kangaroos), *Wallabia* (wallabies), *Thylogale* (pademelons), *Dendrolagus* (tree kangaroos), and *Setonix* (quokkas). There are about 19 genera and 47 species.

Family Petauridae. Terrestrial and arboreal marsupials. First and second digits of the forelimbs are opposable to the other digits. Molars are adapted for chewing leaves. Includes *Pseudocheirus* (ring-tailed possums), *Petaurus* (sugar gliders), *Schoinobates* (greater gliders), *Gymnobelideus* (Leadbeater's possum), and *Dactylopsila* (striped possums). About 5 genera and 25 species.

Family Phalangeridae. Squirrel-sized to cat-sized arboreal species. Includes *Trichosurus* (brush-tailed possum), *Phalanger* (cuscuses), and *Wyulda* (scaly-tailed possum). About 6 genera and 15 species.

†*Family Thylacoleonidae.* So-called Australian marsupial lions, which became extinct less than 26,000 years ago. Teeth remarkably specialized for meat shearing. One genus, *Thylacoleo.*

†*Family Wynyardiidae.* Brushtail-like possums, the oldest fossil marsupials of Australasia, from presumed late Eocene rocks of Tasmania. One species, *Wynyardia bassiana.*

Superfamily Tarsipedoidea (honey possums)

Comprises a single family, Tarsipedidae. Adapted for feeding on nectar of flowers. One species, *Tarsipes spenserae,* of southwestern Western Australia.

Superfamily Vombatoidea

Terrestrial and aboreal Australian marsupials, syndactylism of the second and third digits of the hindfeet. Two extant families; 1 extinct family.

†*Family Diprotodontidae.* Large herbivorous marsupial represented in the Australian fossil record as far back as the Miocene.

Family Phascolarctidae (koalas). Small bearlike Australian marsupials. Adapted to arboreal living and restricted to a diet of the leaves of a few species of eucalyptus. First and second digits of the forelimbs are opposable to the other digits. There is no tail. Pouch opens backward. One species along the eastern Australian coast: *Phascolarctos cinereus.*

Family Vombatidae (wombats). Woodchuck-like marsupials distinct from all others in having a single pair of ever-growing upper and lower incisors and in having rootless, high-crowned cheek teeth. The tail is very short. Includes *Phascolonus,* the giant wombat of the Pleistocene; *Vombatus* (common, or naked-nosed, wombat); *Lasiorhinus* (hairy-nosed wombat). Comprises 2 genera and 4 species.

Critical appraisal. An earlier classification, which is still encountered (see section *The class Mammalia*), attributes ordinal rank to the marsupials. Under that system are listed 6 superfamilies and from 13 to 18 families, according to the authority followed. Various attempts to subdivide the order into suborders to accommodate necessary changes required by accumulated taxonomic findings have not been widely accepted. Along with a resurgence of interest in marsupial relationships has developed a growing body of opinion that the order rests too heavily on the presence of a pouch or its vestige. Such a foundation is as unwarranted as the reduction of all recognized orders of placental mammals into a single order because they all possess a placenta. In fact, as many variations and as great a degree of variations are found within the marsupials as are found in the placental mammals. Current opinion, therefore, favours the recognition of the Marsupialia as a superorder comprising four orders as given above.

(H.M.V.D.)

Insectivora (shrews, moles, hedgehogs, tenrecs, solenodons)

The mammal order Insectivora includes the shrews, moles, hedgehogs, and several lesser known groups. Of about 400 species in the order, nearly 300 are shrews of the family Soricidae. All insectivores are small, as mammals go, the largest being about the size of a small rabbit.

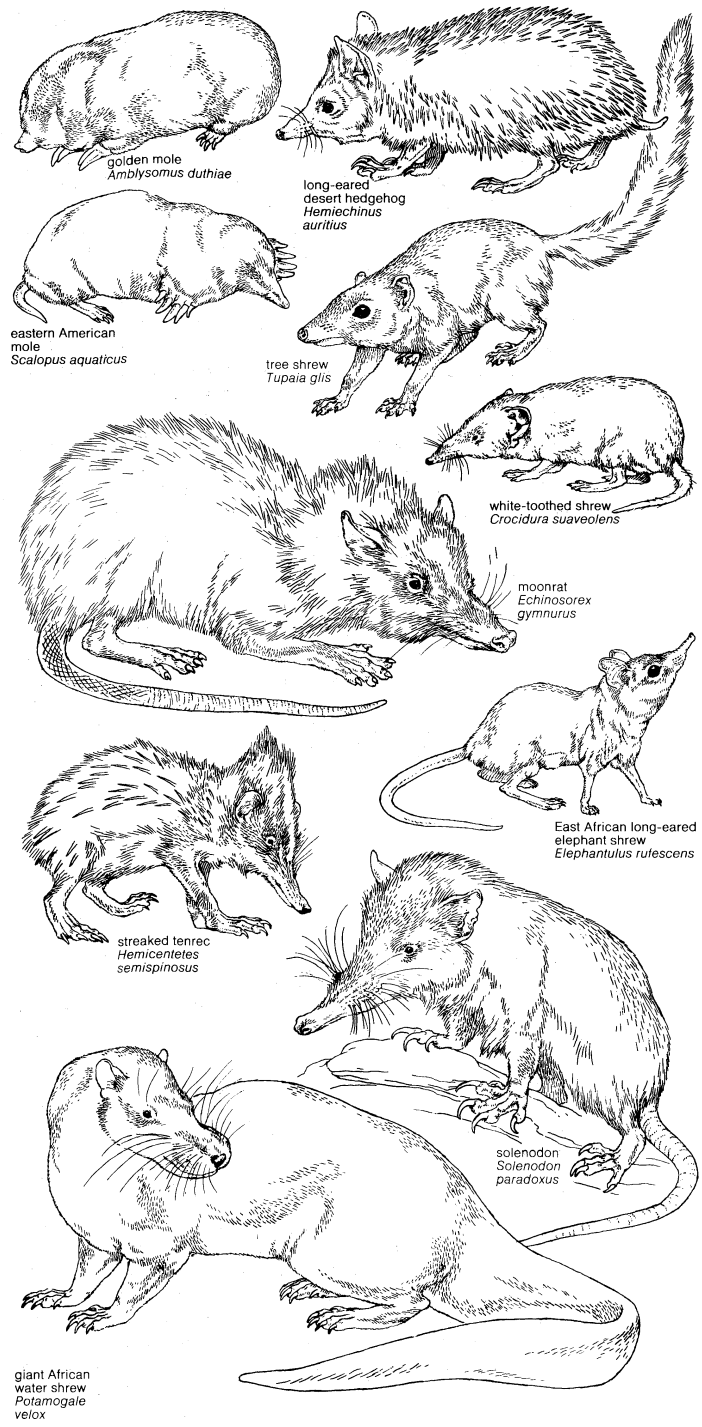


Figure 14: Representative insectivores.

Drawing by R. Keane based on photos courtesy of (golden mole) Herbert Lang through J. Meester, (eastern American mole) *U.S. Fish and Wildlife Service,* (white-toothed shrew) Liselotte Dorfmuller, (streaked tenrec) Howard E. Uible, (solenodon) New York Zoological Society; from (long-eared desert hedgehog, tree shrew, East African long-eared elephant shrew) E.P. Walker, *Mammals of the World,* (giant African water shrew) Grzimeks Tierleben, *Enzyklopadie des Tierreiches;* Kindler Verlag Zurich

Shrews are probably the smallest of all mammals; some species of *Sorex* and *Crocidura* weigh as little as 2.5 grams (0.09 ounce). The general body plan (Figure 14) varies greatly within the order. Many shrews are superficially mouselike but may be distinguished from mice and other rodents by the absence of chisel-like incisor teeth. Typical moles (Talpidae) and golden moles (Chrysochloridae) are cylindrical, burrowing animals with short, thick limbs and reduced tails and eyes. Hedgehogs (Erinaceidae) and some hedgehog-like tenrecs (Tenrecidae) are stocky animals whose upper surfaces are covered with short spines. Elephant shrews (Macroscelididae) are rather delicate animals, with long hindlegs like those of kangaroo rats or

jerboas. Within the larger families the body plan is quite variable. Most insectivores live at or near ground level, except for the squirrel-like tree shrews and several aquatic moles, shrews, and tenrecs.

Importance to man. Small and secretive, most insectivores are rarely seen by man and are of almost no economic importance. The velvety fur has poor durability, but mole skins have occasionally been used for trimming various garments and for complete coats.

The burrows of moles in lawns may cause the collapse of the turf, and those in fields and gardens provide small rodents with access to plant roots. Indirectly, the more abundant insectivores are ecologically important as consumers of invertebrates and occasionally of small vertebrates and as prey for larger animals. None is an important disease carrier, although elephant shrews harbor malaria, and some soricid shrews may carry plague.

Role of shrews in mythology

A few insectivores are the subjects of aboriginal myths. Some Eskimos believe that the red-toothed shrews (*Sorex*) will attack a man and kill him by burrowing through his body to the heart. The hero shrew (*Scutisorex congicus*), whose backbone of fused and arched vertebrae protects it from crushing, is venerated by central African natives, who believe that consumption of its heart will confer the animal's strength upon the eater.

Distribution. Insectivores are widely distributed but are not found in Antarctica, the Australian region, and South America (except for extreme northern regions). Various marsupials found in Australia are ecologically equivalent to certain insectivores, including a marsupial mole. Why insectivores are absent from South America has not yet been satisfactorily explained.

NATURAL HISTORY

Life cycle. *Reproduction.* Temperate zone insectivores, such as shrews, moles, and hedgehogs, have cyclic reproductive periods. Males develop enlarged testes in late winter or early spring that begin to produce sperm. At the same season females develop enlarged ovarian follicles in preparation for ovulation (release of eggs). Because female hedgehogs may undergo several periods of heat during which copulation does not result in pregnancy, it seems likely that ovulation occurs at a specific time, regardless of other reproductive activities. In shrews and moles, by contrast, every mating results in pregnancy; it has thus been supposed that ovulation in these animals is induced by copulation. An inverse relationship seems to exist between gametogenesis (the production of eggs or sperm) and the development of certain cells in the ovaries and testes. Female moles also exhibit a vaginal cycle, in which the vagina opens to the exterior only during the season of heat, pregnancy, and lactation. The majority of tropical insectivores appear to breed throughout the year; presumably, sperm production is continuous, and ovulation can occur at any appropriate time. Some evidence indicates that, in tropical areas with wet and dry seasons, the majority of births take place with the approach of the rainy season and the accompanying increase in the abundance of invertebrates. In West Malaysia, for example, 41 percent of the recorded pregnancies of one species of white-toothed, or musk, shrew (*Suncus*) occur from October through December, and the peak of the monsoons occurs in January; 19 percent of the pregnancies occur from April through June, before the height of the dry season in July. South African golden moles breed during rainy seasons. Some tenrecs become dormant during the Madagascan dry season, so it is likely that their reproductive period is also related to approaching rains.

Courtship and mating have seldom been observed among insectivores. Males and females of most species remain together for a brief period during which mating takes place. Among tenrecs the male may bite the female's neck; copulation then takes place, with the male mounting the female from behind in all known cases. Shrews may remain locked together for up to five minutes after the male dismounts, while the female drags the male after her. The ejaculate of the male short-tailed shrew (*Blarina brevicauda*) includes a waxy plug that blocks the vaginal orifice, preventing the loss of semen.

Gestation periods

The length of the gestation period varies from two or three weeks in shrews to 50 to 60 days in elephant shrews and tree shrews. In part, the length depends upon the condition of the young at birth; shrews and moles, for example, with a gestation of three or four weeks, have naked, blind, and helpless young. Hedgehogs, with a slightly longer period, have young that, although essentially helpless, do have partially developed spines. Tenrecs may have a 50-day gestation; the young have spines and rather quickly develop neuromuscular integration sufficient to enable them at about four days of age to walk and to exhibit at five days the stereotyped tenrec "bucking" defensive behaviour pattern (see below). Newly born elephant shrews have fur, open eyes and ears, and the ability to follow the parent. That a long gestation does not necessarily produce more developed young in insectivores, however, is shown by tree shrews, which bear helpless young after a relatively long gestation period. After giving birth, females of some species, such as tree shrews and some red-toothed shrews, may immediately mate again.

The number of young produced depends in part on species, latitude, and condition of young at birth. Elephant shrews, tree shrews, otter shrews (Potamogalidae), solenodons (Solenodontidae), and golden moles have one to three young. In the hedgehog family the spiny species have one to seven young; the nonspiny tropical Asian gymnures commonly have two. Moles average two to five young, depending on the species. Shrews have more young in temperate areas than in the tropics. Average numbers recorded range from two to three in Africa; two on the island of Guam; three, four, and five in various parts of India; and seven to eight in Great Britain. Various tenrecs may hold the record for number of young among placental mammals, *Centetes* having as many as 25; the presence of more than 30 embryos has been recorded.

Maturation and longevity. Embryonic development has been studied in only a few insectivores. In most shrews, moles, hedgehogs, and tenrecs, the eyes and ears open in about a week, and the babies are weaned in three or four weeks. Tropical white-toothed shrews reach sexual maturity in about five weeks, as do elephant shrews; tenrecs attain this condition in two months, a time that may also apply to other tropical insectivores. Temperate zone shrews, moles, and hedgehogs attain reproductive maturity in the second year of life. The characteristic chain-forming behaviour of young white-toothed shrews (in which the young form a chain behind the mother, each animal gripping with its teeth the rump fur of the one ahead) may begin after the ears open but before the eyelids part, suggesting that this behaviour is organized by sound stimuli.

The average life-span of most wild insectivores, like that of most small mammals, is probably quite short. It has been estimated at six weeks for shrews. Some individuals of populations of temperate zone shrews and moles, however, must live at least one year in order to participate in reproduction. These species seemingly are programmed for a year of healthy life, but few live beyond this; they succumb from wornout teeth. Captive shrews have survived two and one half years, elephant shrews three years, and hedgehogs 10 years.

Response to adverse conditions. As a normal part of their annual cycle, various insectivores have developed strategies for surviving intervals during which food is in short supply. In temperate regions this occurs in winter; in tropical areas dry seasons often are times of privation. Hedgehogs cope with the problem of food shortage by seeking a shelter, rolling themselves into a ball, and becoming dormant. Although the temperature of the extremities drops to that of the surrounding air, that of the heart remains at the level of an active animal; in this way energy is used slowly. In the dormant hedgehog, white blood cells concentrate in unusual numbers around the stomach and intestine, the sugar level in the blood drops by one half, and the level of magnesium in the blood rises to twice the normal value. In a normally dormant hedgehog, the Islets of Langerhans, which are tissues in the pancreas that secrete the hormone insulin, are active. Certain tenrecs are the only other insectivores definitely

Dormancy in hedgehogs

known to become dormant, a state they may enter during the dry season or in cool weather. The body temperature drops to that of the environment, and the respiratory rate may slacken to five or six breaths per minute. One kind of golden mole, *Chrysospalax,* has been reported to become dormant.

Another common stratagem used by many animals to survive lean times, the storage of food, is known for only a few insectivores. Some kinds of shrews store such invertebrates as snails. The European mole (*Talpa europaea*) may paralyze earthworms before burying them in groups in its mound. Energy may be stored as fat by some hedgehogs, tenrecs (*Microgale*), and star-nosed moles (*Condylura*). The last two genera use the tail as a storage organ.

A curious and unexplained aspect of the annual cycle of some temperate zone shrews involves the volume of the brain and the height of the braincase; both decrease to a minimum in midwinter, then increase again.

Behaviour. *Feeding behaviour.* All insectivores feed wholly or partly on animal matter and spend a substantial amount of time foraging. Shrews consume large quantities of invertebrates and occasionally small vertebrates. Moles eat soil invertebrates, including worms. Larger species eat larger prey; hedgehogs, for example, may consume frogs, mice, birds, lizards, and snakes and in captivity will eat bread and milk. The diet of golden moles is not well-known, although some species are known to prefer legless lizards (*Anguis*); others consume earthworms and insect pupae and larvae. Tree shrews are omnivorous, eating many insects as well as fruit; in captivity they thrive on commercial monkey food. Elephant shrews avidly eat locusts and, in captivity, thrive on bird meat. The aquatic tenrec *Limnogale* eats the tubers of the pondweed *Aponogeton,* and also takes fish. The water shrews consume aquatic animals, sometimes even small fish, and the otter shrew eats crabs, fish, and amphibians.

Many insectivores probe through leaf litter or soft soil. Solenodons proceed in this way until the long snout contacts prey; then the forepaws sweep forward and inward, gathering the object into the mouth. So stereotyped is this behaviour that even in captivity the animal captures inert food items in this way. Young solenodons follow the foraging parent and may learn the scent of acceptable food by smelling the mouth of the female. Young tenrecs may also learn to discriminate food by following their mother. Tenrecs, which often forage in groups, become excited at the scent of earthworms and begin to root enthusiastically. The sound of rooting and chewing stimulates other tenrecs in their searching activity. Shrews search through leaf litter or underground passageways of other mammals, locating prey by smell and contact. Moles do much of their hunting in their burrows, which may act as natural traps into which invertebrates fall. Golden moles emerge after rains and forage in the surface soil. Some shrews hold food items in their paws while eating and may carry food to preferred sites to eat it. Tenrecs may use the forefeet to hold down worms while they are torn in pieces.

A number of insectivores are reported to drink water, although elephant shrews probably do not do so. Tree shrews use water for drinking and bathing. Many insectivores may obtain sufficient water from their food or from dew.

Orientation and activity periods. Except for tree shrews and elephant shrews, both of which are primarily diurnal (*i.e.,* active by day), insectivores rely relatively little on vision for orientation. In most species, olfactory, tactile, and auditory cues are most important. Shrews, some tenrecs, and possibly the solenodons use a crude form of

Sound production

echolocation, a process by which objects are located by sound waves reflected back to the animal. Tenrecs produce clicks with the tongue at frequencies from five to seven kilohertz. Spiny tenrecs produce high frequency noises by vibrating together rows of overlapping spines on the middle of the back; the sounds produced in this way probably are used in social communication rather than orientation. Shrews and solenodons produce, probably with the larynx, clicks of 10 to 31 kilohertz (in solenodons) and 25 to 60 kilohertz (in shrews). Many insectivores, when exploring strange surroundings, maintain contact with a solid surface such as a wall or a log, which is probably tested with the facial sensory hairs (vibrissae). After a territory has become familiar, many insectivores are able to traverse it by memory and may become severely disoriented if the landscape changes.

Most insectivores are active at night, although shrews may be active throughout the 24-hour period. Most elephant shrews and tree shrews are diurnal, but at least one genus in each family is nocturnal. True moles and apparently also the little-studied golden moles are active by day or night.

Locomotion. In walking, most insectivores proceed by moving the right forefoot and left hindfoot, then left forefoot and right hindfoot forward together (crossed extension limb synchrony). Solenodons run with a quadrupedal richochetal (springing) gait, in which both forefeet and then both hindfeet strike the ground together. Other large insectivores also probably run in this way. Elephant shrews may hop bipedally, but their locomotion has not yet been studied. Tree shrews are skillful climbers but spend much of their time on the ground; some tenrecs are also partly arboreal. A few insectivores, such as otter shrews, several species of shrews and moles, and one tenrec, are partially aquatic and swim with alternate strokes of the hindfeet. The American water shrew (*Sorex palustris*) has been observed to scamper over the surface of still water, apparently being supported by the surface tension and the fringe of stiff hairs on its feet. The water tenrec (*Limnogale*) has webbed feet, as do the aquatic moles *Desmana* and *Galemys.* Some aquatic insectivores may use a laterally compressed tail in swimming, such as is found in the water shrew *Nectogale elegans,* the water moles *Desmana moschata* and *Galemys pyrenaicus,* and the otter shrew *Potamogale velox.*

Movements in water

Grooming. All insectivores spend some time in self-grooming. The universality of this behaviour suggests that it is necessary for keeping the hair in proper condition for its role in thermoregulation. Solenodons use only the hindfeet in grooming, perhaps because the forefeet are enlarged for digging, and the elongated snout hinders the use of the mouth in grooming. Shrews use the hindfeet and also the tongue. (Why the forefeet are not used by shrews is not clear.) Hedgehogs likewise do not use the front feet for cleaning. The tenrec *Echinops* uses both hindfeet and front feet. Tree shrews of the genus *Tupaia* use all four feet, but the forepaws and mouth are the dominant cleaning organs. These animals also comb the fur with the lower incisors.

Burrowing and nesting. Probably the most complex insectivore shelters are built by moles. A mole may have an underground nest chamber surrounded by concentric rings of tunnels interconnected by radiating ones. Shallow surface tunnels extend from this deeper complex and are evidenced by raised ridges, which mark their course. The mole may also throw up a large mound of earth on the surface that marks the location of its deep tunnel system. Burrows are dug by using the spadelike forefeet in a manner resembling the breaststroke of swimmers. The golden moles may use the same method, but little is known concerning their excavations. Some shrews dig tunnels, but many species use the runs of other animals. Shrews build surface nests, lined with shredded plant material, in which they rest or bear their young. Solenodons construct nest chambers during the breeding season, as do hedgehogs and tenrecs. The last two also build nests for keeping warm when the temperature drops. Elephant shrews often seek shelter among rocks and probably seldom excavate their own burrows; they may, however, occupy those abandoned by rodents. Tree shrews build nests of leaves and other vegetation, often among roots or fallen timber or in a tree cavity above the ground. When frightened, many insectivores immediately flee to a nearby burrow entrance.

Defensive behaviour. Spiny insectivores, such as hedgehogs and tenrecs, defend themselves by rolling into a ball with the spineless undersurface curled inward. Hedgehogs have a specialized muscle that is capable of pulling the spiny dorsal skin about the animal so that only a small unprotected area remains. Spiny tenrecs, when disturbed, exhibit a stereotyped defense behaviour in which the

spines on the head and neck are erected, and the animal "bucks" by throwing its forequarters upward. The action would drive spines into an attacker. Solenodons and tenrecs gape and hiss at intruders, a pattern also seen in the unrelated American opossum, a marsupial. Shrews, when confronted with strangers, open the mouth and emit bursts of high-pitched cries.

Shrews and solenodons have been shown to produce saliva with poisonous properties; it probably functions in quieting large prey animals. Shrew bites may be painful to humans.

Social behaviour. The majority of insectivores are rather antisocial, except at breeding time and in the context of the female–young relationship. Adult shrews and moles tolerate other adults of their species briefly during the period of mating, following which the female and young form a group until the young are full grown; the young rarely appear alone during this interval. Young white-toothed shrews (*Crocidura*) exhibit chain formation, in which each animal grasps the rump fur of another, one grasping the mother, after whom they then trail. Baby shrews are rather indiscriminate and may attempt chain formation with such inappropriate subjects as white mice.

Juvenile behaviour

Solenodons have a prolonged period of juvenile dependency during which the young follow the female, perhaps learning to locate favoured foraging areas and to discriminate in food selection. Adult solenodons approach each other with open mouths, perhaps emitting high-frequency clicks. Contact involves one animal closing its mouth over the snout of the other, then pressing the snout against its own flanks and back. Several solenodons may occupy a burrow. Various tenrecs forage in large groups which probably consist of one or more females and young. Elephant and tree shrews live solitarily or in groups of two or three, consisting of a female and her young. Hedgehogs may also form temporary groups of female and young.

Hedgehogs and spiny tenrecs perform an unexplained "self-anointing" behaviour, which is triggered by the presence of such substances as urine or certain chemicals. The animal licks the material until a frothy spittle is produced; in some species this is then spread over the body by the tongue in typical hedgehogs (*Erinaceus*) or by the mouth and forefeet in tenrecs.

Ecology. *Predation upon insectivores.* Shrews are preyed upon by a variety of birds, mammals, and snakes, but seemingly the strong odour produced by the flank gland of shrews makes them unpalatable to some predators. Skulls of shrews are commonly found in owl pellets (regurgitated indigestible remains of prey). Little is known about the predators of elephant and tree shrews. True moles and golden moles may be safe from many predators because of their subterranean habits; spiny hedgehogs and tenrecs may, for the same reason, be relatively safe. In Madagascar some tenrecs are preyed upon by man. Solenodons seemingly suffered little predation on Cuba and Hispaniola until man introduced dogs, cats, and mongooses there. These carnivores have drastically reduced solenodon populations, probably because the solenodons had evolved no defenses against such predation.

Territoriality. Each adult shrew occupies its own territory of about 0.1 to 0.4 hectare (0.25 to one acre). About 40 percent of its behaviour involves avoiding other shrews. When shrews from adjacent territories encounter one another, they engage in mutual avoidance if possible. Avoidance is made easier by an oily odoriferous substance exuded by the flank glands, which helps to warn off other shews of the same species. A similar kind of territoriality probably is practiced by moles. Population structure of most other insectivores has not yet been studied. Tree shrews apparently defend nest territories, and males may mark these areas by smearing onto branches the yellow exudate from throat and chest glands.

Habitats. Insectivores utilize diverse habitats. Shrews occupy the ground surface in forests and grasslands but are less common in deserts, although *Notiosorex* and *Diplomesodon* occupy New and Old World deserts, respectively. Moles are largely confined to North Temperate Zone forests and meadows, especially in areas of deciduous forests and adjacent prairies. Hedgehogs live in forests, grasslands, and deserts of the Old World. Spineless hedgehogs (Echinosoricinae) live in tropical rain forests of Asia. Golden moles occupy many habitats in southern Africa. Otter shrews live in riparian situations (*i.e.,* on the banks of natural water courses), in the Congo Basin and environs. Elephant shrews prefer African savanna areas, but some are forest dwellers. The tenrecs live in most habitats on Madagascar; some are partly arboreal. Tree shrews are largely terrestrial in tropical Asian forests, but most can climb well and are partly arboreal. Solenodons are strictly terrestrial foragers in the forests of Cuba and Hispaniola.

FORM AND FUNCTION

The order Insectivora is the most difficult of living mammalian orders to define because the groups included retain traits of primitive placental mammals and are not very specialized. Although each family has specializations, they are not such as to suggest that each group should be raised to ordinal level. Furthermore, many of the families are as old as many mammalian orders, and the evidences of common relationship are often obscured.

Skull. The skull of insectivores is typically low in that the braincase does not rise abruptly from the level of the rostrum (the bony support for the snout), which tends to be long and tapered. Seen from above, the skull is triangular in outline. The cerebral cortex of the brain is not well developed, and the olfactory bulbs are large, in conjunction with the well-developed sense of smell. In addition, the elongated rostrum supports muscles that activate a still longer, highly mobile snout with well-developed tactile capabilities that are useful in probing for food. The middle ear cavity is partly or wholly enclosed by the tympanic bone. In such forms as shrews and hedgehogs, the tympanic bone (surrounding the middle ear) is ringlike and attached to the base of the skull (basicranium) only at one point, probably the primitive situation. Other forms, such as moles and golden moles, have an expanded tympanic bone fused to the basicranium to form an auditory bulla (an enclosed chamber). The bulla of tree shrews and elephant shrews includes an entotympanic bone, a trait also seen in primates. Many insectivores have small external ear pinnae (the projecting "shells" of the ear); this is nearly or quite lacking in burrowing forms such as moles but may be rather prominent in such forms as desert hedgehogs. In shrews and tenrecs, at least, sensitivity to high-frequency sounds is well developed.

Presence of tympanic bone

Orbital fossae (eye sockets) and eyes are variably developed in insectivores, most species having relatively small eyes. In moles and golden moles, in which the eyelids are fused shut, probably only differences in light intensity can be detected. Large eyes and well-developed vision are present only in tree and elephant shrews. In the former the orbital fossa is completely encircled with bone.

Dentition. Insectivores have from 26 to 44 teeth, the latter being the maximum for placental mammals. Incisors range from three to six pairs. The second lower incisors of solenodons are long and grooved and transmit venom from the submaxillary salivary glands, the ducts of which end at the bases of these teeth. The upper and lower first incisors of shrews are provided with one or more accessory cusps. These specialized teeth close so as to provide the animal with a delicate tool for picking up small invertebrates. The incisors of tenrecs may also be provided with accessory cusps, but tenrecs lack the other dental specializations of shrews. The lower incisors of tree shrews form a comblike structure and are used for grooming the fur. The canine teeth, though present in insectivores, are usually small, rarely extending much above the level of the other teeth; they are not strong and elongated as those in carnivores (an exception is the Palearctic mole *Talpa*). The canine teeth of hedgehogs and moles have double roots, an unusual condition among mammals. Premolars range from the simple one-cusped structures seen in shrews and some moles to multicusped molariform teeth. The molariform premolar is usually the fourth or, more rarely, the third.

Molars

Insectivoran molars are of a primitive placental type. The crown supports several usually conical cusps arranged in a triangular pattern (called protocone, paracone, and

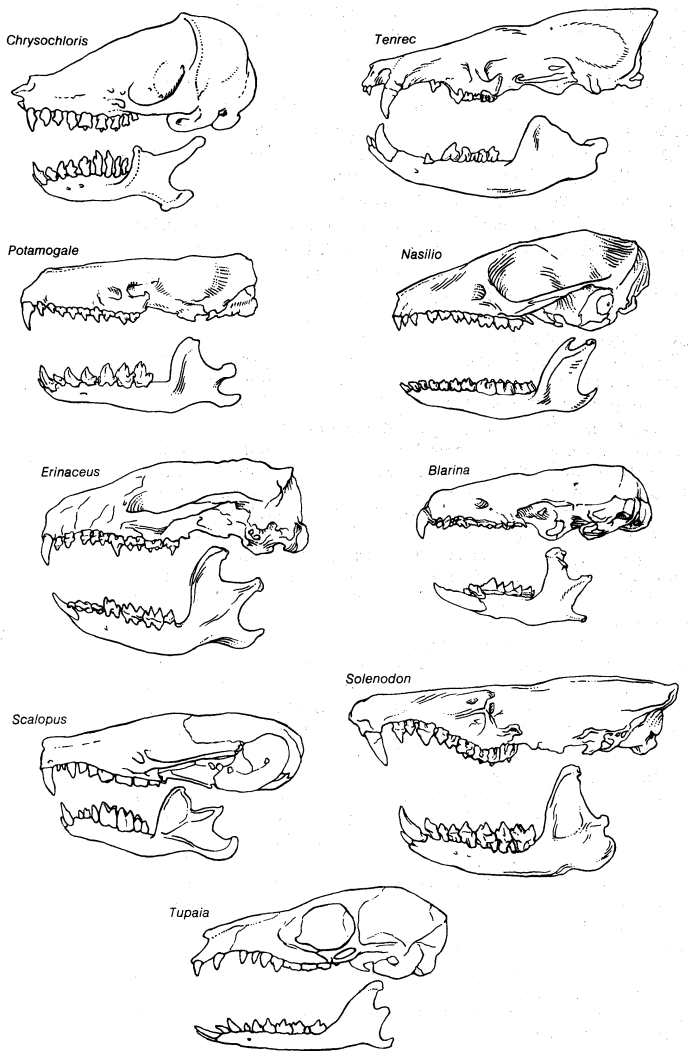


Figure 15: Skulls (cranium and lower mandible) of representative insectivores.

Drawing by R. Keane based on (*Chrysochloris, Tenrec, Potamogale, Nasilio*) P.-P. Grasse, *Traite de Zoologie*, vol. 17 (©1955); Masson & Cie, Paris; (*Blarina, Scalopus*) E.R. Hall and K.P. Nelson, *The Mammals of North America*, copyright © 1959, The Ronald Press Company, New York; (*Tupaia*) *Proceedings of the U.S. National Museum*, vol. 45

metacone on the upper teeth; protoconid, paraconid, and metaconid on the lower). The upper triangle is often squared off by the addition of a posteromedial hypocone (*i.e.*, on the rear inner corner). To the lower triangles is added a posterior shelf (talonid) frequently made up of cusps called hypoconid and entoconid. These molars close so as to allow some degree of both slicing and crushing action; the term tuberculosectorial is applied to them. Omnivorous forms tend to emphasize the squaring up of the teeth and the crushing, rather than the slicing, function by lowering the cusps so that the closing surface is made up of low hillocks. This type, the bunodont molar, is best developed in hedgehogs. When the triangular part of each crown is emphasized and elevated, the result generally indicates a carnivorous diet; such teeth are seen in tenrecs, otter shrews, golden moles, and solenodons. In these four groups the paracone and metacone (outer cusps, forming the base of the upper triangle) are close together and medially removed from the outer edge of the tooth. Such a molar type is called zalambdodont. Moles and shrews have converted the paracone and metacone into V-shaped ridges, such that the two produce a W-shaped pattern on the upper molar (termed dilambdodont). The result is that there are four, rather than two, cutting edges on the upper molar. This specialization, seen also in insectivorous bats, is seemingly especially useful in chopping up small invertebrates. The molars of tree shrews and elephant shrews are also dilambdodont; in the former the appearance is sufficiently similar to that of true shrews to suggest some dependence on insects. The teeth of elephant shrews show a curious parallel to those of artiodactyls in being somewhat high crowned, with considerable molarification of the premolars. The extent to which elephant shrews are vegetarians, as their dentition suggests they may be, has not been measured. Shrews may replace their milk teeth while still in the uterus.

Limbs. The appendicular skeleton of insectivores retains many traits seen in primitive placental mammals. All except the otter shrews retain a clavicle connecting scapula (dorsal part of pectoral girdle) to sternum (breastbone). Since absence or reduction of the clavicle usually characterizes mammals that are specialized for rapid leaping locomotion, such as artiodactyls (antelopes, pigs, and cattle) and perissodactyls (horses, rhinoceroses, and tapirs), its absence in otter shrews, which are reported to be clumsy on land, is perplexing. The humerus (upper bone of the forelimb) of insectivores retains an entepicondylar foramen (an opening in the articular surface), absent in many more specialized mammals. The radius and ulna (bones of the forearm) are separate and well developed. The manus (forefoot) is rarely specialized, usually possessing five digits and a relatively full complement of wristbones (carpals). In some forms (shrews and hedgehogs) one wristbone, the centrale, is missing or fused with an adjacent bone, as it is in most primates. Moles have broad forefeet, in which supernumerary bones are closely knit to form a rigid plane, as in the blade of a shovel. Golden moles have the hand digits reduced to four, two of which are greatly enlarged and bear huge, picklike claws. The pelvic girdle shows no peculiarities. In a few families (shrews, moles, golden moles, and otter shrews) the pubic bones do not form a symphysis (union); in others a short or long symphysis is present. The femur (thighbone) is marked by a laterally placed third trochanter (a ridge) for muscular attachment, presumably a primitive trait but one seen in a variety of other mammals. The tibia and fibula (lower bones of the hindlimb) are usually fused distally (near the ankle), but in solenodons this fusion takes place only with old age and in some tenrecs and all tree shrews not at all. Fusion reduces the lateral mobility of the foot but promotes more efficient forward and backward motion. In hedgehogs and elephant shrews the toes may be reduced to four. All insectivores except elephant shrews have plantigrade foot posture; *i.e.*, the entire sole is placed down in walking. In some species, however, the heel is frequently lifted from the ground.

Fusion of tibia and fibula

Tail. The tail structure varies according to the behaviour of the animal. Burrowing insectivores have little or no tail, in common with most other burrowing animals. Climbing and running species have well-developed tails, haired or nearly naked. The slow-moving hedgehogs generally have short tails. In aquatic forms, such as aquatic shrews, moles, and potamogales, lateral compression of the tail seemingly suits it for steering functions.

Fur texture and coloration. Specializations of the hair are seen in hedgehogs and some tenrecs, which possess

From (A,B,C,D) A.S. Romer, *Vertebrate Paleontology*, copyright 1966, University of Chicago; (E) *Bulletin of the American Museum of Natural History* (1942), courtesy of the American Museum of Natural History

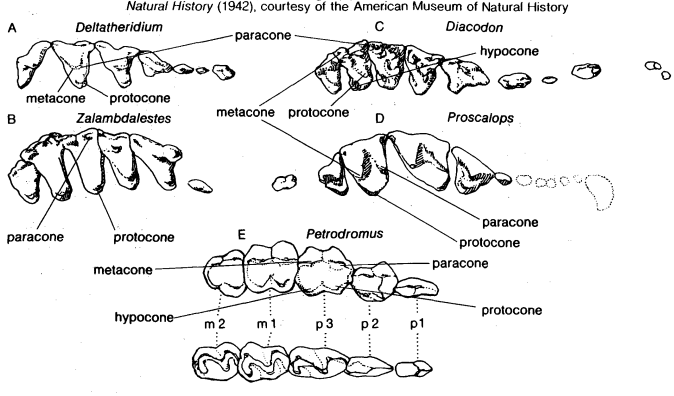


Figure 16: Major molar types in insectivores.
(A) An early zalambdodont placental. (B) A Cretaceous leptictid. (C) An early hedgehog with omnivorous bunodont tooth type. (D) A dilambdodont mole. (E) A putatively vegetarian elephant shrew (see text).

spiny hair. Burrowing species have reversible fur, which can lie down either anteriorly or posteriorly as the animal moves forward or backward in the tunnel. The fur of golden moles has a distinctive metallic sheen or iridescence. Few insectivores are strikingly coloured; most are brown, gray, or black with little countershading. Exceptions occur in diurnal species, which may be more brightly coloured and have some indication of patterns.

Digestive and reproductive specializations. Tree shrews and elephant shrews possess a cecum (a blind pouch) at the junction of small and large intestines. An intestinal cecum is not present in carnivorous insectivores, since it serves for storage and continued digestion of plant food. Elephant shrews, whose teeth suggest a diet consisting largely of plants, have a large cecum; tree shrews, which probably eat relatively less plant food, have a comparatively small cecum.

A true scrotum containing the testes is present only in tree shrews; in some other forms, such as moles and tenrecs, the testes lie close to the surface in the perineal region (the area ventral to the anus) but do not leave the abdominal cavity. A penis bone (baculum) is found in tenrecs and moles. Elephant shrews have a trifurcate penis, a trait that led some early students to ally them with the marsupials, some of which have a bifurcate structure. The form of the uterus is Y-shaped, or bicornuate, as in many other mammals.

Metabolism. Although insectivoran physiology has been studied in connection with special abilities, such as dormancy in hedgehogs, it is not well-known for the order.

Metabolism in shrews

Of great interest are the smaller shrews, which, because of their minute size and relatively large ratio of body surface to volume, have an exceptionally high rate of metabolism, necessitated by the high rate of energy loss in the form of heat. As a result shrew life consists largely of a frenetic search for food. A captive female *Sorex* ate 3.3 times her own weight every 24 hours.

EVOLUTION AND CLASSIFICATION

Evolution. In late Cretaceous and early Cenozoic times (beginning about 100,000,000 years ago) a number of primitive placental mammals existed that did not display the specialized characteristics of modern orders; they possessed tritubercular (three-cusped) cheek teeth, with the paracone and metacone well separated and located laterally on the occlusal (closing) surface. These animals, the Leptictidae, some of which persisted until Oligocene times (about 38,000,000 years ago), were close structurally to Eocene hedgehogs and may well be ancestral not only to many modern insectivores, especially hedgehogs, shrews, and moles, but also to a variety of other modern mammals, such as ungulates and primates. These insectivores, in which the paracone and metacone are well separated and are near the lateral edge of the molar, are thought by some paleontologists to have descended from Late Cretaceous leptictids and to form a separate lineage from the zalambdodonts. The fossil record is fairly clear in showing common ancestry of shrews and moles, and probable derivation from late Eocene or early Oligocene hedgehog-like ancestors (perhaps 40,000,000 years ago). Some paleontologists believe that the presence of zalambdodont molars in tenrecs, otter shrews, and golden moles is indicative of independent descent from the zalambdodont deltatheridiids, a family of doubtful ordinal affinities, of Late Cretaceous time. It has recently been suggested that solenodons also descended from a hedgehog-like ancestor, which reached the Greater Antilles in early Tertiary time. Tree shrews are known from Paleocene time (about 65,000,000 years ago) and probably come closest among living mammals to representing the appearance of the early Cenozoic primitive placentals. Fossil elephant shrews from the African Miocene (about 26,000,000 years ago) offer few clues to the origin of the Macroscelididae, which were in all likelihood derived from the unspecialized Early Cenozoic placental group. Tenrecoids and golden moles have an exclusively African and Malagasy fossil history, also stemming from the Miocene.

Classification. *Distinguishing taxonomic features.* Insectivores are characterized by the lack of distinctive features seen in other mammalian orders rather than by the possession of unifying morphological traits. All possess a low braincase, a rather long conical rostrum (snout) and unspecialized legs, feet, and locomotion. They can be distinguished from similarly built rodents by the lack of the gnawing incisors (followed by a diastema, or space) that characterize the rodents and from similarly built marsupials by the lack of an abdominal pouch.

Annotated classification. Taxonomists disagree on the relative degrees of relationship among insectivoran families (except that shrews, moles, and hedgehogs are conceded to form a cohesive phyletic unit); the classification omits subordinal and superfamilial groupings.

ORDER INSECTIVORA

Primitive placental mammals characterized by a low braincase, conical snout, and lacking locomotor or dental specializations of other mammalian orders; 77 genera, 406 species.

Family Erinaceidae (hedgehogs and gymnures)

Eocene to Recent; Eurasia and Africa. Spiny pelage (covering) present (subfamily Erinaceinae, hedgehogs) or lacking (Echinosoricinae, gymnures); zygomatic arch (below eye orbit) present in skull; molars quadrate, bunodont (see above *Dentition*); auditory bulla a tympanic ring, not fused to cranium. Terrestrial, omnivorous, often hibernating or estivating. Hedgehogs are Eurasian and African; gymnures native to the Asian tropics, often shrewlike. Ten genera, 14 species.

Family Talpidae (moles)

Eocene to Recent; Holarctic. Manus very broad and specialized for digging; molars with W-shaped crests; zygomatic arch present; auditory bulla fused to cranium; eyes minute; ears without pinnae; humerus bone extremely broad and short. Burrowing (fossorial), sometimes aquatic. Fur of some commercial value. Fifteen genera, about 22 species.

Family Soricidae (shrews)

Eocene to Recent; Africa, Eurasia, North America, northern edge of South America. First upper and lower incisors procumbent, multicusped; zygomatic arch lacking; bulla a tympanic ring, unfused to skull; molars with W-shaped crests; small size. Terrestrial, aquatic, or subfossorial foragers in ground litter or water for small invertebrates. Twenty-four genera, approximately 291 species.

Family Solenodontidae (solenodon, almiqui)

Pleistocene to Recent; Cuba and Haiti. Molars zalambdodont; zygomatic arch incomplete; incisors canine-like, lower ones grooved; long scaled tail; long mobile snout. Terrestrial, perhaps nearing extinction. Two genera, 2 species.

Family Tenrecidae (tenrecs)

Miocene to Recent; Madagascar. Molars zalambdodont; zygomatic arch incomplete; bulla composed of a tympanic ring and basicranial elements; form variable from nearly tailless, spiny, rabbit-sized *Tenrec* to long-tailed, soft-furred, shrewlike *Microgale. Oryzorictes* is molelike, while *Limnogale* is muskrat-like and a vegetarian. Nine genera, 20 species.

Family Potamogalidae (otter shrews)

Recent; West Africa. Webbed feet, otter-like form; zalambdodont molars. Riparian, aquatic, feeding on aquatic vertebrates and invertebrates. Two genera, 3 species, often included as a subfamily of Tenrecidae.

Family Chrysochloridae (golden moles)

Miocene to Recent; Africa. Two digits on front feet with enormously enlarged claws; proximal segments of arm recessed into sides of thorax; heart and lungs elongated; zalambdodont molars; complete zygomatic arch composed of maxillary rather than jugal bone. Eyes covered with skin; tail rudimentary; pelage with metallic sheen or iridescence. Fossorial. Five genera, 11 species.

Family Macroscelididae (elephant shrews)

Miocene to Recent, Africa. Cecum well-developed; feet greatly elongated; molars quadrate and high-crowned; eyes large; auditory bullae complete, including entotympanic bone; zygomatic arch complete; resembling long-snouted kangaroo rats or jerboas. Diurnal, terrestrial, saltatory, animalivorous and phytophagous. Five genera, about 28 species.

Family Tupaiidae (tree shrews)

Paleocene to Recent, tropical southeast Asian mainland, Sumatra, Borneo, Philippines. Small cecum; orbital fossa ringed with bone; zygomatic arch complete; molars quadrate with W-shaped crests; eyes large; audital bulla includes entotympanic bone; generally like a long-nosed squirrel in appearance and behaviour. Terrestrial and arboreal; diurnal or nocturnal. Five genera, 15 species.

In addition to the above, 14 families of primitive extinct mammals are sometimes assigned to the Insectivora. These families, with their geologic time spans, are listed below.

Family Dimylidae. Oligocene–Paleocene.
Family Picrodontidae. Paleocene.
Family Zalambdalestidae. Cretaceous.
Family Leptictidae. Cretaceous–Oligocene.
Family Adapisoricidae. Paleocene–Miocene.
Family Pantolestidae. Paleocene–Oligocene.
Family Apternodontidae. Eocene–Oligocene.
Family Apatemyidae. Paleocene–Oligocene.
Family Mixodectidae. Paleocene.
Family Anagalidae. Eocene–Oligocene.
Family Paroxyclaenidae. Eocene–Oligocene.
Family Ptolemaiidae. Oligocene.
Family Pentacodontidae. Paleocene–Eocene.
Family Plesiosoricidae. Eocene–Pliocene.

Critical appraisal. It is possible, though current opinion does not support it, that insectivores with zalambdodont teeth (tenrecs, solenodons, golden moles) form a monophyletic unit descended from either pre-tritubercular Mesozoic (Jurassic, about 190,000,000 years ago) pantotheres or the zalambdodont deltatheridiids of late Cretaceous and Paleocene times. More likely, zalambdodont teeth were derived from the tritubercular type seen in leptictids. Zalambdodont and dilambdodont teeth represent different ways of increasing the length of shearing edge of a tritubercular tooth. Tupaiids and macroscelidids share a number of traits that have led some authorities to separate them as a group called Menotyphla, while grouping the other insectivores as the Lipotyphla. Some authorities have considered the colugos (Cynocephalidae) to belong in the Insectivora, but most now place them in a separate order, Dermoptera. Tupaiids share some traits with primitive primates and have been included in that order by some prominent authorities (see, for instance, the section *Primates*). Elephant shrews have many of these same traits, however, and would have to be included in the Primates as well. Some mammalogists treat the elephant shrews as a distinct order, Macroscelidea (see section *Mammalia*). A growing body of evidence suggests that soricids, talpids, erinaceids, solenodons, tenrecs, otter shrews, and possibly chrysochlorids form a naturally related group to which the ordinal or subordinal term Lipotyphla should apply. This group, as well as elephant shrews, tree shrews, primates, bats, and others, apparently have been separate from one another since earliest Tertiary times, and consistency may in the end dictate assignment of a separate ordinal name to each. (J.Fi.)

The Lipotyphla

Chiroptera (bats)

Bats, which comprise the mammalian order Chiroptera, are the only mammals to have evolved true flight. This ability, coupled with the benefits deriving from their system of acoustic orientation (so-called bat sonar), has made the group a successful one in numbers of species and individuals. About 900 species are currently recognized, belonging to some 174 genera. Many species are enormously abundant. Observers have concluded, for example, that some 100,000,000 female Mexican free-tailed bats (*Tadarida brasiliensis mexicana*) form summer nursery colonies in Texas, where they produce about 100,000,000 young in five large caves. Adult males of this species, although equal to the females in numbers, may not range as far north as Texas. (Individuals of the species also range widely throughout tropical America.) Thus, bats of one species alone number at the very least in the hundreds of millions of individuals.

GENERAL FEATURES

Most bats are insectivorous. Little is known of the spectrum of insect species consumed, but the quantities are formidable. The Mexican free-tailed bats of Texas have been estimated to consume about 20,000 tons of insects per year. Bats would thus seem to be important in the balance of insect populations and possibly in the control of insect pests. Some bats feed on fruit and aid in dispersing seeds; others feed on pollen and nectar and are the principal or exclusive pollinators of a number of tropical and subtropical plants. The true vampires of tropical America feed on the blood of large birds and mammals, occasionally becoming significant pests of livestock and sometimes serving as carriers of rabies.

Certain aspects of the physiology of some bats, particularly those involving adaptations for long hibernation, daily lethargy, complex temperature regulation, acoustical orientation, and long-distance migrations, are of interest to experimental scientists.

In tropical countries, in particular, large colonies of bats often inhabit houses and public buildings, attracting attention by their noisiness, guano (droppings), and collective odour. In the West, bats have been the subject of unfavourable myths; in parts of the Orient, however, these animals serve as symbols of good luck, long life, and happiness.

Noxious bats

Diversity of structure. All bats have a generally similar appearance in flight, dominated by the expanse of the wings, but they vary considerably in size. The order is usually divided into two well-defined suborders: the Megachiroptera (Old World) and the Microchiroptera (worldwide). Among members of the Megachiroptera, a flying fox, *Pteropus vampyrus,* may have a wingspread of about 1.5 metres (about five feet) and a weight of about one kilogram (2.2 pounds). The largest insectivorous bat is probably *Cheiromeles torquata;* it weighs about 250 grams (about nine ounces). The largest of the carnivorous bats (and the largest bat in the New World) is *Vampyrum spectrum,* with a wingspread of over 60 centimetres (24 inches). The tiny Philippine bamboo bat, *Tylonycteris pachypus meyeri,* has a wingspread of barely 15 centimetres (six inches) and weighs about 1.5 grams (about 0.05 ounce).

Bats vary in colour and in fur texture. Facial appearance, dominated by the muzzle and ears, varies strikingly with family and often with genus. In several families, a

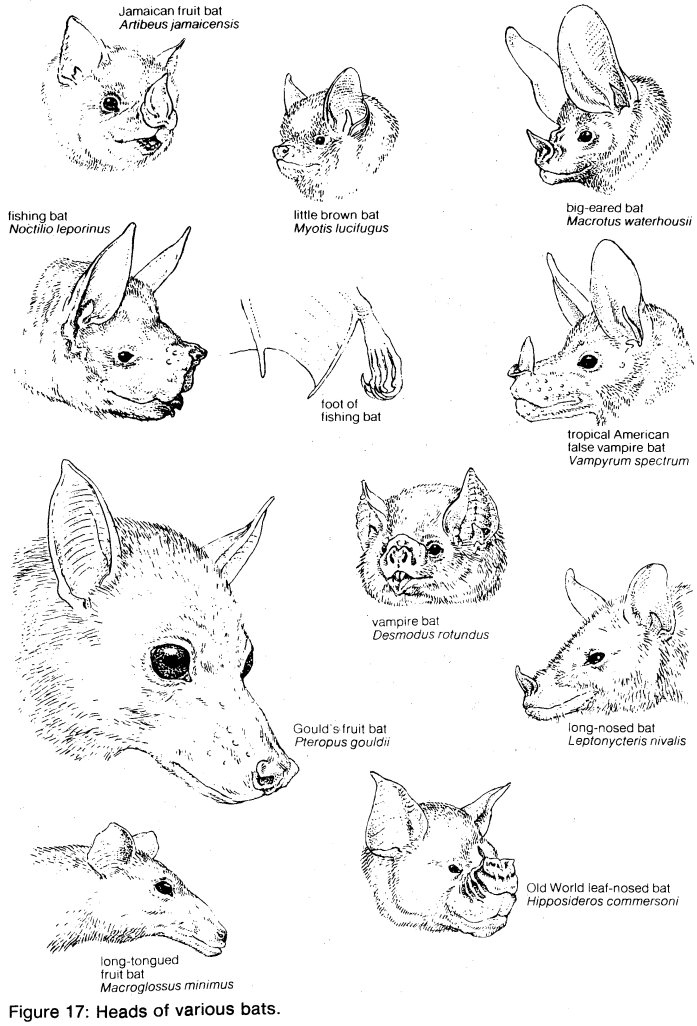


Figure 17: Heads of various bats.

complex fleshy adornment called the nose leaf surrounds the nostrils. Wing proportions are modified according to modes of flight. The tail and the interfemoral (between the legs) membrane also differ, perhaps with feeding, flight, and roosting habits. Finally, at the roost, bats vary in the postures they assume, particularly in whether they hang suspended or rest on the wall and in the manner in which the wings are folded and used.

Distribution. Bats are particularly abundant in the tropics. In West Africa, for example, 31 genera embracing 97 species have been catalogued; in the United States 15 genera, totalling 40 species, are known. Of the 17 living families, three (the Vespertilionidae, Molossidae, and Rhinolophidae) are well represented in the temperate zones. A few phyllostomatid species range into mild temperate regions. Several vespertilionids range well into Canada.

The Vespertilionidae are found worldwide, excepting only the Arctic and sub-Arctic, Antarctic, and isolated islands. The genus *Myotis* has a range almost equal to that of the order. The Molossidae and Emballonuridae also encircle the earth but are restricted to the tropics and subtropics. The Rhinolophidae extend throughout the Old World, the Hipposideridae and Pteropodidae throughout the Old World tropics, and the Phyllostomatidae throughout the New World tropics and slightly beyond. The other families have more restricted ranges.

IMPORTANCE TO MAN

Bats are important to man primarily because they affect other living things through predation, pollination, and seed dispersal. Although losses of commercially valuable fruit are normally small, bananas and figs must, in some cases, be protected by early harvest or by nets. Vampire bats are considered serious pests of livestock in some parts of tropical America, because the small wounds they cause provide egg-laying sites for botflies. Vampires may also transmit a disease of cattle—trypanosomiasis. As is true of mammals in general, bats seem susceptible to rabies, but most are not considered serious carriers of the disease.

Bat guano as fertilizer

The guano deposits of insectivorous bats are used for agricultural fertilizer in many countries and, in the past, were used as a source of nitrogen and phosphours for munitions. Large guano deposits, in addition, cover and thus preserve many archaeologically interesting artifacts and fossils in caves.

In some parts of Southeast Asia and on some Pacific islands, flying foxes (*Pteropus*) are hunted for food. Small bats are also widely but irregularly eaten.

In species and numbers, bats comprise an important and, on the whole, nonintrusive part of wildlife. Several zoos have established interesting exhibits of bats; indeed, some flying foxes and fruit bats have been exhibited in European zoos since the mid-19th century. Bats are interesting pets and have been kept widely for research purposes but require somewhat specialized care.

NATURAL HISTORY

Life cycle. Details of the life cycle are known for only a few species from North America or Europe. In these, there is an annual cycle of sexual activity, with birth taking place between May and July. In males, the testes, normally located in the abdominal region, descend seasonally into the scrotum, and active spermatogenesis occurs. In females, sexual receptivity may be associated with egg maturation and release. Tropical bats may exhibit a single annual sexual cycle or may be diestrous (*i.e.*, with two periods of fertility) or polyestrous (with many).

The sexual cycles of entire populations are closely synchronized, so that almost all mating activity occurs within a few weeks. The periods of development (gestation), birth, lactation, and fledging are also usually synchronized. Gestation varies in duration: five or six months in *Pteropus*, over five months in *Desmodus*, three months in some small *Hipposideros*, and six or seven to 14 weeks in several small vespertilionid genera. The length of gestation may be influenced by ambient (surrounding) and body temperature.

In several North American and Palaearctic (northern Eurasian) vespertilionids and rhinolophids that hibernate, copulation occurs in the fall, and the sperm are stored in the female genital tract until spring. Ovulation, fertilization, and implantation occur after emergence from hibernation, when the female again has available an abundant food supply and warm roost. The favourable environmental conditions greatly enhance the young bat's chances of survival.

Most bats bear one young, but the big brown bat (*Eptesicus fuscus*) may bear twins, and the red bat (*Lasiurus borealis*) bears litters of one to four.

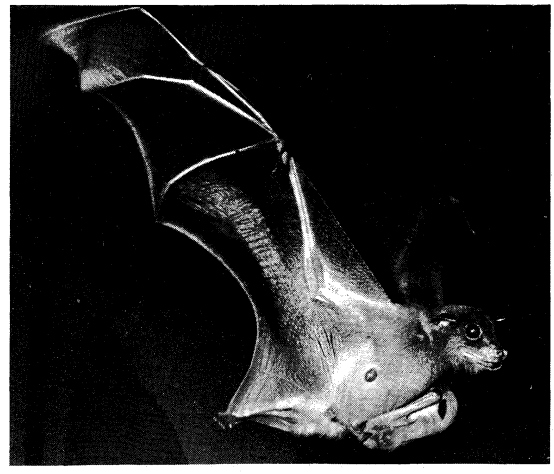
At birth, the young, who may weigh from one-sixth to one-third as much as the mother, usually have well-developed hindlegs with which they hold on to their mother or to the roost. The wings are very immature. The young, nude or lightly furred, are often briefly blind and deaf. Female bats normally have one pectoral (at the chest) or axillary (at the armpit) mammary gland on each side. Several families that carry their young while foraging also have a pair of false pubic nipples, which the infant may hold in its mouth when its mother flies. The infants are nourished by milk for a period of about five or six weeks in many small Microchiroptera and for five months in *Pteropus giganteus*. By two months of age, most of the Microchiroptera have achieved adult size, having begun to fly and forage three or four weeks earlier.

Nursery roosts

In many species, females late in pregnancy migrate to special nursery roosts, in which large numbers of pregnant females may aggregate, usually to the exclusion of nonpregnant females, males, and bats of other species. In some cases, the nursery roosts seem to be chosen for their high temperature, which may derive from the sun, the bats themselves, or from decomposing guano. When foraging, some bats (*Erophylla*) leave their infants hanging quietly, one by one, on the cave wall or ceiling. In other cases (*Tadarida brasiliensis, Natalus*), the closely spaced infants may move about and mingle on the wall. Among others, especially molossids in buildings, the infants may form play groups and rough-and-tumble on the floor. Some bats carry their young with them. Generally, each mother, on returning to her roost, seeks out her own child by position, smell, and acoustical exchange. In the huge nursery colonies of *Tadarida brasiliensis mexicana* and *Miniopterus schreibersii*, however, the mothers appear to feed the first one or two infants they encounter on returning. This is called the milk-herd system.

Some bats achieve sexual maturity in the first year; others in the second year. Infant mortality appears to be high. Developmental and genetic errors and disease take their toll, but accidents seem to cause more serious losses—the young may fall from the ceiling or have serious collisions in early flight attempts. A fair number of bats probably fail to make the transition from dependent infants to self-sufficient foragers.

T.S. Lal—P.I.P.



Mother dog-faced fruit bat (*Cynopterus sphinx*), with clinging young, in flight.

Life-span

Adult bats have low mortality. Predation is rarely serious, especially for cave-dwelling species. Disease, parasitic infestation, starvation, and accidents apparently take small

tolls. There are records of several big brown (*Eptesicus fuscus*), little brown (*Myotis lucifugus*), and greater horseshoe bats (*Rhinolophus ferrum-equinum*) that have lived more than 20 years. Probably many bats in temperate climates live more than 10 years. Longevity has not been established for tropical species.

Several factors probably contribute to the unusual longevity of bats. Generally isolated roosts and nocturnal flight substantially protect them from predation, from some elements of weather, and from exposure to the sun. The largely colonial way of life may ensure that entire populations experience contagious infection and subsequent immunity; indeed, such a pattern may, in the past, have hastened adaptation to disease. The persistent use of various seasonal roosts probably ensures isolation and security, food and water supply, and access to mates. Many bats, moreover, drop their body temperature at rest. Not only is there a probability that this conserves some cellular "machinery," since metabolism is reduced, but fewer hours need be spent in actively seeking food and water.

Behaviour. *Activity patterns.* Nocturnal activity is a major feature of the behavioural pattern of bats: nearly all species roost during the day and forage at night. Carnivorous bats, vampires, and perhaps fishing bats may have an advantage at night over inactive or sleeping prey. In addition, nocturnal flight protects bats from visual predators, exposure to the sun, high ambient temperature, and low relative humidity. The large area of naked wing skin might mean that bats would absorb rather than radiate heat, if they were active during the day. They would also lose body water required for temperature regulation and would then be forced to forage near water or to carry extra water in flight.

The nocturnal-activity pattern in bats is probably kept in synchrony with changing day lengths by their exposure to light at dusk or dawn. Bats often awaken and fly from the cave exit well before nightfall. Should they be too early, their internal clock may be reset. A few species of bats, *Lavia frons* and *Saccopteryx bilineata,* may forage actively during the day, but little is yet known of their special adaptations.

Locomotion. Flight is the primary mode of locomotion in all bats, although the flight styles vary. Some groups (the Molossidae, for example), adapted for flight in open spaces and often at high altitudes, have long, narrow wings, swift flight, and a large radius of turning. Other bats (the Nycteridae, Megadermatidae, and the Glossophaginae), adapted for hovering as they pick prey off vegetation or feed on flowers, have short, broad wings, slow flight, and a small radius of turning. Some bats take flight easily from the ground: members of the genus *Macrotus* do so simply by flapping, vampires (*Desmodus*) leap into the air and then spread their wings and fly. The molossids, however, roost well above the ground since, on takeoff, they fall before becoming airborne.

Speed in flight

Though flight speeds in nature are hard to measure, four vespertilionid species, carefully observed, have been timed on the average at 18.7 to 33.3 kilometres (11.7 to 20.8 miles) per hour. In flight, the posture of each of the four fingers incorporated into the wing is under precise and individual control. Finger and arm postures, which determine the shape, extension, and angle of the wings, govern such actions as turning, diving, landing, and hovering. Except when interrupted by insect catches or obstacles, bat flight paths are straight. Insects may be pursued and captured at a rate of up to two per second; during each catch, the flight path is interrupted and thus appears erratic.

In many cases (especially in the families Nycteridae, Megadermatidae, Rhinolophidae, and Hipposideridae), there is little locomotion other than flight. These bats may move across the cave ceiling from which they hang, by shifting their toehold, one foot at a time. A few genera (especially among the Pteropodidae) may crawl along branches, in a slothlike posture, using their thumb claws as well as their feet. The Emballonuridae and Rhinopomatidae hang on vertical surfaces suspended by their hind claws, but with their thumbs and wrists propped against the surface. In this orientation, they can scramble rapidly up or down and forward or backward, as well as sideways.

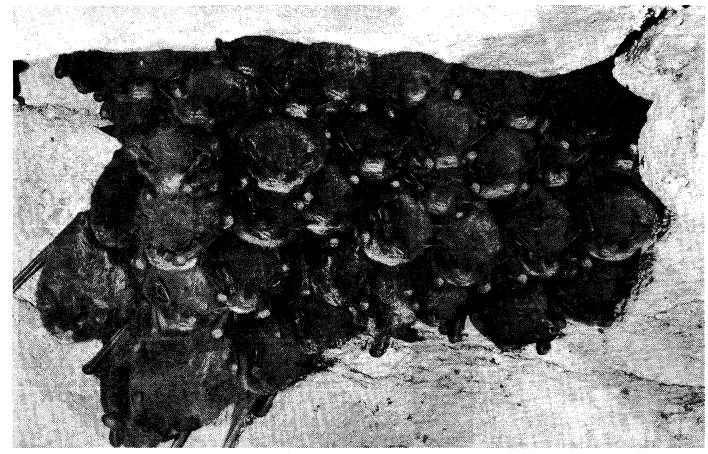
Bats of many genera (Vespertilionidae, Molossidae, Noctilionidae, Desmodontidae) walk or crawl on either horizontal or vertical surfaces using hindfeet, wrists, and thumbs. Many move freely either backward or forward, a convenience for entering and leaving crevices. The vampires may also leap from roost to roost. The Thyropteridae and Myzopodidae, as well as the vespertilionid *Tylonycteris,* have specialized wrist and sole pads for moving along and roosting on the smooth surface of leaves or bamboo stalks.

Bats are not known to swim in nature except, perhaps, by accident. When they do fall into the water, however, they generally swim competently.

Roosting. Bats choose a variety of diurnal roosts. Each species favours a particular kind of roost, though this varies with sex, season, and reproductive activity. Many bats favour isolated or secure roosts—caves, crevices in cliff faces, the interstices of boulder heaps, tree hollows, animal burrows, culverts, abandoned buildings, portions of buildings inaccessible to man (*i.e.,* roof, attic, hollow wall), and the hollow core of bamboo stalks. Some species roost externally—on tree trunks or in the branches of trees, under palm leaves, in unopened tubular leaves, or on the surface of rocks or buildings. For some, the darkness, stability of temperature and humidity, and isolation from predators provided by caves and crevices seem essential. Others prefer the heat and dryness of sun-exposed roosts. Many species choose special nursery or hibernation roosts. Buildings are so widely exploited by bats (especially Vespertilionidae, Molossidae, and Emballonuridae) for regular diurnal or nursery roosts that the numbers of many species have probably become more abundant since the advent of architecture. Many bats also occupy nocturnal roosts, often rocky overhangs or cave entrances, for napping, for chewing food, or for shelter from bad weather.

Bats are usually colonial; indeed, some form very large cave colonies. Generally, large colonies are formed by bats that roost in dense clusters, pressing against one another, although many roost spaced out, not touching. In trees, *Pteropus* may form outdoor camps numbering hundreds of thousands of individuals. Many species form smaller groups of several dozen to several hundred. Less commonly, bats are solitary; sometimes, the adult female roosts only with its most recent child. Occasionally one sex is colonial, and the other is apparently solitary. Some species regularly form mixed colonies (*e.g., Mormoops* and *Chilonycteris* with *Leptonycteris, Monophyllus, Carollia*). The advantages of colonial or solitary life and the factors that govern colony size in bats with colonial predilection have not yet been established.

Allan Roberts



Indiana bats (*Myotis sodalis*) hibernating on a cave roof.

The roost requirements of many bats, rather precise in terms of light, temperature, and humidity, limit their distribution in space. Some of the Megachiroptera strikingly defoliate the trees on which they roost.

Elaborate communities of other animals are often satellites of cave-bat colonies. Among these are cave crickets, roaches, blood-sucking bugs, a variety of parasites (*e.g.,*

fleas, lice, ticks, mites, and certain flies), and dermestid beetles and other insects that feed on cave-floor debris—guano, bat and insect corpses, and discarded pieces of food or seeds. Molds and other fungi are also conspicuous members of the cave-floor community. Bats and their excretions alter the cave environment by producing heat, carbon dioxide, and ammonia.

Migration. Many bats of temperate climates migrate annually to and from summer roosts and winter hibernation sites, with an individual often occupying the same roosts in seasonal sequence each year. Members of the same species may converge on a single hibernation cave or nursery roost from many directions, indicating that the choice of migration direction to and from these caves cannot be genetically determined. Migration time probably is genetically determined (*i.e.,* instinctive) and influenced also by weather conditions and the availability of food. Nothing is known of how migration goals are recognized or how their location is learned by succeeding generations. Female young, of course, are born at a nursery roost and may memorize its location, but how they know where to go at other times of year is not clear.

Separation of male and female migrants

Female Mexican free-tailed bats migrate from Central Mexico to Texas and adjacent states each spring, returning south in the fall. Mating probably occurs in transient roosts during the spring flight. The migration is believed to remove pregnant (and lactating) females to a region of high food supply where they need not compete with males of their own species. Presumably they return to Mexico for its suitable winter climate and food supply and to meet their mates.

The North American red and hoary bats (*Lasiurus borealis* and *L. cinereus*) and the silver-haired bat (*Lasionycteris noctivagans*) migrate in the fall from the northern U.S. and Canada to the southern states, returning in spring. Little is yet known of energy storage, navigation, or other specializations for migrations.

Orientation. Bats of the suborder Microchiroptera orient acoustically by echolocation ("sonar"). They emit short, high-frequency pulses of sound (usually well above the range of human hearing) and listen to the echoes returning from objects in the vicinity. By interpreting returning echoes, bats may identify the direction, distance, velocity, and some aspects of the size or nature or both of objects that draw their attention. Echolocation is used to locate and track flying and terrestrial prey, to avoid obstacles, and possibly to regulate altitude; orientation pulses may also serve as communication signals between bats of the same species. Bats of the megachiropteran genus, *Rousettus,* have independently evolved a parallel echolocation system for obstacle avoidance alone. Echolocation pulses are produced by vibrating membranes in the larynx and emitted via the nose or the mouth, depending upon species. Nose leaves in some species may serve to channel the sound.

The principle of echolocation

The echolocation signals spread in three dimensions on emission, the bulk of the energy in the hemisphere in front of the bat or in a cone-shaped region from the nostrils or mouth. When the sound impinges on an intervening surface (an insect or a leaf, for example) some of the energy in the signal is reflected or scattered, some absorbed, and some transmitted and reradiated on the far side; the proportion of sound energy in each category is a function of wavelength and of the dimensions, characteristics, and orientation of the object. The reflected sound spreads in three dimensions, and some portion of it may impinge on the bat's ears at perceptible energy levels.

Bats' external ears are generally large, probably enhancing their value for detecting direction of incoming signals, and their middle and inner ears are specialized for high-frequency sensitivity. In addition, the bony otic (auditory) complex is often isolated acoustically from the skull, probably enhancing signal comparison by both ears. The threshold and range of hearing have been studied in several genera of bats, and, in each case, the region of maximum sensitivity coincides with the prominent frequencies of the outgoing echolocation signals.

The characteristics of echolocation pulses vary with family and even with species. Echolocation pulses of a substantial number of bat species have been analyzed in terms of frequency, frequency pattern, duration, repetition rate, intensity, and directionality. The prominent frequency or frequencies range from 12 kiloHertz (one kiloHertz is equivalent to 1,000 cycles per second) to about 150 kiloHertz or more. Factors influencing frequency may include bat size, prey size, the energetics of sound production, inefficiency of propagation of high frequencies, and ambient noise levels.

Orientation pulses may be of several types. The individual pulse may include a frequency drop from beginning to end (frequency modulation, FM) or the frequency may be held constant (CF) during part of the pulse, followed by a brief FM sweep; either FM or CF pulses may have high harmonic content. The pulse duration varies with the species and the situation. In cruising flight, the pulses of the Asian false vampire (*Megaderma lyra*) are 1.5 milliseconds (0.0015 second), those of Dobson's mustache bat (*Chilonycteris psilotis*) 4 milliseconds, and those of the greater horseshoe bat 55–65 milliseconds. In goal-oriented flight, such as the pursuit of an insect or the evaluation of an obstacle or a landing perch, the pulse duration is systematically altered (usually shortened) with target distance, sometimes ending with pulses as short as 0.25 millisecond.

During insect pursuit, obstacle avoidance, and landing manoeuvres, there are three phases of pulse output design: search, approach, and terminal. The search phase, during which many bats emit about 10 pulses per second, precedes specific attention to a target. In the approach phase, which starts when the bat detects an object to which it subsequently devotes its attention, the bat raises the pulse rate to about 25 to 50 per second, shortens the pulses with decreasing distance, and often alters the frequency pattern. The terminal phase, which often lasts about 100 milliseconds, is characterized by extremely short pulses, repeated as rapidly as 200 or more per second, and ceases as the bat intercepts the target or passes it (the stimulus being, perhaps, the cessation of echoes); another search phase follows. During the brief terminal phase (a fraction of a second), the bat is engaged in final interception (or avoidance) manoeuvres and appears to pay little attention to other objects.

The use of echolocation to gain sensory information requires integration of the vocal and auditory centres of the brain, in addition to sensitive ears. Not only must the nervous system of the bat analyze in a few thousandths of a second the reflected, and thus altered, form of its own pulse, but it must separate this echo from those of other individuals and from others of its own pulses. All of this must be done while the animal (and often the target) is moving in space. In the laboratory, bats have been found to be able to identify, pursue, and capture as many as two fruitflies (*Drosophila,* about three millimetres [0.12 inch] long) per second, and to locate and avoid wires as fine as 0.1 or even 0.08 millimetre (0.004 or 0.003 inch) in diameter.

Capabilities of bat sonar

Research has provided some information on the mechanisms of bat sonar. There is evidence that the multiple frequencies of FM or harmonic patterns serve in determining target direction. The relative intensities of the various frequencies will be different at the two ears, allowing the animal to determine target direction when three or more frequencies are received. Target velocity may be measured by constant-frequency bats through the use of the Doppler shift, a change in perceived frequency due to the relative motion of the bat and its target. Changes in pulse–echo timing may provide information on target distance and velocity. The ratio of useful signal to background noise is increased by several mechanisms, including specializations of the middle ear and its ossicles (tiny bones), isolation of the cochlea (the area where sound energy is converted into nerve impulses), and adaptations of the central nervous system.

Food habits. Most bats feed on flying insects. In some cases, prey species have been identified from stomach contents or from discarded pieces under night roosts, but such studies have not yet provided an adequate measure of the spectrum of bat diets.

Insects are identified and tracked in flight by echolocation. Large insects may be intercepted with the wing membranes and pulled into the mouth. Some moths, however, (Noctuidae, Arctiidae) are able to avoid capture by bats.

Some genera of bats (*Macrotus, Antrozous, Plecotus, Nycteris*) feed on arthropods, such as large insects, spiders, and scorpions, which they find on the ground, on walls, or on vegetation. These bats may either land on and kill their prey before taking off with it or pick it up with their teeth while hovering.

Fishing by bats

Three genera (*Noctilio, Pizonyx,* and *Myotis*) include at least one species that catches small fish and possibly crustaceans. All fish-eating species also feed on flying insects or have close relatives who do so. Each is specialized in having exceptionally large hindfeet armed with long, strong claws with which the fish are gaffed.

The Megachiroptera and many of the phyllostomatid genera feed on a variety of fruits, often green or brown in colour; usually such fruits are either borne directly on wood or hang well away from the bulk of the tree and have a sour or musky odour.

The pteropodid subfamily Eonycterinae (and some other fruit bats) and the phyllostomatid Glossophaginae feed, at least in part, on nectar and pollen. Many tropical flowers, adapted for pollination by these bats, open at night, are white or inconspicuous, have a sour, rancid, or mammalian odour, and are borne on wood, on pendulous branches, or held beyond or above the bulk of the plant. The phyllostomatid Phyllonycterinae may also feed on flowers.

Several phyllostomatid and megadermatid genera are carnivorous, feeding on small rodents, shrews, bats, sleeping birds, tree frogs, and lizards. The true vampires, which feed on the blood of large mammals or birds, land near a quiet prospective victim, walk or jump to a vulnerable spot on it where the skin is relatively exposed—the edge of the ear or nostril, around the anus, or between the toes, for example—make a scooping, superficial bite from which the blood oozes freely, and lap the blood with very specialized tongue movements. Each vampire requires about 15 millilitres (about one cubic inch) of blood per night.

The interaction of bats with their food, be it insects, fruit, or flowers, probably has a substantial impact on some biological communities. Many plants are dependent on bats for pollination; other plants benefit from seed dispersal by bats. Moths of two families are known to take evasive or protective action on hearing bat pulses nearby, an adaptation that implies heavy predation.

Maintenance behaviour. Bats are meticulous in their grooming, spending a fair part of the day and night combing and grooming their fur and cleansing their wing membranes. Generally they comb with the claws of one foot, while hanging by the other; they remove the combings and moisten their claws with their lips and tongue. On the wing membranes, in particular, they use the mouth meticulously, perhaps oiling the skin with the secretions of dermal (skin) glands while cleansing it.

Social interactions. Bats often segregate by sex. As noted, in many species, pregnant females occupy special nursery roosts until their young are independent. In some species, the sexes occupy the same general roost but gather in separate clusters. In others, the sexes intermingle or arrange themselves into a pattern within a group—the females centrally, for example, and the males peripherally. Sexual segregation during foraging has been reported for several species. Among bats that migrate over long distances, such as Mexican free-tailed, red, and hoary bats, the sexes may meet only briefly each year.

FORM AND FUNCTION

Anatomical specializations. Bats are mammals with front limbs modified for flight. The fingers, other than the thumb, are greatly elongated and are joined by a membrane that also extends from the posterior border of the forearm and upper arm to the side of the body and leg as far as the ankle or foot. Only the thumb, and occasionally the index finger, end with a claw. The wing membrane consists of two layers of skin, generally darkly pigmented and naked, between which course blood vessels and nerves.

Structure of the wing

When not fully extended, the wing skin is gathered into wrinkled folds by elastic connective tissue and muscle fibres. Some of the fingers, especially the third, fold over when the bat is not in flight; the wing may then be quite tightly folded or may partly enfold the bat's undersurface. The thumb, always free of the wing membrane, is used for walking or climbing in some species; in others, it is used for handling food. Bats that walk often have pads or suction disks on their thumbs or wrists or both, and many female bats use the thumbs for suspending themselves, hammock fashion, when giving birth.

Wing shape, governed by the relative lengths of the forearm and the fingers, varies greatly, in adaptation to flight characteristics.

Most bats have a membrane, consisting of skin like

Drawn by R. Keane, based on *Natural History* (October 1958)

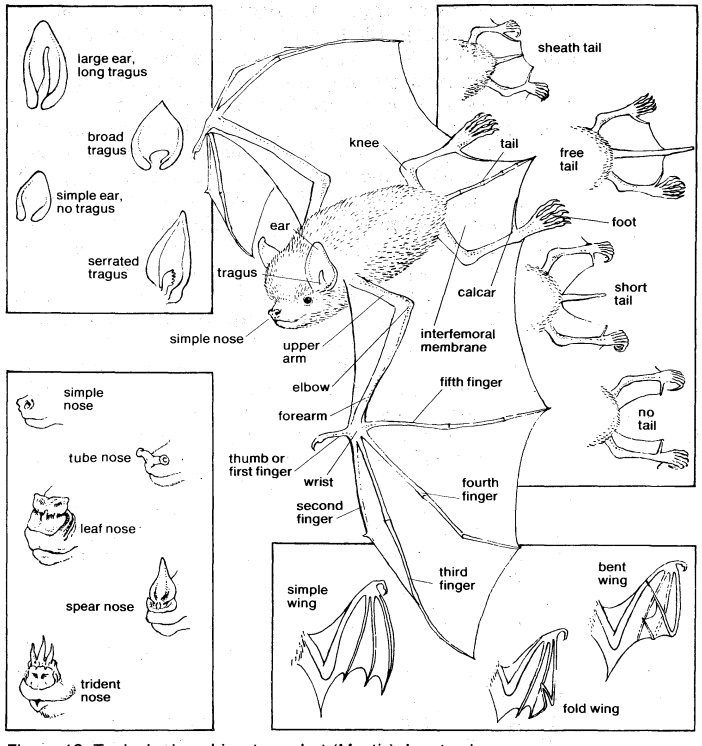


Figure 18: Typical microchiropteran bat (*Myotis*). Insets show variations of structures found in other bats.

the wings, that extends between their legs (interfemoral membrane). In the midline, the interfemoral membrane is usually supported, at least in part, by the tail, and the distal edges are often shaped in flight by greatly elongated heel bones. The interfemoral membrane, especially well developed in insectivorous, carnivorous, and fish-eating bats, is less well developed or even absent in the vampires and in fruit- and flower-feeding bats. Many bats in flight, on catching large prey, bring the membrane forward and by flexing the neck and back tuck the prey against and into the membrane while taking a new tooth hold. By this manoeuvre, the bat takes hold of the victim head first and is able to kill or disable it promptly.

At rest, a bat's head is its most striking feature. The pinna (projecting portion) of the external ear usually is extremely large and often funnel shaped. In several genera that feed on terrestrial arthropods, the ears are particularly oversized, probably for ultraprecise directional assessment. The tragus (a projection on the anterior side of the auditory canal) or antitragus (on the posterior side) may also be conspicuous. The ears are often highly mobile, sometimes flicking back and forth in phase with the production of sonar signals. In some species the ears are immobile, but, in all cases, they probably function jointly for directional analysis.

Shape of the face

Bats often have a muzzle of somewhat rodent-like or foxlike proportions but, in many, the face has a pushed in, pug-dog appearance. In the nectar feeders, the snout is elongated to house the long, extensible tongue. Many bats

have a facial ornament, the nose leaf, consisting of skin and connective tissue surrounding the nostrils and extending as a free flap or flaps above the nostrils and in front of the face. The complexity and shape of the nose leaf varies with family; its presence correlates with nasal emission of orientation signals. Thus, it is supposed that the nose leaf influences sound output, perhaps by narrowing the beam, but evidence is sparse.

A bat's neck is likely to be short and relatively immobile, the chest and shoulders large and well muscled, the hips and legs slender.

Most bats are well furred (except for the wing membranes), generally in shades of brown, tan, gray, or black on top and in lighter shades ventrally. Red, yellow, or orange variants occur in many species. Among species that roost in the open, speckled or mottled fur, bright or light-coloured spots or stripes, or bright red, yellow, or orange shading on the pendant head, neck, and shoulders are common. Mottled fur may blend with lichen-covered bark or rock. Bright spots may simulate the speckled sunlight of the forest canopy as seen from below. Stripes probably break up contours. The colouring seen while the animal is hanging may be a kind of countershading for concealment or it may enhance the bat's simulation of a ripening fruit or a dead leaf. Many bats who roost externally hang from a branch by one foot, which then looks like a plant stem.

Many bats have large dermal (skin) glands (the location of which depends on family), which secrete odorous substances that may serve as species or sex recognition signals. Some glands may also supply oils for conditioning the skin or waterproofing the pelage.

Lowering of body temperature

Thermoregulation. Although some bats maintain fairly even body temperatures, a large number undergo periodic raising or lowering of their temperature. Many of the vespertilionids and rhinolophids and a few molossids drop their body temperature to ambient temperature shortly after coming to rest (a condition called heterothermy). They raise their temperature again on being aroused or when readying themselves for nocturnal foraging. When fully active, bats have a body temperature of about 37° C (98.6° F). During pregnancy, lactation, and juvenile growth, bats probably thermoregulate differently, more closely approximating stability. The drop in body temperature, if the ambient temperature is relatively low, results in the bat's assuming a lethargic state. Energy is conserved by "turning down the thermostat," but the bat is rendered relatively unresponsive to threats by predators or weather. Heterothermic bats, therefore, generally roost in secluded, isolated sites, often in crevices, and have specializations for arousal. One or more sensory systems and the brain remain sensitive at low temperatures and initiate the necessary heat production for arousal. Heat is generated by the metabolism of brown fat and by shivering.

Many bats that exhibit daily torpor hibernate during the winter and must, therefore, store energy as body fat in the fall, increasing their weight by 50 to 100 percent. They must also migrate from the summer roost to a suitable hibernation site (often a cave); *i.e.,* one that will remain cool and humid but will not drop below freezing throughout the winter. Large populations often aggregate in such caves. Hibernation also involves the suspension of temperature regulation for long periods; adaptations of circulation, respiration, and renal function; and suspension of at least some aspects of diurnal activity. Bats of hibernating species generally court and mate in the fall when the males are at their nutritional peak.

Bats of several tropical families are homeothermic. A spectrum of degrees of homeothermy and heterothermy probably will be discovered in the order.

Digestion and water conservation. Digestion in bats is unusually rapid. They chew and fragment their food exceptionally thoroughly and thus expose a large surface area of it to digestive action. They may begin to defecate 30 to 60 minutes after beginning to feed, thus reducing the load that must be carried in flight.

Some bats live in sun-baked roosts without access to water during the day. They may choose these roosts for their heat, thus conserving body heat, but it is not yet known how they hold their body temperature down without using water. In the laboratory, bats die if body temperature rises above about 40–41° C (104°–106° F).

Use of the eyes

Senses. Bats have been considered in folklore to be blind. In fact, the eyes in the Microchiroptera are small and have not been well studied as yet. Among the Megachiroptera, the eyes are large, but vision has been studied in detail only in *Pteropus.* These bats are able to make visual discriminations at lower light levels than can man. The Megachiroptera fly at night, of course, and some genera fly below or in the jungle canopy where light levels are very low. Except for *Rousettus,* none are known to orient acoustically.

Studies of several genera of Microchiroptera have revealed that vision is of use in long distance navigation and that obstacles and motion can be detected visually. Bats also presumably use vision to distinguish day from night and to synchronize their internal clock with the local day-night cycle.

The senses of taste, smell, and touch in bats do not seem to be strikingly different from those of related mammals. Smell is probably used as an aid in locating fruit and flowers and possibly, in the case of vampires, large vertebrates. It may also be used for locating an occupied roost, colleagues of the same species, and individuals of the correct sex. Many bats depend upon touch, aided by well-developed facial and toe whiskers and possibly by the projecting tail, to place themselves in comforting body contact with rock surfaces or with colleagues in the roost.

EVOLUTION AND PALEONTOLOGY

The fossil record of bats prior to the Pleistocene (about 2,500,000 years ago) is limited and reveals little about bat evolution. Most fossils can be attributed to living families. Skulls and teeth compatible with early bats are known from the Paleocene (about 60,000,000 years ago), but these fossils may equally well have been from insectivores, from which bats are clearly separable only on the basis of adaptations for flight. By the middle Eocene (45,000,000 years ago) bats with full flight had evolved.

Two main evolutionary lines of bats

The Chiroptera are readily divided into two suborders—Megachiroptera and Microchiroptera. The Megachiroptera orient visually and exhibit a number of primitive skeletal features. The Microchiroptera orient acoustically. It is not certain that they have a common origin. The suborders either evolved separately from flightless Insectivora or diverged very early in chiropteran history.

The two principal geographic centres of bat evolution appear to be the Australo-Malaysian region, with about 290 species, and the New World tropics, with about 230 species. Comparable ecological niches in the Old and New World are largely occupied by different genera of bats, usually of different families.

CLASSIFICATION

Distinguishing taxonomic features. The order Chiroptera is defined by true flight. The elongated finger bones support the wing membrane. Bats are also characterized by their generally small size, marked pectoral specialization for flight, and relatively weak pelvic and leg development. The ulna is reduced, claws are absent on fingers except for the thumb (and occasionally the second finger), and the knee is directed posteriorly and laterally. The maximum complement of milk teeth is 22 and of permanent teeth, 38; the minimum of permanent teeth is 20. The dental formula indicates the number of pairs of upper and lower incisors, canines, premolars, and molars, respectively, and total number of teeth.

Annotated classification.

ORDER CHIROPTERA
Eighteen living families, 174 genera, and about 900 living species.

Suborder Megachiroptera
One family.

Family Pteropodidae (flying foxes and Old World fruit bats)
Generally large, fruit- or flower-feeding; lack acoustic orientation (except *Rousettus*); ears small, eyes large, vision well-developed; generally roost in trees, often colonial; often show countershading, cryptic markings, or bright fur colours or patterns. Index finger generally clawed, tail short or lacking, inter-

femoral membrane reduced. Muzzle simple in appearance (except *Hypsignathus*), generally cannot walk but can move along branches in hanging posture. Forearm length varies from 37 mm (1.46 inches; *Macroglossus*) to 220 mm (8.66 inches; *Pteropus vampyrus*). Teeth modified for fruit- and flower-feeding. Dental formula $\frac{(1-2)\,.\,1\,.\,3\,.\,(1-2)}{(0-2)\,.\,1\,.\,3\,.\,(2-3)} = 24\text{–}34$. Old World tropics and subtropics, including many Pacific islands; 39 Recent genera, about 154 species.

Suborder Microchiroptera

Seventeen families.

Family Rhinopomatidae (mouse-tailed bats)

Small, insectivorous; tail very long and largely free beyond a narrow interfemoral membrane, forearm very long, ears large, small nose leaf, primitive shoulder girdle. Dental formula $\frac{1\,.\,1\,.\,1\,.\,3}{2\,.\,1\,.\,2\,.\,3} = 28$. Store fat (probably on a seasonal basis). Roost on vertical surfaces, probably not in total darkness or isolation. Tropical distribution from northern Africa through southern Asia as far as Sumatra; 1 genus, 3 species. Generally considered to be the most primitive of living Microchiroptera.

Family Emballonuridae (sheath-tailed or sac-winged bats)

Small to medium sized. Ears large but simply shaped, eyes small, muzzle sharp but plain; tail short, perforating dorsal surface of well-developed interfemoral membrane. Several genera have a glandular pouch in the wing extension, anterior to the arm. Relatively unspecialized shoulder girdle and arm articulation. Dental formula $\frac{2\,.\,1\,.\,2\,.\,3}{3\,.\,1\,.\,2\,.\,3} = 34$, $\frac{1\,.\,1\,.\,2\,.\,3}{3\,.\,1\,.\,2\,.\,3} = 32$, or $\frac{1\,.\,1\,.\,2\,.\,3}{2\,.\,1\,.\,2\,.\,3} = 30$. Insectivorous; roost on vertical surfaces, such as tree trunks, cliff faces, cave entrances, and walls. Some favour buildings, especially belfries; some densely colonial but not touching one another; others form small groups or are solitary. Hang suspended from toes with wrists propped against wall. Worldwide tropical distribution (excluding West Indies and some Pacific islands); 12 genera, about 50 species; each genus restricted to either Old or New World.

Family Nycteridae (slit-faced or hollow-faced bats)

Small to medium sized, the humerus and pectoral girdle not greatly specialized, skull with peculiar nasal fossa (depression), cleft nose leaf, and a deep midline facial cleft behind and above the nostrils. Ears large, wings broad, tail long with bifid (split) end, calcars (heel bones) greatly elongated, tail and calcars support well-developed interfemoral membrane. Dental formula $\frac{2\,.\,1\,.\,1\,.\,3}{3\,.\,1\,.\,2\,.\,3} = 32$. Insectivorous, mostly preying on terrestrial forms or those resting on vegetation, rocks, or walls. Cannot walk. Roosts usually dark and humid, some species roosting externally in jungle canopy. Generally form small nontouching colonies, but some are solitary. Distributed through most of tropical Africa, Malaysia, and Indonesia; 1 genus, 13 species.

Family Megadermatidae (false vampires)

Moderately large bats; external ears very large and fused across midline; tragus bifid; nose leaf large with truncated end; eyes relatively large; wings broad, interfemoral membrane well-developed and supported distally by heel bones, no external tail. Females bear false inguinal nipples. Premaxillae lacking; dental formula $\frac{0\,.\,1\,.\,(1-2)\,.\,3}{2\,.\,1\,.\,2\,.\,3} = 26\text{–}28$. Insectivorous, principally on terrestrial arthropods, as in Nycteridae; at least 2 species, *Megaderma lyra* and *Macroderma gigas*, also feed on small vertebrates hunted and taken in the same fashion as arthropod prey. Cannot walk. Form small nontouching colonies usually in dark, secluded caves or abandoned buildings. *Lavia frons* is at least partly diurnal and roost in trees in the savanna and open forest. Central Africa, Southeast Asia, and Australia; 4 genera, 5 species.

Family Hipposideridae (Old World leaf-nosed bats)

Small to large bats. Complex nose leaf with subordinate leaflets and compartments; large ears, widely separated and highly mobile, antitragus well-developed; interfemoral membrane well-developed, tail generally projects slightly beyond distal edge of membrane. Often bear glandular pouch on forehead; females have false inguinal nipples. Dental formula $\frac{1\,.\,1\,.\,2\,.\,3}{2\,.\,1\,.\,2\,.\,3} = 30$. Colour usually drab brownish but red phases not uncommon. Insectivorous, usually on flying insects. Mostly colonial, nontouching, and roosting in humid caves, tree hollows, culverts, or buildings. A few roost externally in branches of trees, a few are solitary. Do not walk. Old World tropics; 9 genera, about 60 species, the genus *Hipposideros* particularly successful in numbers of species and individuals throughout the Old World tropics.

Family Rhinolophidae (horseshoe bats)

Small to moderately large size. Complex nose leaf; large, highly mobile ears, well-developed antitragus; wings short and rounded; well-developed interfemoral membrane, supported by tail; calcanea (backs of heels) weak. Fur generally brown (red phases occur); females bear false inguinal nipples. Dental formula $\frac{1\,.\,1\,.\,2\,.\,3}{2\,.\,1\,.\,3\,.\,3} = 32$. Dark, humid roosts selected, especially caves, but tree hollows, buildings, and culverts as well; generally colonial, nontouching; cannot walk. Insectivorous, usually on flying insects. Old World, including parts of western Europe, the U.S.S.R., and Japan; 2 genera, about 70 species, the genus *Rhinolophus* one of the most successful in species and numbers.

Family Noctilionidae (bulldog bats)

Medium sized. Muzzle heavy but unadorned; lips full; internal cheek pouches; ears large, pointed, and mobile; wings long and narrow. Dental formula $\frac{2\,.\,1\,.\,1\,.\,3}{1\,.\,1\,.\,2\,.\,3} = 28$. Tail well-developed, extending to midpoint of large interfemoral membrane, which is pierced dorsally by tail tip, membrane supported distally by very well-developed calcars and calcanea. Feet large, or very large (*N. leporinus*). Colour dark brown to rufous orange dorsally; musky odour. Walk well, often roost in crevices, tree hollows, attics, grottoes, and caves; colonial, in touching clusters. *Noctilio labialis* feeds on flying insects, as does *N. leporinus*, which also gaffs fish. Tropical America; 1 genus, 2 species.

Family Mormoopidae

Small; insectivorous on flying insects; some walk; all lack nose leaf but have elaborate lip leaves; colour from brown through orange, red, and yellow. Densely colonial in dark caves, colonies often number tens of thousands. Tail and interfemoral membrane well-developed. Dental formula $\frac{2\,.\,1\,.\,2\,.\,3}{2\,.\,1\,.\,3\,.\,3} = 34$. Tropical Central and South America; 3 genera, 9 species.

Family Phyllostomatidae (American leaf-nosed bats)

Small to large size. Nose leaf simply shaped, ears often large and mobile, wings generally short and broad, tail and interfemoral membrane quite varied (from absent to well-developed); dental formula varied, from 26 to 34; fur colour and patterns varied. Insectivorous (the insects eaten may be flying or terrestrial types), carnivorous, fruit- and flower-feeding species. Generally do not walk. Colonial, often densely so; generally roost in touching clusters in caves, tree hollows, buildings, culverts, or in the open under bridges, eaves, in the crests of palm trees, or on the underside of palm leaves. Flight swift and straight to hovering. Southwestern U.S. through tropical America; 47 genera, about 120 species.

Subfamily Phyllostomatinae. Medium to large. Nose leaf often of striking size, ears large. Insectivorous, fruit-feeding, or carnivorous. Varied dental formulas. Subfamily may be an artificial grouping.

Subfamily Glossophaginae. Small to medium sized. Small nose leaf, snout elongated; tongue elongated and extensible. Wings broad; hovering flight. Feed on pollen, nectar, fruit.

Subfamilies Carolliinae, Strunirinae, Stenoderminae. Medium sized. Fruit-feeding; some may be insectivorous. Nose leaves well-developed. Many Stenoderminae have white or light facial stripes, may roost in the open; 2 species alter palm leaves as roosts.

Subfamily Phyllonycterinae. Small bats, endemic to West Indies, probably fruit- and flower-feeding. Legs and feet exceptionally well-developed and body flexible. Beige, maize, or light brown; cave-dwelling; nose leaves very small; snout and tongue moderately elongated.

Family Desmodontidae (vampire bats)

Medium-sized. Small nose leaf. Teeth highly specialized for cutting skin, cheek teeth reduced; dental formula in *Desmodus* $\frac{1\,.\,1\,.\,1\,.\,1}{2\,.\,1\,.\,2\,.\,1} = 20$. Hindlegs and thumbs very well adapted for walking and jumping. Tail absent; interfemoral membrane reduced. Feed on blood of large mammals or birds; roost in caves, hollow trees and culverts; colonial. Most of tropical America, excluding West Indies; 3 genera, 3 species.

Family Natalidae (funnel-eared bats)

Small, slenderly built. Gray, buffy, yellow, or reddish; fur deep. Well-developed tail and interfemoral membrane. Ears large snout plain: dental formula $\frac{2\,.\,1\,.\,3\,.\,3}{3\,.\,1\,.\,3\,.\,3} = 38$. Natalids walk clumsily but do not enter crevices; they are cave-dwelling, colonial in nontouching groups; feed on flying insects. Central America, and northern South America, West Indies; 1 genus, 4 species.

Family Furipteridae (smoky bats)
Small, delicately built. Thumb vestigial; snout plain; tail long, ending short of distal edge of well-developed interfemoral membrane; legs long; feet small. Dental formula $\frac{2 \cdot 1 \cdot 2 \cdot 3}{3 \cdot 1 \cdot 3 \cdot 3} = 36$. Biology unknown; probably insectivorous. Northern South America; 2 genera, 2 species.

Family Thyropteridae (disk-wing bats)
Second finger reduced to rudiment, base of thumb and sole provided with sucking disk, simple muzzle, ears large; dental formula $\frac{2 \cdot 1 \cdot 3 \cdot 3}{3 \cdot 1 \cdot 3 \cdot 3} = 38$. Insectivorous, roost alone or in small groups, often in still furled banana leaves. Biology poorly known. Central America and northern South America, excluding West Indies; 1 genus, 2 species.

Family Myzopodidae (Old World sucker-footed bats)
Small, plain muzzle, large ears with peculiar mushroom-shaped lobe, dental formula $\frac{2 \cdot 1 \cdot 3 \cdot 3}{3 \cdot 1 \cdot 3 \cdot 3} = 38$. Thumb and sole with adhesive disks, vestigial thumb claw; tail extends free beyond interfemoral membrane; specialized scapulo-humeral articulations. Probably insectivorous (biology unknown); endemic to Madagascar; 1 species.

Family Vespertilionidae (common bats)
Small to medium sized. Muzzle plain; eyes small; ears moderate to large, tragus well-developed; dental formula varied, $\frac{1 \cdot 1 \cdot 1 \cdot 3}{2 \cdot 1 \cdot 2 \cdot 3} = 28$ to $\frac{2 \cdot 1 \cdot 3 \cdot 3}{3 \cdot 1 \cdot 3 \cdot 3} = 38$. Wings generally long and moderately narrow; tail and interfemoral membrane well-developed. All walk well, often entering crevices. Insectivorous, some on flying, some on terrestrial insects; a few genera (*Pizonyx* and some *Myotis*) eat fishes. Mostly roost in caves, attics, barns, hollow trees, boulder heaps, the twig work of birds' nests, or roof thatching; a few (*Lasiurus*) roost in branches, on tree trunks, or in hollow core of bamboo (*Tylonycteris*). Generally colonial in dense, touching clusters; a few solitary. Many temperate species hibernate and migrate; many drably coloured; several that roost externally are either spotted (*Euderma*) or speckled (*Lasiurus*). Family worldwide to tree line, including many oceanic islands; some genera (*Myotis, Eptesicus,* and *Pipistrellus*) also worldwide; 35 genera, about 290 species.

Family Mystacinidae (New Zealand short-tailed bats)
Small, with simple vespertilionid-like head; highly adapted for walking. Wings fold exceptionally compactly; thumb and toe claws long and sharp; tail perforates interfemoral membrane dorsally. Dental formula $\frac{1 \cdot 1 \cdot 2 \cdot 3}{1 \cdot 1 \cdot 2 \cdot 3} = 28$. Feeds on flying and terrestrial insects. Biology poorly known. New Zealand; 1 species.

Family Molossidae (free-tailed bats)
Robustly built, small to very large. Tail projects well beyond well-developed interfemoral membrane; ears large, rather immobile, often fused to one another, and of very unusual shapes; lips and snout often heavy, eyes tiny. Wings very long and narrow; legs well-developed for walking; toes often bear bristles; dental formula varied from $\frac{1 \cdot 1 \cdot 1 \cdot 3}{1 \cdot 1 \cdot 2 \cdot 3} = 26$ to $\frac{1 \cdot 1 \cdot 2 \cdot 3}{3 \cdot 1 \cdot 2 \cdot 3} = 32$. Often have a conspicuous odour; fur short, usually dark brown or black. Many highly colonial, as many as millions clustering in dense, touching groups; roost pressed against fellows or walls, often in crevices. Occupy caves, tree hollows, and buildings. Walk exceptionally well; many prefer hot, dry roosts. Feed on flying insects; some migrate. Worldwide in tropics and subtropics, with a few species ranging into mild temperate regions; 11 genera, about 90 species (the genus *Tadarida* worldwide).

(A.N.)

Primates (lemurs, lorises, tarsiers, monkeys, apes, and hominids, including man)

The mammalian order Primates, which includes the prosimians, monkeys, apes, and man, has long excited human interest, because it contains man's closest relatives. The obvious similarity of monkeys and apes to man was noted long before any genetic relationship was postulated. Recognizing that monkeys possess intelligence substantially exceeding that of other familiar mammals, man has valued these animals as companions and pets for centuries. In recent decades, in addition to becoming increasingly popular as pets, nonhuman primates have served as substitutes for man in situations, such as experimental medicine and space science, which require near-human biological reactions.

The two main groups of the order Primates are the prosimians (lemurs, lorises, and tarsiers) and the anthropoids (monkeys, apes, and man). As classified at present, the suborder Prosimii consists of six families; the Tupaiidae (tree shrews, though these are considered by some authorities to belong in the order Insectivora, a view elaborated in the section *Insectivora*), Lemuridae (lemurs), Indriidae (indri, sifaka, and avahi), Daubentoniidae (aye-aye), Lorisidae (galagos and lorises), and Tarsiidae (tarsiers). Collectively, the prosimian suborder is frequently referred to as lower primates.

The suborder Anthropoidea, sometimes called "higher" primates, also comprises six families: Callitrichidae (marmosets and tamarins), Cebidae (South American monkeys other than marmosets), Cercopithecidae (African and Asian monkeys), Hylobatidae (lesser apes: siamangs and



Figure 19: Representative prosimians.

gibbons), Pongidae (great apes: orangutan [orangutan], gorilla, and chimpanzee), and Hominidae (men, living and extinct). The two geographically separated stocks, the Old World and the New World monkeys, are often referred to as catarrhines and platyrrhines, respectively, terms which derive from the shape of the nose (see below, *Form and function*).

GENERAL CONSIDERATIONS

Historical background of primate studies. The inclusion of man in the order provides the rationale for the vigour with which this group has been studied by scientists since the time of Galen of Pergamum. Aristotle and Hippocrates, in the 3rd and 4th centuries BC, recognized the similarity of man and apes, but it was Galen who demonstrated the truth of this kinship by dissection. He wrote, "the ape is likest to man in viscera, muscles, arteries, veins, nerves and in the form of bones." It should be noted that Galen was in fact referring to monkeys, not the true apes, which were unknown to western man until the 15th century. None of these early scientists saw any evolutionary significance in the similarity of man and "apes," a correspondence that they regarded as purely coincidental. An inkling of the truth of man's relationship with primates must have penetrated the mind of St. Albertus Magnus, probably the best leading naturalist of the Middle Ages, who produced a classification of animal life in his book *De animalibus.* Albertus' classification, which placed man between "apes" (monkeys) on the one hand and "animals" on the other, provides the first whiff of the "missing-link" concept, which later was to befog the issue of man's place in nature.

The Dark Ages were aptly named as far as knowledge of primates is concerned. The first evidence of a renaissance of interest was in the time of Vesalius, the great Belgian anatomist of the 16th century, who published a comparative anatomy of man and "apes" in order to confound the precepts of Galen. He did not succeed in disproving Galen's assertion that "apes were likest to man" but, unwittingly, he succeeded in stirring up an interest in the biology of primates that has never since flagged. The first true ape studied as a scientific specimen was a chimpanzee dissected by Edward Tyson, an English anatomist, in 1699. Tyson's specimen, which he called the "Orang-Outang, sive Homo Sylvestris," is to this day housed in the British Museum (Natural History), mounted in a standing position that reflects Tyson's belief that he had discovered the pygmy, a race of humans known since the time of the ancient Greeks. Tyson wrote of his "pygmie" that it was "no man, nor yet a common ape but a sort of animal between both." It never occurred to Tyson or his contemporaries, who believed that all animals had been created independently in their current image, that man, apes, and monkeys were connected by common evolutionary descent. In 1758, Linnaeus—the father of animal and plant classification—added the lemurs and bats to the monkeys, apes, and man and called the whole assemblage the Primates. It is to be noted, however, that Linnaeus was sufficiently perceptive to see that man was a primate. His conclusion was regarded as a grave blow to human dignity, and it was followed by new classifications such as that of Blumenbach in 1776, placing man in a separate order. Man was not again considered part of the primate order until a century later when the English anatomist St. George Mivart in the climate of post-Darwinian thought published his classification of primates.

Discovery of the first true ape

The first evolutionist was a French scholar of the late 18th century, the Chevalier de Lamarck, who saw animal life as an uninterrupted continuity in which old species were transformed into new species in a sequence of increasing complexity and perfection. In 1821 Baron Georges Cuvier, a rabid anti-evolutionist, had the historic distinction of describing *Adapis,* the first fossil primate genus ever recognized. Fossils such as *Adapis,* Cuvier believed, were the remains of animals destroyed by past catastrophes (floods, earthquakes, etc.) and living animals were new stocks divinely created to fill the vacuum, a view consistent with the widely held notion that species were immutable. During the early 19th century a number of geologists and biologists questioned the doctrine of immutability, but it was not until 1859, with the publication of Charles Darwin's "On the Origin of Species by Means of Natural Selection," that positive evidence was provided, along with a sound alternative theory. The Darwinian contention that not only had man evolved, but he had evolved from a simian ancestor resulted in acrimonious debate among scientists, theologians, philosophers, and laymen. As influential zoologists and anatomists rose to support Darwin, the truth of man's primate consanguinity began to be accepted, if not actually relished. Today, few scientists deny that man and the lower primates belong in the same order; in fact, much current research is directed toward closing the apparent gap between the highest of the nonhuman primates, the chimpanzee and gorilla, and man.

In times past, the public image of primates was largely dictated by prevailing religious beliefs. In Asian countries, where primates abound, monkeys have for a long time been regarded with various degrees of deference that, among Hindus in India, for instance, amounts to worship. In Europe and North America, where monkeys and apes are totally absent, no religious sect has attached divine significance to them and, in fact, the reverse has been the case, monkeys at various times having been regarded as the personification of evil and depravity, familiars of the devil. This image, however, is fading as a result of enlightened instruction in schools and the advances in naturalistic presentation of primates in zoos. The nonhuman primate is becoming generally accepted by laymen as an animal of peculiar interest to man with many amusing and endearing qualities.

Religious views toward primates

The qualities of primates, however, are less endearing to farmers and agriculturists in certain parts of the world. In South Africa, the chacma baboon (*Papio ursinus*) competes with domestic sheep for grazing lands and is an occasional predator of lambs; in West and Central Africa native crops are subject to daily assaults by forest-living monkeys; and, in India, macaques, which have been accorded a semi-sacred status, live alongside man in towns and villages and are parasitic upon him for their food and shelter.

Scientific interest in nonhuman primates, their structure at all levels and their way of life, is currently in the ascendancy. Their value as research animals increases year by year, so that at present over a quarter of a million wild monkeys are consumed annually by laboratories in the study of human diseases, in the production of vaccines, in the experiments of organ transplantation, in the testing of drugs and, even, for the clinical trials of new cosmetics. Clearly their scientific usefulness raises important problems of conservation of primate stocks in the wild. Other disciplines whose researches depend upon observations and upon experimentation with nonhuman primates include those of endocrinology, neurology, psychology, and sociology. Much is being learned as a result of such studies that is of great significance for man and his future betterment.

Size range and adaptive diversity. Members of the order Primates show a remarkable range of size and adaptive diversity. The smallest primate (in linear dimensions) is probably the dwarf or mouse lemur (*Microcebus murinus*) of Madagascar, which measures approximately 13 centimetres (about five inches) in head and body length; the most massive is certainly the gorilla (*Gorilla gorilla*), whose head and torso length may reach 100 centimetres (39 inches; the comparable dimension in man may exceed this figure, but the gorilla is considerably more robust). In terms of body weight, the range is even more striking. The pen-tailed tree shrew (*Ptilocercus lowii*) weighs on average 46 grams (1.62 ounces); the gorilla's weight may be more than 3,000 times as great, varying from 140 to 180 kilograms (about 300–400 pounds).

Size and weight

In terms of habitat, primates are tropical animals and occupy two major vegetational zones: the tropical forest and woodland–grassland vegetational complexes. Each of these zones has produced in its resident primates the appropriate adaptations, which are not remarkable for their

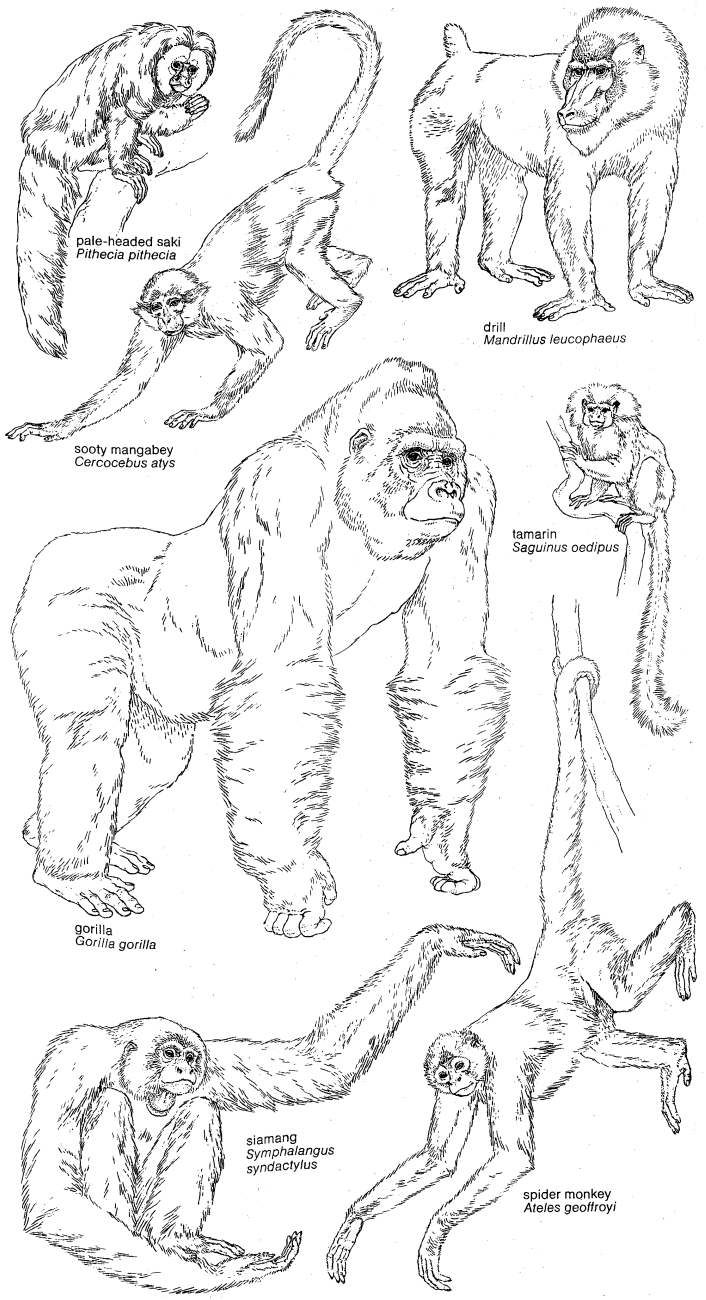


Figure 20: Body plans of representative anthropoids.
Drawing by R. Keane

diversity. There is perhaps more diversity of bodily form to be seen amongst forest-living species than among savanna inhabitants. One of the explanations of this difference is that it is the precise pattern of locomotion rather than the simple matter of habitat that governs overt bodily adaptations. Within the forest there are a number of ways of moving about. An animal can live on the forest floor or in the canopy, for instance, and within the canopy it can move in three particular ways: (a) by leaping—a function principally dictated by the hindlimbs; (b) by arm swinging (brachiation)—a function particularly of the forelimbs; (c) by quadrupedalism—a function equally divided between the forelimbs and the hindlimbs. On the savanna, or in the woodland-savanna biome, which substantially demands adaptations for ground-living locomotion rather than those of tree-living, the possibilities are limited. If bipedal man is discounted, there is a single pattern of ground-living locomotion, which is called quadrupedalism. Within this category there are at least two variations on the theme: (a) knuckle-walking quadrupedalism, and (b) digitigrade quadrupedalism. The former gait is characteristic of the African apes (chimpanzee and gorilla), and the latter of baboons and macaques, who walk on the flats of their fingers. After man, the Cercopithecinae are the most successful colonizers of nonarboreal habitats.

The structural adaptations of primates resulting from locomotor differences are considered below in more detail, but they do not prove to be very extensive. Primates are a homogeneous group in terms of morphology, and it is only in the realm of behaviour that differences between primate taxa are clearly discriminant. It can be said that the most successful primates (judged in terms of the usual criteria of population numbers and territorial spread) are those that have departed least from the ancestral pattern of structure, but furthest from the ancestral pattern of behaviour. "Manners makyth man" is true in the widest sense of the word; in the same sense, manners delineate primate species.

Distribution and abundance. The nonhuman primates have a wide distribution throughout the tropical latitudes of Africa, India, Southeast Asia, and South America. Within this tropical belt, which lies between latitudes 25° N and 30° S, they have a considerable altitudinal range. In Ethiopia the gelada baboon (*Theropithecus*) is found living at altitudes up to 5,000 metres (16,000 feet). Mountain gorillas of the Virunga Volcanoes are known to travel across mountain passes at altitudes of more than 4,200 metres (13,800 feet) when travelling from one high valley to another. The howler monkeys of Venezuela (*Alouatta seniculus*) live at 2,500 metres (8,000 feet) in the Cordillera da Merida and in northern Colombia the dourocouli (*Aotus*) is found in the tropical montane forests of the Cordillera Central.

Temperate zone species

A few primate species occur naturally in temperate latitudes. The rhesus monkey (*Macaca mulatta*), for example, is found near Peking, and Japanese macaques (*Macaca fuscata*) are to be found on the Shimokita Peninsula, a hatchet-shaped promontory of the Japanese island of Honshu on latitude 41° N, where snow covers the ground for four or five months of the year. Macaques are particularly hardy animals and can withstand cold climates in the wild and captivity extremely well. The limiting factor to the occupation of temperate zones by monkeys is the winter food supply, combined with the short daylight hours available for foraging.

Primates in general are still plentiful in the wild; for the majority of common species there are few serious conservation problems at present. A number of species, particularly the orang-utan, the gorilla, some of the Madagascan lemurs, and some South American species, such as the golden marmoset, however, are in serious danger of extinction unless their habitats can be preserved in perpetuity and predatory habits of man kept under control.

NATURAL HISTORY

Reproduction and life cycle. The stages of the life cycle of primates vary considerably in duration. Among the most primitive members of the group, these stages are broadly comparable to those of other mammals of similar size. Higher in the phylogenetic scale they are substantially extended. The life span of a lemur is about 15 years, of a monkey 25–30 years, of a chimpanzee 40 years, and of man 70–75 years. The greatest differential between man and prosimians, indeed between the prosimians and apes and between apes and man, is in the duration of the infant and juvenile stages combined. The least differences occur in gestation period which, despite the general belief, can not be consistently correlated with adult body size. Gibbons, which weigh considerably less than macaques, have a 20 percent longer gestation period.

The clear trend toward prolongation of the period of juvenile and adolescent life is probably to be associated with the corresponding trend towards a progressive elaboration of the brain. The extended period of adolescence means that the young remain under adult (primarily maternal) surveillance for a long period, during which time the juvenile acquires, by example from its mother and peers, the knowledge that will allow it to become properly integrated as a fully adult member of a complicated social system. One might expect, therefore, a close correlation between the period of adolescence, the brain size, and the

complexity of the social system; and, insofar as the latter factor can be assessed, this appears to be the case.

Breeding periods. The reproductive events in the primate calendar are copulation, gestation, birth, and lactation. Owing to the long duration of the gestation period, these phases occupy the female primate (among higher primates anyway) for a full year or more; then the cycle starts again. The female does not usually come into estrus (physiologically receptive period) until the infant of the previous pregnancy is weaned.

The reproductive patterns of prosimians differ from those of anthropoids. Prosimians show one or more discrete breeding seasons during the year, during which time they may undergo more than one reproductive estrous cycle. The breeding seasons are separated by periods of anestrus, which in galagos are accompanied by changes in the skin of the vulva (the external organ of the genitalia), which closes over, completely sealing the vagina. In the Sudan estrus occurs only twice yearly in the African bush baby (*Galago senegalensis*), during December and August; but in captivity, breeding seasons may occur at any period in the year. In the wild, birth seasons are closely correlated with the prevailing climate, but in captivity under uniformly equable laboratory conditions, this consideration does not apply. In its native Madagascar the ring-tailed lemur (*Lemur catta*) has only a single breeding season during the year; conception occurring in autumn (April) and births taking place in late winter (August and September); but in zoos in the Northern Hemisphere a seasonal inversion of the birth period occurs in which it shifts to late spring and early summer. These examples suffice to indicate the influence of environmental factors on the timing of the birth seasons.

Reproductive cycles in the Anthropoidea continue uninterrupted throughout the year, though seasonality in births is characteristic of primate species living either outside the equatorial belt (5° north and south of the Equator) or at high altitudes in equatorial regions, where dry seasons and seasonal food shortages occur. Seasonality of births in macaques (*Macaca* species) has been documented in Japan, on Cayo Santiago in the Caribbean (where an introduced population thrives under semi-natural conditions), and in India. Observations on langurs in India and Ceylon, on geladas in Ethiopia, and on patas monkeys in Uganda have also demonstrated seasonality in areas with well-marked wet and dry seasons. Anthropoids within the equatorial belt display birth peaks rather than birth seasons. A birth peak is a period of the year in which a high proportion of births, but not by any means all, are concentrated. Equatorial primates such as guenons, colobus monkeys, howlers, gibbons, chimpanzees, and gorillas might be expected to show a pattern of births uniformly distributed throughout the year, but population samples are as yet too small to make this assumption. In captivity, under both laboratory and zoo conditions, anthropoids breed throughout the year with little evidence of seasonality. Even in man there is evidence of high birth peaks. In Europe, the highest birth rates are reached in the first half of the year, and in countries in the Southern Hemisphere in the second half. This may, however, be a cultural rather than an ecological phenomenon, for marriages in certain western countries reach a peak in the closing weeks of the fiscal year, a fact which undoubtedly has some repercussions on the birth period.

Distribution of births through the year

Gestation period and parturition. The period during which the growing fetus is protected in the uterus is characterized by a considerable range of variation among primate species, but shows a general trend towards prolongation as one ascends the evolutionary scale. The range of prosimians overlaps that of the anthropoids, which is never shorter than 140 days. Tree shrews, for example, have a gestation period of 41–50 days, lemurs 132–134 days, macaque monkeys 146–186 days, gibbons 210 days, chimpanzees and gorillas 250–290 days, and man (on the average) 267 days. Even small primates like bushbabies have a gestation period considerably longer than those of nonprimate mammals of equivalent size, reflecting the increased complexity and differentiation of primate structure compared with that of nonprimates. Although in primates there is a general trend towards evolutionary increase in body size, there is no absolute correlation between body size and the duration of the gestation period. Marmosets, for example, are considerably smaller than spider monkeys and howler monkeys but have a slightly longer pregnancy (howler monkeys 139 days; marmosets 140–150 days).

An extraordinary and somewhat inexplicable difference exists between the dimensions of the pelvic cavity and the dimensions of the head of the infant at birth in monkeys and man on the one hand, apes on the other. The head of the infant ape is considerably smaller than the pelvic cavity, so that birth occurs easily and without prolonged labor. When the head of the infant monkey engages in the pelvis the fit is exact, and labour may be a prolonged and difficult affair as it is in man, in whom similar conditions prevail. Human parturition, however, is generally a much more extended process than that of monkeys. Like the human infant, the monkey is born head first. Twin births are as rare in Old World monkeys and apes as they are in man, but certain New World monkeys (*e.g.,* marmosets) habitually produce twins.

Condition of the newborn primate

The degrees of maturation and mother dependency at birth are obviously closely related phenomena. Newborn primate infants are neither as helpless as kittens, puppies, or rats, nor as precocial as the newborn gazelles, horses, and other savanna-living animals. With a few exceptions, primate young are born with their eyes open and are fully furred. Exceptions are the lowly tree shrews and mouse lemurs, which carry their young in their mouths in the manner of many mammals which bear altricial infants. Primate life being peripatetic, it is axiomatic that the infants must be able to cling to the mother's fur. The young of most higher primates have fully prehensile hands and feet at birth and are thus able to grasp the maternal fur without assistance; only man, chimpanzee, and gorilla need to support their newborn infants. In man this stage persists throughout the nursing period (six months or more), but in gorillas it lasts only four weeks and in chimpanzees for an even shorter period.

It seems likely that the difference between the African apes and man in respect of postnatal grasping ability is related to the acquisition in man of bipedal walking. One of the anatomical correlates of the human gait is the loss of the grasping function of big toe, which is aligned in parallel with the remaining digits. Such an arrangement precludes the use of the foot as a grasping extremity. The human infant and, to a lesser degree, the gorilla infant must depend largely on its grasping hands to support itself unaided. The fact that man is habitually bipedal and that, consequently his hands are freed from locomotor chores, may also be a contributory factor; the human mother can move about and at the same time continue to support her infant. Selection for postnatal grasping, therefore, has not had the high survival value that it has in nonhuman primates in which the survival of the infant depends on its ability to hold on tightly. It is a well-known fact, of course, that newborn human infants can support their own weight, for short periods, by means of their grasping hands. Clearly adaptations for survival are not wholly lacking in the human species. Perhaps there are involved cultural factors which have the effect of suppressing natural selection for early infant grasping ability. The first may be that the social evolution of a division of labour between the sexes and a fixed home base has allowed the mother to park her infant with other members of the family as baby-sitters; the second that in more peripatetic communities, the invention of infant-carrying devices, such as the papoose technique of North American Indians, has made it unnecessary for the infant to support itself. Whatever the biological or cultural reasons, the human infant is less precocial than the young of any other primates.

Once the primate infant learns to support itself by standing on its own two (or four) feet, the physical phase of dependency is over; the next phase of psychological dependency lasts much longer. Man, metaphorically, is tied to his mother's apron strings for much longer periods than are the nonhuman primates. The reasons for this are discussed below. According to Adolph Schultz, the anthropologist whose comparative anatomical studies have

illuminated knowledge of nonhuman primates during the last 40 years, the juvenile period of psychological maternal dependency is 2½ years in lemurs, 6 years in monkeys, 7–8 years in apes, and 14 years in man.

Growth and longevity. The characteristic pattern of weight-growth curves of man is also shared by higher nonhuman primates such as chimpanzees (*Pan troglodytes*) and rhesus monkeys (*Macaca mulatta*). There seems little reason to doubt that similar curves will be found to exist in other higher primates not yet studied. It has been suggested that the growth curves of primates are closely akin to the basic curves of mammals in general.

The trend toward the prolongation of postnatal life as a characteristic of primates affects all postnatal life periods including infantile, juvenile, adult, and the period of senescence. The period which shows the least change in apes and man is that of gestation. The relative stability of the gestation period is a clear indication that, developmentally speaking, in the degree of structural complexity and cellular differentiation, there is little difference between apes and man. The differences that undoubtedly exist can probably be attributed to cultural factors. There is no reason to suppose that apes are not built to last equally as long as man; it is simply that the natural hazards of gorilla or chimpanzee life are not buffered by the blessings of civilization, except under conditions of captivity, in which case life spans of 40 years and more are known. The potential life span of the chimpanzee has been estimated at 60 years.

Average and potential life span

The characteristic growth spurts of human infants in weight and height also occur in nonhuman primates but start earlier in the postnatal period and are of shorter duration. Primates differ from most nonprimate mamals by virtue of a delayed puberty in both sexes until growth is nearly complete; in male macaque monkeys, as in man, the growth spurt precedes the onset of puberty. To postpone this latter event is of selective advantage because it ensures that the juvenile males in a primate society are not sexually competitive with adult males until they are large enough and strong enough to contribute fully to the responsibilities of the society in which they live by taking part, for instance, in defense against predators.

Locomotion. Primate locomotion, being an aspect of behaviour which arises out of anatomical structure, shows much of the conservativeness and the opportunism that characterizes the primates. Primates with remarkably few changes in their skeletons and musculature have adopted a bewildering variety of locomotor patterns. The "natural" habitat of primates—in the historical sense—is the canopy of the forest. Although many primates have adopted the ground as their principal foraging area during the day, given the opportunity, they return to the trees to sleep at night. Trees provide cover from the climate and protection from predators; they are of course also a source of food. Only the gelada (*Theropithecus gelada*), the hamadryas baboon (*Papio hamadryas*) of the mountainous regions of Ethiopia, and the chacma baboon (*P. ursinus*), which lives on the rocky coast of the Cape of Good Hope, South Africa, are ground sleepers; even these animals seek the protection of the cliffs and rocky precipices of their habitats at night. No primate sleeps totally unprotected; as a consequence of their relative immunity from predation, primates are heavy sleepers.

The essential arboreality of primates has guaranteed the relative uniformity of the locomotor apparatus. Even man, who has long since abandoned the trees as his principal lodging place, has only partially lost the physical adaptations for tree climbing; his hands remain in the arboreal mold. Only his feet have lost their primitive prehensility in adapting to bipedal walking. Primate locomotion can be classified on behavioral grounds into four major types (see Table 1). Within these major categories there are a number of subtypes, and within these subtypes there are an infinite number of variations between species and, by virtue of individual variability, within species. The differences between the four major categories lie principally in the degree to which the forelimbs and hindlimbs are used to climb, swing, jump, and run.

Locomotor types

Vertical clinging and leaping, for instance, is primarily a function of the hindlimbs, as is bipedalism, whereas brachiation is performed exclusively with the forelimbs. Quadrupedalism involves both forelimbs and hindlimbs, of course, although not to an equal extent. Some quadrupeds are hindlimb dominated; in others, the forelimb and the hindlimb are equally important. The hindlimb-dominated primates, such as the langurs and colobus monkeys, employ a large element of leaping in their movements, a less notable feature of the more generalized quadrupeds such as the species of the genus *Cercopithecus*. The quadrupedal category is inevitably somewhat of a grab bag, and the gaits included in it have not yet been studied critically. One subtype, here designated as slow climbing, differs profoundly from the other subtypes of the category, being somewhat ponderous and devoid of elements of leaping or jumping. The species in this category are all arboreal nocturnal prosimians.

As many authorities who have studied locomotion in free-ranging primate species have pointed out, the classifications of locomotion into categories is a somewhat artificial procedure. A chimpanzee shows a variety of different gaits according to the circumstances of the environment, and brachiation may form only a minor part of its repertoire; this holds true also for the langurs and colobus monkeys, which are placed in the subtype Old World semibrachiation. Although the categories are phrased in behavioral terms, their implications are strictly anatomical. Brachiation is the mode of locomotion for which the animal is specifically adapted; the anatomical correlates of brachiation are quite unmistakable and can be determined equally as well in fossil bones as in living animals. A

Table 1: Locomotor Classification

category	subtype	activity	primate genera
Vertical clinging and leaping		leaping in trees and hopping on the ground	*Avahi, Galago, Hapalemur, Lepilemur, Propithecus, Indri, Tarsius*
Quadrupedalism	slow climbing type	cautious climbing—no leaping or branch running	*Arctocebus, Loris, Nycticebus, Perodicticus*
	branch running and walking type	climbing, springing, branch running and jumping	*Aotus, Cacajao, Callicebus, Callimico, Callithrix, Cebuella, Cebus, Cercopithecus, Cheirogaleus, Chiropotes, Lemur, Leontideus, Phaner, Pithecia, Saguinus, Saimiri, Tupaia*
	ground running and walking type	climbing, ground running	*Macaca, Mandrillus, Papio, Theropithecus, Erythrocebus*
	New World semibrachiation type	arm swinging with use of prehensile tail; little leaping	*Alouatta, Ateles, Brachyteles, Lagothrix*
	Old World semibrachiation type	arm swinging and leaping	*Colobus, Nasalis, Presbytis, Pygathrix, Rhinopithecus, Simias*
Brachiation	true brachiation	gibbon type of brachiation	*Hylobates, Symphalangus*
	modified brachiation	chimpanzee and orangutan type of brachiation	*Gorilla, Pan, Pongo*
Bipedalism		striding	*Homo*

Source: From J.R. and P.H. Napier, *A Handbook of Living Primates.*

purely behavioral classification, based on observations in the living, can have little value in paleontological diagnosis. It is quite arguable that although pure arm swinging is rare in Old World semibrachiators, it is nevertheless one of the principal survival mechanisms of the group, an adaptive advantage that in critical circumstances provides these species with an "edge" over their competitors. In some instances it may well be the case that the particular category is of evolutionary significance only. The gorillas, for instance, whose anatomy is that of a brachiator, in fact, seldom brachiate; adult males probably never do. Gorillas are large animals and bulk-food eaters, forced by their dietary demands to obtain their food on the ground; living gorillas have adapted to a terrestrial life by adopting a compromise gait in which they walk quadrupedally bearing their weight on the backs of their knuckles. This is a form of quadrupedalism (inasmuch as all four limbs are employed) but, once again, the anatomy of their bodies indicates that in past time gorillas (or their evolutionary forebears) were smaller animals which lived and moved in trees. The modern gorilla differs widely from its historical counterpart in behaviour but, structurally, its kinship with its arboreal ancestor is overtly apparent.

The relation of locomotion to habitat changes

Changes in climate and geography during the evolutionary history of primates may also have led to structural atavisms in the anatomy of living primates. Many chimpanzees now living in woodland-savanna conditions in Africa, where the trees are widely spaced and generally unsuitable for the classic arm swinging progress of a brachiator, have adopted a largely ground-living life. Like the gorilla, the chimpanzees are also knuckle walkers, but given an environment like that of a zoo with a cage specially designed with a plentitude of overhead bars and ropes, they brachiate freely and frequently. Habitat changes in the past undoubtedly account for many apparent anomalies in the locomotor classification of present-day primates.

When the subject of primate arboreal locomotion is studied in evolutionary terms, through the medium of fossils, it becomes clear that locomotor categories are not discrete, but that they constitute a continuum of change from a hindlimb-dominated gait to a forelimb-dominated one. The best single indicator of gait, one that has the added advantage of being strictly quantitative, is the intermembral index. Briefly, the index is a ratio expressed as percentage of arm length to leg length; an index over 100 indicates relatively long arms. The intermembral index provides a model by means of which the locomotion of an early primate can be inferred by determination of the intermembral index of the fossil skeleton. Animals do not necessarily fall discretely into categories. Species with indices lying between those of clearly recognizable locomotor types represent transitional types, whose style of locomotion manifests features of both of the bracketing categories. Some lemurs have indices that fall between 65 and 75, but their gait is a combination of vertical clinging and quadrupedalism. The ring-tailed lemur in fact shows an intermediate form of behaviour; it is generally quadrupedal but, given the right environmental milieu, can be observed in vertical clinging and leaping. The South American spider monkeys (*Ateles*), whose index lies between 100 and 108, show a type of locomotion that contains the elements of both quadrupedalism and brachiation.

The intermembral index

Applying the intermembral index to fossil primates, it has become apparent that the earliest primates living in the Eocene Epoch, from about 54,000,000 to 38,000,000 years ago, moved about in the manner of modern vertical clingers and leapers. Quadrupedalism was well-established during the Miocene Epoch (about 26,000,000 to 17,000,000 years ago) when the two major environmental types of quadrupedal gait—the terrestrial and the arboreal—were established, with indices in the region of 85–100 and 75–85, respectively. Brachiation, associated with a high intermembral index, was established as a way of arboreal life at the end of the Miocene and the beginning of the Pliocene.

The origin of bipedalism, which implies the appearance of manlike creatures, is indeterminate. There is direct evidence of bipedalism extending back 1,500,000 to 2,000,000 years, and certain indirect evidence (see below *Evolution and paleontology*) suggests that bipedalism might have evolved in a modified form some 10,000,000 to 14,000,000 years ago.

Locomotion of primates is an area of primate biology of great interest to scholars. It is not only that locomotion is a basic factor of primate life (primates must move to eat and must eat to live) but it is also an indicator of evolutionary progress. To study the origins of primate locomotion is to study primate evolution. Because of its immediacy for survival, natural selection has acted strongly on locomotor behaviour and the locomotor apparatus; changes in locomotor pattern may thus be said to provide principal milestones on the road to modern primates.

Bipedalism, far from being a unique possession of mankind, as is generally thought, is a basic possession of the order Primates. With one or two exceptions (one of which is the tree shrew, which may not even be a primate), all primates sit upright. Many stand upright without supporting their body weight by their arms, and some actually walk upright for short periods. The view that the possession of uprightness is a solely human attribute is untenable; man is merely one species among the 189 that constitute the order who has exploited the potential of his ancestry to his ineffable advantage.

Walking

Chimpanzees, gorillas and gibbons, macaques, spider monkeys, capuchins, and others are all frequent bipedal walkers. To define man taxonomically as "bipedal" is not enough; to describe him as habitually bipedal is nearer the truth, but habit as such does not leave its mark on fossil bones. Some more precise definition is needed. Man's walk has been described as striding, a mode of locomotion defining a special pattern of behaviour and a special morphology. Striding, in a sense, is the quintessence of bipedalism; it is a means of travelling during which the energy output of the body is reduced to a physiological minimum by the smooth, undulating flow of the progression. It is a complex activity involving the joints and muscles of the whole body in its performance and it is likely that the evolution in the human gait took place gradually over a period of 10,000,000 years or so.

The pattern of locomotion of man's ancestors immediately preceding the acquisition of bipedalism has long been a matter of controversy, and the question has not yet been resolved. The evidence derived from anatomical, physiological, and biochemical studies for the close affinity of chimpanzees and gorillas with man would imply that he evolved from a brachiating ancestry. But the many differences between the apes and man, particularly in the structure of the limbs and teeth, have persuaded the majority of leading anthropologists that the theory of a brachiating origin of man is untenable. The issue is still hotly debated, the more so since current immunological studies of serum proteins are tending to emphasize the recentness of the separation of pongid and hominid stocks (the lines that led, repectively, to apes and man) in phylogeny. An alternative to the "brachiating theory" must inevitably be some sort of "quadrupedal theory." Various authorities have proposed other solutions: semibrachiation (a form of quadrupedalism with emphasis on arm swinging), a formal quadrupedal gait, and even a form of locomotion similar to that of *Tarsius* and other clingers and leapers. At the present time, there is insufficient information to elucidate the phylogeny of man's bipedal gait. It can, however, be assumed that it must have involved a large measure of truncal uprightness, a good deal of leaping, and a sizable element of arm swinging. There is no living primate that demonstrates, precisely, this combination of locomotor characteristics, although the arboreal monkeys of the subfamily Colobinae theoretically fit the bill better than most.

Ecology and diet. Few primate species are found today in temperate latitudes. The limiting factors appear to be the general climate and the duration of the daylight hours. The higher the latitude, the greater is the fluctuation between summer and winter temperatures and the shorter are the hours of winter daylight. The combination of quiescent growth periods of fruits, leaves, and grasses, a frozen top soil, and a restricted diurnal period all make

conditions extremely hard for monkeys, which depend on foraging in the trees, on the ground, or under it for their food supply. No monkey lives, or has ever lived, in the zones of permafrost, but it is probable that in times past they flourished in temperate regions. Without the harassment of man, monkeys no doubt could have survived in the temperate zones; their foraging, albeit condensed into a temperate-zone day of seven hours or so, would have supplied them with the food they required, providing they were free to forage without the threat of human persecution. Dietary adaptability of primates is greater than realized. On the Shimokita Peninsula of the Japanese Island of Honshu and in the Shiga Mountains of the same island, Japanese macaques survive from late November until April on incredibly meagre rations. The diet of these monkeys, during the months when snow covers the ground and when the last remaining summer seeds and leaves have vanished, is bark and the fleshy cambium layer that lies beneath it. The Japanese macaques, cold and hungry, ride out the worst that winter, at the latitude of 41° N, offers. Near Moscow, several species of African primates have been kept for several years in outdoor facilities in which winter temperatures go well below −18° C (0° F). It is true that they are well provisioned but, nevertheless, it is almost incredible that a tropical species should be sufficiently adaptable and possess adequately effective homeostatic mechanisms to survive in this degree of climatic adversity. Ruggedness of this sort is a primate characteristic, a survival mechanism that operates in the worst conditions, and one which provides a clue to the real reason for the primacy of the primates among mammals.

Resistance to cold

Diet. The diet of primates is a factor of their ecology that, during their evolution, has clearly played an important role in their dispersion and adaptive radiation as well as in the development of the teeth, jaws, and digestive system. Diet is also closely related to locomotor pattern and to body size.

The principal food substances taken by primates may be divided into vegetable (fruits, flowers, leaves, nuts, barks, pith, seeds, grasses, stems, roots, and tubers) and animal (birds, birds' eggs, lizards, small rodents and bats, insects, frogs, and crustacea). The flesh of larger mammals (including primates) is not listed as an important item of nonhuman primate diet, although it is taken by certain species such as baboons and chimpanzees in special circumstances that are not yet fully understood.

While in many mammalian groups, diet is selective and specific to the order, among primates it is difficult to establish any hard and fast rules. Although there are decided preferences for certain food items, catholicity is more characteristic than specificity. Generally speaking, primates are omnivorous, as the physiology of their digestive system attests. Relatively few examples of dietary specialization are to be found. The so-called leaf-eating monkeys, a sobriquet that embraces the whole of the subfamily Colobinae, are by no means exclusively leaf eaters and according to season include flowers, fruit, and seeds in their diet. The howler monkeys of the New World have a similar dietary preference.

Broadly, however, certain overall dietary preferences are discernible. The leaf-eating langurs have already been mentioned. The apes (other than the gorilla) are substantially fruit eaters; the percentages of fruit in the diet of the gibbon and the three great apes are as follows: gibbon 62 percent, orang 50 percent, chimpanzee 67 percent, and gorilla 2–15 percent.

Many of the smaller nocturnal prosimians such as tarsiers, galagos, dwarf lemurs, sportive lemurs, the aye-aye, slender lorises, and the South American night monkey are substantially insectivorous.

The larger diurnal prosimians (*e.g.,* typical lemurs, the sifaka, and the indri) and the large nocturnal prosimians (the Asian slow loris and the African potto) are more vegetarian, including both fruit and leaves in their menu. It seems apparent that size, rather than activity rhythm, governs the nature of the primate diet. The small marmosets of the South American genus *Callithrix* have exclusively diurnal rhythms and are insectivorous, while the slightly larger, but equally diurnal, tamarins (*Saguinus*) are more omnivorous.

The most widespread monkey populations are those of the subfamily Cercopithecinae. Many ground-living species are included in this group, the members of which are notable for the extreme catholicity of their diet; meat eating as a regular pattern, however, has not been reported among them.

Size in relation to food habits. In evolutionary terms, increase in size has probably played a large part in determining the direction of primate evolution. Fossil evidence of early primates of the Paleocene (about 65,000,000 years ago) suggests that these were small, forest-living creatures, not yet specialized as tree climbers. Their bodily form suggests that they possessed the build and the gait of squirrels. Their molar teeth bore high, pointed cusps but were neither as tall nor as pointed as those of the insectivore-like ancestors of the preceding Cretaceous epoch, whose molars were ideally adapted for cracking the hard, chitinous exoskeletons of insects. This fact suggests that the reduction of the molar cusps was associated with the adoption of a fruit-eating habit. Although this has some validity as a generalization, it should not be taken too literally. All primates include insects in their diet and there are many almost exclusively insectivorous forms, which have reduced the height and acuity of their molar cusps. Increasing body size, a trend which is clearly apparent throughout primate evolution, would have been associated with the adoption of supplementary sources of food. Insect eating is a laborious way for a large animal to obtain nourishment unless it happens to possess the extensible tongue of those mammalian "vacuum cleaners," the anteaters.

The shift from insect to fruit eating

An increase in size and the gradual addition of bulk foods to the diet would, in turn, have affected the habitat and the pattern of locomotion of primates. Fruit eating is more easily accomplished in trees than on the ground and this growing preference amongst primates may have been accompanied by the tree-climbing habit. Once such a habit was established, natural selection would have operated to espouse the suitable physical adaptations which would facilitate such a behaviour. Suitable adaptations in this case would have been the facility to climb, leap, and balance in trees.

It is noteworthy that during evolution the development of a prehensile foot preceded that of a prehensile hand. Vertical clinging primates such as the tarsiers or small squirrel-like quadrupeds like the marmosets and tamarins—all of which have prehensile feet, but not prehensile hands—by remaining small, have never been subject to the same evo-

From J.R. Napier and P.H. Napier, *A Handbook of Living Primates* (1967); Academic Press

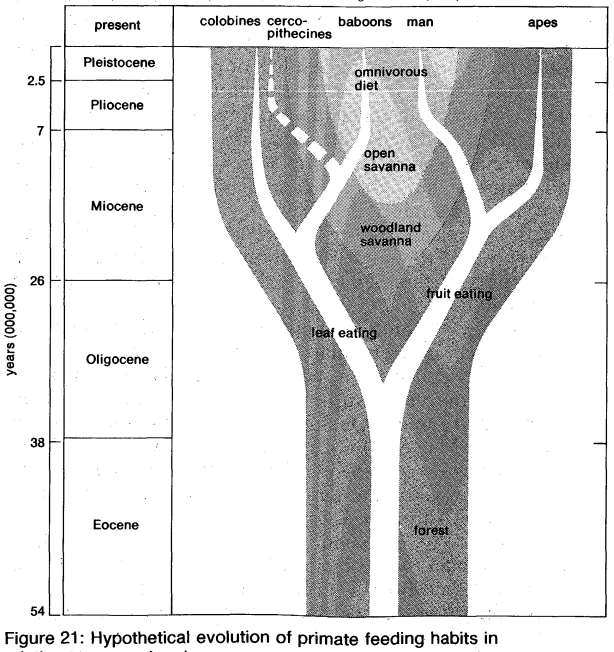


Figure 21: Hypothetical evolution of primate feeding habits in relation to vegetational areas.

lutionary pressures that have impinged on larger primates. A large arboreal primate without prehensile hands is at a considerable disadvantage in moving about in the canopy of trees, but a small one suffers little disadvantage. Amid the large and firm branches, size is no particular hazard, but at the periphery of the crown, where the fruit is most abundant and the branches are slender and flexible, the risk of falling is increased. It is, therefore, likely that the combination of an increase in body size, associated with the inevitable shift towards a bulk diet, led first to the evolution of a grasping hand, then to the appearance of a prehensile hand, and finally to an opposable thumb. Four prehensile extremities are more effective than two in defying gravity.

Such adaptations of the forelimbs would have had the effect of equalizing the role of the limbs. The limbs of vertical clingers are functionally disparate, the lower pair being dominantly propulsive, the upper, secondary and purely supportive. The limbs of quadrupeds, on the other hand, are more homogeneous, both pairs having a propulsive function during running. Thus, it would seem that the transition in locomotor grade between vertical clinging and leaping and quadrupedalism came about as adaptation to increased body size. Size, diet, ecology, locomotion, and anatomical structure provide a constellation of causes and effects that are critical factors in the evolution of the primates.

Effect of forest stratification. An aspect of primate ecology that has not yet been fully investigated is the effect of the stratification of forests on the adaptations (particularly the locomotor adaptations) and on the speciation of primates. In recent years it has become apparent that forests are not simply a random consociation of trees but are an ordered system of plants under the control of climatic, photobiotic, and edaphic (soil) factors.

Forest structure and traffic ways

The chief physiognomic features of rain forests, the ancestral home of primates and the principal habitat (in numerical terms) of nonhuman primates today, are the evergreen broad-leaved trees that, collectively, form a closed canopy, so opaque to sunlight that the forest floor is in perpetual twilight. Epiphytes and thick-stemmed lianas drape the trees, linking one crown to another and providing aerial pathways for monkeys to pass from tree to tree through a continuum of interlacing branches, a three-dimensional maze that provides home, restaurant, shopping districts, and highways for primates. Three strata of the canopy of rain forests are usually recognized by botanists: an understory, a middle story, and an upper story. The understory is often "closed," the crowns of the constituent trees overlapping one another to form a dense, continuous, horizontal layer. The middle story is characterized by trees that are in lateral contact but do not overlap; and the highest story by tall trees, some 50 metres (160 feet) or more, which form a discontinuous layer of umbrella-shaped crowns. The occasional "emergent" forest giant may tower above the highest layer of the canopy. There is some evidence, much of it conflicting, that zonation of forest primates occurs within the forest canopy. The stratification of forest is extremely variable and the number of layers tends to diminish from three to two in secondary forest, dry deciduous forest, and montane forest; and from two to one as temperate zone, tropical woodland, or montane woodland supervenes.

Tropical grasslands, or savannas, are also the homes of primates in Africa and Asia; no savanna-living primates exist in South America. Tropical grasslands comprise a mixture of trees and grasses. The proportion of trees to grass varies directly with the rainfall; areas of high seasonal rainfall support single story woodlands of tall trees while lush grasses form the ground vegetation; but where rainfall is both seasonal and low, the trees consist of stubby xerophilous (dry-loving) shrubs and short, tussocky grasses. The principal primate fauna of the savanna biome are the ground-living species; in Africa, the grivet *(Cercopithecus aethiops),* baboons (*Papio* species), and the patas monkey (*Erythrocebus patas*); and in Asia the macaques (*Macaca* species) and the Hanumān langur (*Presbytis entellus*).

Tropical montane forests or tropical rain forests at high altitude also abound in primates in Africa, Asia, and South America. In equatorial Africa certain primate species have colonized the montane-savanna regions, or moorlands, where the rugged mountainous terrain and seasonal food scarcity support herds of geladas and hamadryas baboons. These high mountaineers of Africa have no ecological counterparts in Asia or South America.

As a measure of the physiological adaptability of primates it may be noted that various populations have at different times been subjected to relocation in alien environments by man. Some of these expatriates are the green monkeys (*Cercopithecus sabaeus*) of St. Kitts in the West Indies, the mona monkeys (*C. mona*) of Grenada, the macaques on Mauritius, and the Barbary apes (*Macaca sylvana*) of Gibraltar.

FORM AND FUNCTION

Distinguishing features. The basis of the success of the order Primates is the relatively unspecialized nature of their structure and the highly specialized plasticity of their behaviour. This combination has permitted the primates throughout their evolutionary history to exploit the wide variety of novel ecological opportunities that have come their way. Although there are a few highly specialized species among the lower primates (the aye-aye, the tarsier, the potto, and the lorises, for instance), the higher primates, the anthropoids, are extremely conservative in their structure; morphologically speaking, they have maintained a position in the evolutionary midstream and have avoided the potential stagnation of specialized life near the bank. Specialization is not always a liability; in times of environmental stability the specialized animal enjoys many advantages, but in a rapidly changing world it is the less specialized animals that are more likely to survive and flourish.

The importance of a large brain

The plasticity of primate behaviour is largely a function of the brain. The primate brain is distinguished by its relatively large size compared with the size of the body as a whole; it is also notable for the complexity and elaboration of the cerebral cortex, the function of which is to receive, analyze, and synthesize the incoming impulses from the sense organs and to convert them into appropriate motor actions, which in turn constitute behaviour.

It has been observed that there is no distinguishing feature which characterizes all primates except a negative one—their lack of specialization. St. George Mivart defined the order Primates in the following terms:

> . . . an unguiculate, claviculate, placental mammal with orbits encircled by bone; three kinds of teeth at least at one time of life; brain always with a posterior lobe and a calcarine fissure; the innermost digits of at least one pair of extremities opposable; hallux with a flat nail or none; a well-developed caecum; penis pendulous; testes scrotal; always two pectoral mammae.

When these criteria are examined one by one the inevitable conclusion is reached that, singly, none of these characters is unique to primates nor do all primates demonstrate every single character. Collectively, however, more primates than nonprimates show this particular constellation of characters. Perhaps a more meaningful way to look at primate characteristics is in the form of evolutionary trends. This approach is quantitative, inasmuch as it is recognized that not all adaptations are developed in primates to the same degree. A list of evolutionary trends with behavioral connotations is also a somewhat more functional inventory than Mivart's essentially morphological one and, furthermore, is more applicable to the study of fossil primates.

Binocular vision

Primates are essentially arboreal animals whose limbs are adapted for climbing, leaping, and running in trees. Active arboreal life requires the mechanical assistance of a long tail and sensitive, grasping hands and feet with opposable thumbs and big toes, to aid in climbing and to ensure stability on slender branches high above the ground. Active arboreal locomotion also requires a much more accurate judgment of distances than life on the ground; this is facilitated by the development of stereoscopic vision, the anatomical basis of visual judgments in depth. The forward-facing eyes of primates are adaptations for this type of visual precision. A highly developed sense of smell is not nearly as important for animals leading

an arboreal life as it is for those on the ground. Primates thus have a much reduced olfactory mechanism; noses are shorter, the turbinal (or scroll) bones of the nose are reduced in number and complexity compared with most nonprimate mammals.

Primates, being on the whole omnivorous animals, have evolved a multipurpose dentition; incisors to cut, canines to pierce and tear, and molars and premolars to crush and grind. Such a dental arrangement equips its possessor with the power to cope with many different sorts of food—insect food, vegetable food, or flesh—and provides a major contribution for the generalized way of life that is so characteristic of the order.

A basic adaptation for the primate way of life (including the unique specialization of man—upright walking) is verticality of the trunk. Above the level of the tree shrews, the ability to hold the body upright is a universal primate possession; indeed paleontological evidence indicates that truncal uprightness was a primate characteristic as far back as the early Eocene, some 54,000,000 years ago.

Above all, the principal evolutionary trend of primates has been the elaboration of the brain, particularly of that portion of the cerebral hemispheres known as the neopallium or neocortex. A neocortex is characteristic of higher vertebrates, such as mammals, which operate under the control of multiple sources of sensory input. In many mammals the olfactory system dominates the senses and the cerebral hemispheres consist largely of paleocortex—the "smell brain"—of lower vertebrates. The arboreal habit of primates has led to a dethronement of the olfactory sense and the accession of a tactile, visually dominant sensory system. This evolutionary trend has resulted in the dramatic expansion and differentiation of the neocortex.

General structure. *Vertebral column and posture.* All primates retain a clavicle or collarbone (lost in many mammalian groups), a separate radius and ulna in the forearm, and a separate tibia and fibula in the lower leg. The single exception to this is the tarsier, in which the fibula becomes fused to the tibia.

The primate vertebral column shows a basic mammalian pattern of components including an "anticlinal" vertebra, situated in the mid-thoracic (upper-back) region of the spinal column and marking the transition between the forelimb and hindlimb segments. In a galloping greyhound the anticlinal vertebra is at the apex of the acute curve of the back. An anticlinal vertebra is characteristic of all quadrupeds and is seen in all primate families except the hominids and the pongids, whose posture is upright or semi-upright. The evolutionary trend in the vertebral column is toward shortening the lumbar, sacral, and caudal regions. Extreme shortening in the lumbar region, with complete loss of the caudal (tail) vertebrae, occurs in gibbons, the great apes, and man. Prehensility of the tail is a specialization of certain New World monkeys, but appears also for a brief period in the infants of many Old World monkey groups, in which it provides an important mechanical aid to survival.

Hands and feet. With two exceptions, all primates have retained five digits on hand and foot. The exceptions are the spider monkeys of South America and the colobus monkeys of Africa, which have lost or reduced the thumb. This appears to be an adaptation for locomotion, the rationale for which is not fully understood at present.

All (except tree shrews) possess prehensile (grasping) hands, and (except man) prehensile feet. The hands of Anthropoidea show a greater range of precise manipulative activity than those of Prosimii. Prosimians lack the functional duality of the hands of anthropoids. Duality in hand function has been described for man and nonhuman primates in terms of precision and power grips. The power grip of prosimians is very well developed but the precision grip is lacking. Among Anthropoidea, the New World monkeys show a considerable advance over prosimians in tactile sensitivity but they possess less functionally effective hands in prehensive terms than Old World monkeys. The critical component of the prehensile hand in terms of skilled manipulation is the opposable thumb: a thumb, that is to say, that is capable of being moved freely and independently. The movement of opposition is a rotary movement in which the thumb, swinging about its own axis, comes to face the under or palmar surface of the tips of the fingers. The opposable thumb is the basis of the precision grip which, though present to some extent in all primates, is particularly highly developed in man. Opposability is present to some degree or other in most primates but varies considerably in its functional effectiveness as an instrument of fine manipulation. The baboons and man are pre-eminent in this respect. The apes, having short thumbs and long fingers, are handicapped in relation to delicate manual dexterity but are adept in the coarser elements of hand use, particularly in relation to tree climbing and branch swinging.

Power versus precision grips

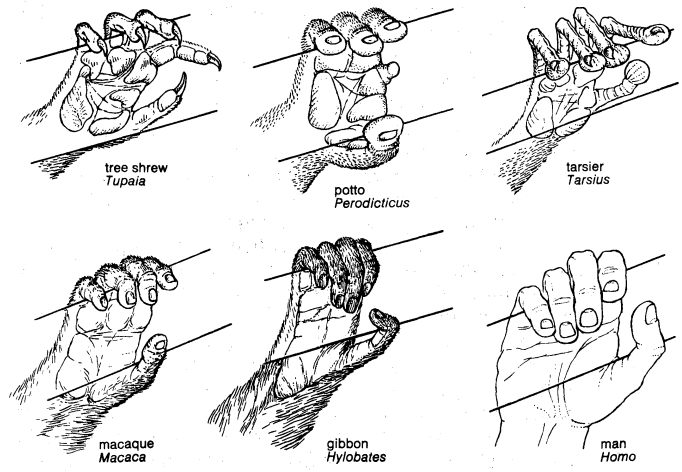


Figure 22: Shape and function of the hand in certain primates.
Drawing by R. Keane

Teeth. A dentition with different kinds of teeth (heterodonty)—incisors, canines, and cheek teeth—is characteristic of all primates and indeed of mammals generally. Heterodonty is a primitive characteristic and primates have evolved less far from the original pattern than most mammals. The principal changes are a reduction in the number of teeth and an elaboration of the cusp pattern of the molars.

The primitive mammalian dental formula is assumed to have been $\frac{3 \cdot 1 \cdot 4 \cdot 3}{3 \cdot 1 \cdot 4 \cdot 3} = 44$ teeth (the numbers being the numbers respectively of pairs of incisors, canines, premolars, and molars in the upper and lower jaws). No living primate has retained more than two incisors in the upper jaw and only in tree shrews is the primitive number of three incisors retained in the lower jaw. The incisors are subject to considerable variation in prosimians. Characteristically, the upper incisors are peglike, one or other pair often being absent; in the lower jaw the incisors show a peculiar conformation which has been likened structurally and functionally to a comb. The dental "comb" of prosimians is composed of the lower canines and lower incisors compressed from side to side and slanted forwards; the dental comb appears to be primarily a device for grooming the fur. Canines are present throughout the order, but show remarkable variation in size, shape, projection, and function. Characteristically the teeth of cercopithecines have a function in the maintenance of social order within the group as well as an overtly offensive role; their function as organs of digestion is relatively unimportant. They are large and subject to sexual dimorphism, being larger in males than females. Man's canines are small and are almost wholly dental in function, with no size differences between the sexes.

Evolutionary reduction of the cheek teeth

The trend in the evolution of the cheek teeth has been to increase the number of cusps and reduce the number of teeth. Both molars and premolars show this tendency. No living primate has four premolars; prosimians and New World monkeys have retained three on each side of each jaw, but at the anthropoid level two premolars are the invariable rule. The primitive premolars are uniform in shape and are unicuspid, but in primates the most posterior premolar tends to evolve either one or two extra

cusps (molarization), an adaptation providing an extended tooth row for a herbivorous diet. In species with large upper canines, the most anterior lower premolar assumes a peculiar shape known as sectorial, functioning as a hone for the scythelike canine. In man, whose canines are small and unremarkable, the first and second premolars are identical in shape and two-cusped.

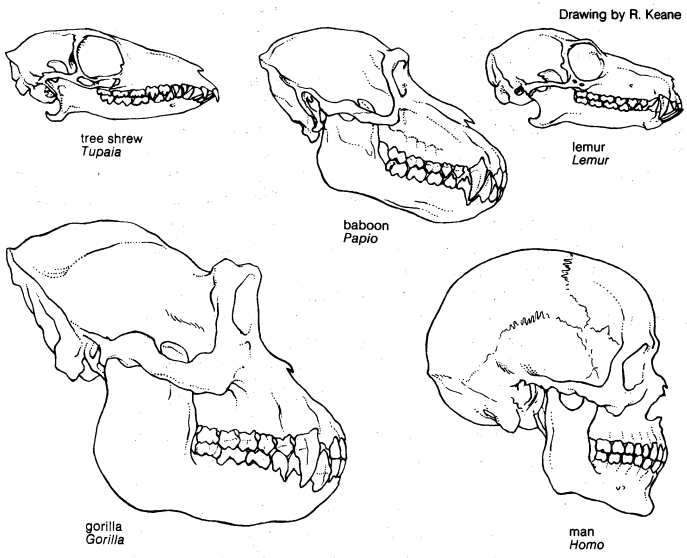


Figure 23: Skulls and teeth of primates.

The trend in the morphology of the molars has been to increase the primitive three cusps to four or five. The apes and man are more specialized in this feature than are the monkeys or the lower primates, having four cusps on the molar crown in the upper jaw and five cusps on the lower. A further tendency has been to reduce the molar series from three to two in both jaws, principally in the lower.

Snouts, muzzles, and noses. The reduction of the snout in primates is a correlate of the diminution of the olfactory sense (see below *Sensory reception*). To a great extent visual acuity and manual dexterity have replaced the sensitive, inquiring nose found in so many nonprimate mammals. A marked reduction in the complexity of the "scroll" bones of the nose, the richness of the innervation of the olfactory mucous membrane, and the sensitivity of the moist tip of the nose—the rhinarium—is associated with the reduction in length of the primate snout. Although the trend in primate evolution is towards a dethronement of the primacy of the sense of smell, there are still some good snouts to be seen in those lower primates, that retain a naked moist rhinarium attached to the upper lip.

Prosimians depend for many aspects of their social and reproductive behaviour on olfactory signals which are effected by means of special scent glands distributed in different regions of the body but congregated principally in the anal and perineal regions. Marking behaviour, the placing of scent at various points in the environment, is a prominent feature of the prosimian repertoire of communication. Marking behaviour ceases to be of much importance above the prosimian level, and in anthropoids the structures by which olfactory signals are given and received are diminished. Monkeys, apes, and man, though they may be lower down the scale of olfactory excellence than prosimians, rely in considerable measure on the sense of smell; all higher primates, for instance, including man, sniff at unfamiliar items of food before placing them in the mouth.

The platyrrhine–catarrhine division

The shape of the nose of higher primates is one of the most reliable means of distinguishing Old World monkeys from New World monkeys at a glance. In New World monkeys or Platyrrhines (platyrrhine means "flat nosed") the nose is broad and nostrils are set wide apart, well separated by a broad septum, and point sideways; in Catarrhines (catarrhine: "downward nosed") the nostrils are set close together, point forwards or downwards, and are separated by a very narrow septum.

Some Old World monkeys, particularly those that have adopted a ground-living way of life, such as baboons, mangabeys, and the mandrills of the subfamily Cercopithecinae, appear to have re-adopted a long snout during their evolution. This structure, however, is not primarily olfactory in function but seems, rather, to be more closely related to the large size of the jaws and the prominence of the canine teeth; it should be considered a dental muzzle rather than an olfactory one.

Sensory reception and the brain. Among mammals in general, the olfactory system is the primary receptor for environmental information; consequently the brain of most mammals is dominated by the olfactory centres. In primates the sense of smell is considerably less important than the well-developed visual system and highly refined sense of touch. The primate brain is enlarged in the specific areas concerned with vision (occipital lobes) and touch (parietal lobes), thus taking a characteristic shape throughout the higher primates.

Touch. The skin of the primate hand is well adapted for tactile discrimination. Meissner's corpuscles, the principal receptors for touch in hairless skin, are best developed in the apes and man, but can be found in all primates except the tree shrew; structurally correlated with a high level of tactile sensitivity are certain anatomical features of the skin of the hands and feet, such as the absence of pads on the palms and soles, characteristic of mammals in general, and the presence of a finely ridged pattern of skin corrugations known as dermatoglyphics (the basis for fingerprints).

Eyes and vision. The evolutionary trend towards frontality of the eyes has not proceeded as far in the prosimian as in the anthropoid families. Tree shrews retain set of the eyes like that of the primitive ancestor, the central axis of each bony orbit being 140° apart. In lemurs this angle is considerably less, 60°–70°, and in the anthropoid families of apes and monkeys the divergence has been reduced to 20°. It should be noted that the axes of the eyeballs (as distinct from the bony orbits) in apes and monkeys are, in fact, parallel.

Colour vision

Colour vision is of considerable advantage to arboreal animals living on fruits and insects. Species that have both rod and cone receptors in their retinas include all of the Anthropoidea with the exception of some of the nocturnal species (*Aotus,* the night monkey of South America, for example), and the family Lorisidae and some of the Lemuridae among the prosimians. Rods are the more primitive retinal cells and respond to low intensity light, while cones, which respond to high light intensity, have a greater resolving power, and thus provide for a finer type of visual discrimination. Cones, moreover, are adapted for colour discrimination.

Nervous integration. The elaboration of touch and vision supplements the senses of smell, hearing, and taste, providing the primate with a sensory armament of great range and flexibility. The primate central nervous system is sufficiently refined to deal with the elaborate bombardment of environmental information reaching it. Association areas provide the connections between the input and output centers of the brain—the motor and sensory cortex. Association areas are the memory banks where the memory of past experience is encoded in the infinitely complicated plexiform arrangement of the neurons, the

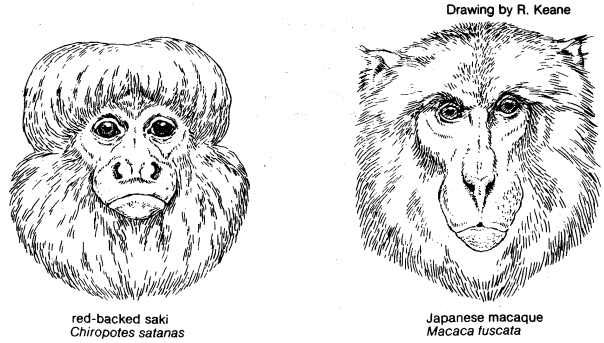


Figure 24: Head of New World monkey (saki) compared with head of Old World monkey (macaque).

brain cells, and their processes. All sensory impulses reaching the cortical centers of the central nervous system are routed through the association areas for conditioning, as it were, before reaching the effector side or output side where the appropriate response is initiated in the cells of the motor cortex. The more highly developed the association areas of the brain are, the more specific and appropriate is the behaviour and the more versatile is the animal in facing environmental demands.

The brain. The principal evolutionary trend in brain development has been towards elaboration. The neocortex of higher primates possesses highly developed associative functions, an aptitude for receiving, analyzing, and synthesizing the sensory input from visual, olfactory, auditory, gustatory, and tactile receptors and converting them into the appropriate motor responses.

Relative versus absolute brain size

The brain of Anthropoidea is larger, both absolutely and relatively, than that of prosimians. For instance, the weight of the simplest anthropoid brain, that of a marmoset, is three times greater than the brain weight of the bush baby, a prosimian of comparative size. This quantitative increase, in part, is attributable to the elaboration of the regions of the neocortex concerned with tactile and visual sensitivity; and in part to the elaboration of the intrinsic pathways connecting one part of the brain with another. The large brain of man is attributable not so much to an increased nerve cell content as to an increase in the size of the nerve cells and to a greater complexity of the connections linking one cell to another.

The external form of the anthropoid cerebral cortex is characterized by a complicated pattern of folds and fissures (sulci and gyri) in the brain surface. The fissural pattern is seen in its simplest form in the marmosets, but in the larger and more advanced members of the New World monkey infraorder, the capuchins (*Cebus*) for instance, the cerebrum is richly convoluted. Gyri and sulci are well marked in Old World monkeys and in the apes, the complexity of the pattern closely approximating the tortuous mazelike pattern seen in man.

Reproductive system. *Male and female genitalia.* The functions of the individual organs of reproductive systems are fairly uniform throughout the primates, but in spite of this physiological homology there is a remarkable degree of variation in minor detail of organs between groups, particularly in the external genitalia, which, by their variation, provide a morphological basis for the reproductive isolation of the species. There could be no more effective barrier to mating between different species than incompatability of the male and female sex organs.

Among the characteristics of the primate order as listed by Mivart, the penis is described as "pendulous" and the testes as "scrotal." In contrast to other mammals, the primate penis is not attached to the abdominal wall but hangs free. The testes, with a few exceptions among the prosimians, in which they are withdrawn seasonally, lie permanently in the scrotal sac. In all primates except modern man (there is no evidence one way or the other regarding early representatives of the genus *Homo*), tarsiers, and some South American monkeys, the penis contains a small bone called the baculum, a typically mammalian character. The uterus of female primates shows all grades of transition between the two-horned (bicornuate) uterus, typical of most mammals, to the single-chambered (simplex) uterus of the higher primates and man.

Variations between primate taxa are demonstrated most strikingly by the glans penis, the scrotum, and perineum of the male, and the clitoris and the labial folds of the female vulva. The length and form of the clitoris, which when elongated mimics the penis, as in spider monkeys for instance, is a potent source of confusion in determining the sex of certain New World primates. The coloration of the male scrotum in forest-living primates, particularly of the genera *Cercopithecus* and *Mandrillus,* shows an infinite range of variations and provides a species-recognition signal of considerable effectiveness.

The external appearance of the genitalia undergoes seasonal variation in a number of primates. In the male, swellings of the testis and colour changes of the scrotum occur, and, in the female, swelling and coloration of the vulva and perineal region herald ovulation, sometimes most obtrusively. Turgidity and excessive vascularity of the tissues of the perineum are probably characteristic of all mammals, but there are certain primate species in which this engorgement reaches monstrous proportions, notably baboons, mangabeys, some macaque species, and chimpanzees. Regions other than the primarily genital areas may also be affected by the hormones circulating at certain periods of the reproductive cycle. For instance, in the gelada baboon, the skin on the front of the female chest, which normally bears a string of caruncles resembling the beads of a necklace, becomes engorged and brightly coloured. A German zoologist, Wolfgang Wickler, has suggested that this is a form of sexual mimicry, the chest mimicking the perineal region. The observation that geladas spend many hours a day feeding in a sitting posture provides a feasible, Darwinian explanation of this curious physiological adaptation.

Mimicry of bottom by chest colours

Placenta. The placenta, a characteristic of all Eutherian mammals, is a vascular structure that permits physiological interchange of blood and body fluids between the mother and the fetus and the breakdown products of the fetal metabolism; it also provides a two-way barrier preventing the passage of some, but not all, noxious substances and organisms such as bacteria and viruses from one individual to the other and is the source of hormones such as estrogens.

The placenta is a flat discoid-shaped "cake" in man, some of the other higher primates, and the tarsier. In many monkeys it is bidiscoidal, having two linked portions. The placenta is intimately attached on its outer surface to the endometrium, the lining of the uterus, by fingerlike processes (villi) that embed themselves in the endometrium where the vascular connections between the two circulations are achieved. The connection between fetal and maternal circulations appears as two distinct types among primates, a distinction which is believed to have had an important effect on the evolution of the order. In the first type (epitheliochorial), found in the prosimian suborder, several cellular layers separate the maternal and fetal bloodstreams, thus limiting the passage of molecules of serum proteins. In the second type, hemochorial placenta, the relationship is much more intimate, there being no fetal and maternal cellular layers separating the two circulations, and serum proteins can easily pass. A hemochorial type of placentation is found in tarsiers, apes, monkeys, and man.

The placenta is shed at birth in all primates, and, in all other than civilized man, is eaten by the mother.

EVOLUTION AND PALEONTOLOGY

Renewed interest in primate origins. Beginning in the 1950s there was a notable expression of interest in primate paleontology. Since then, hardly a year has passed without the announcement of some new major discovery. New sites have been opened up and old discoveries redescribed and reallocated. New techniques in geological dating, in palynology (the study of fossil pollen), in paleoclimatology and paleoecology and in the archeological interpretation of fossil sites, have helped to lift primate paleontology into the forefront of the life sciences and have aroused public interest to an unprecedented level. The popularity of all aspects of the evolution of man is reflected, for instance, in the spate of books covering this field published during this period.

Important recent sites of primate fossils

The African continent has contributed the greatest share of significant early finds—Rusinga and Songhor in Kenya, Olduvai in Tanzania, and Sterkfontein, Kromdraai, Swartkrans, Makapan in South Africa are names with which every anthropology student is familiar. Later, the names Omo, Kanapoi, Lothagam, Napak, Moroto, Laetoli, and Hadar—all East African sites of hominid and pongid discoveries in the 1970s and '80s—became equally well known.

Elsewhere, pieces of this colossal worldwide jigsaw have been discovered in Europe, notably in Hungary, France, and Italy; in the Siwālik Hills of northwest India; in the ever prolific middle Eocene Bridger Beds of North Amer-

ica; and in Colombia and Bolivia. These latter discoveries were particularly welcome because of the poverty of fossils of New World monkeys.

While new discoveries have clarified the human story, older ones, that had only served to cloud it, have been repudiated. Piltdown man was shown unequivocally to be a fake in 1953; *Oreopithecus* of Monte Bamboli, Italy, has been relegated to a family remote from human ancestry, and Galley Hill Man in England has been shown to be a recent intrusion (a burial) into middle Pleistocene gravels. The questionable finds from the remoter geological period of the Eocene and Oligocene have also been reexamined, with the result that a number of confusing fossils have been dismissed. *Anagale,* long thought to be a fossil tree shrew, is now regarded as a small carnivore; others, such as the supposed tarsioid-anthropoid annectent form, *Parapithecus fraasi,* have been reallocated within the order Primates.

Progress in constructing the phylogeny of higher primates and man has been bedevilled by a number of controversies concerning taxonomy and nomenclature, and, as a result, internal disagreements have developed which mitigate against the sort of international cooperation that is essential for so vast a project as the construction of a human phylogeny. New-school and old-school taxonomists have come into conflict. But with the rapid advances in genetics, in the new concepts of biological species and in population anthropology, a fresh equilibrium is slowly being acquired, as the pendulum swings between the reactionary "lumping" of such taxa as genera and the traditional "splitting," in which every new discovery was provided with a new generic name. The most extreme example of taxonomic lumping in primate paleontology was in a 1965 revision by E.L. Simons and D.R. Pilbeam of the Miocene–Pliocene genus *Dryopithecus;* from the 28 genera recognized prior to 1965, the authors recommended the recognition of three genera only: *Dryopithecus, Ramapithecus,* and *Gigantopithecus.* All the rest were absorbed into the three genera, a result initially satisfying but not yet subjected to the test of time.

The primate fossil record. *Cretaceous.* The known range of the primate order was extended back from the middle Paleocene to the Late Cretaceous (about 75,000,000 years ago) by the discovery in Montana of one premolar and four molar teeth, representing two species of insectivore-like primates that were assigned in 1965 to a new genus, *Purgatorius.* This diagnosis is based on the characters of four molar and one premolar tooth and is not by any means universally accepted.

Paleocene. The best known early primate species, of which, *inter alia,* a complete skull and a partial postcranial skeleton are available, belongs to the genus *Plesiadapis* from the Paleocene (about 60,000,000 years ago) of Cernay, France, and Colorado. The skull shows a number of dental specializations including procumbent rodentlike incisors in the upper and lower jaw and the absence of antemolar teeth, but the molar teeth show the unequivocal primate affinities of this species. Rodent affinities have been suggested for *Plesiadapis,* but most authorities agree that it is a primate, however aberrant.

Eocene. The known Eocene fossil families include the Tarsiidae (tarsiers), the Adapidae, considered probable ancestor of lemurs and lorises, and the Omomyidae, possible ancestors of the monkeys and apes.

Possible ancestors of the lemurs

The family Adapidae contains two North American genera, *Notharctus* and *Smilodectes,* which are well represented in the fossil deposits of the Bridger Basin, Wyoming. *Adapis,* a closely related genus, was widely distributed in Europe, as were *Anchomomys* and *Pronycticebus. Notharctus* and *Smilodectes* are not thought to be antecedent to living lemurs, principally on zoogeographical grounds, since egress from the North American continent for tropical species was effectively cut off by ecological changes soon after the beginning of the Eocene (about 54,000,000 years ago). *Notharctus* was not unlike the modern lemurs in size and general appearance. On both morphological and zoogeographical grounds, *Adapis* and *Pronycticebus* or *Anchomomys* could have provided the stem from which the living lemurs and lorises evolved.



Figure 25: Eocene fossil lemur *Notharctus.*
Drawing by R. Keane

Representatives of the Omomyidae have been found in North America, in Europe, and in Asia. According to American paleontologist Elwyn L. Simons, the Omomyidae are probably the key family for tracing the phylogeny of the higher primates, the apes and monkeys. The family was common to North America and Eurasia, a fact which reinforces the presently held view that omomyids provided the stock from which New World monkeys and Old World monkeys, apes, and man arose. Their principal claim to antecedency (other than the circumstances of their zoogeography) is the dental formula: $\frac{2 \cdot 1 \cdot 3 \cdot 3}{2 \cdot 1 \cdot 3 \cdot 3} = 36$ teeth (18 pairs).

The Eocene Tarsiidae, represented by the European species *Necrolemur antiquus,* found in the Quercy deposits of France, are likely to contain the ancestor of the modern genus *Tarsius.* The tarsier is indeed a "living fossil" in the best sense of that overworked term. What is known of the skull and skeleton of *Necrolemur antiquus* indicates that it was almost identical with that of the living *Tarsius.*

Oligocene. Information on primate evolution during the Oligocene Epoch (about 38,000,000 to 26,000,000 years ago) rests principally on discoveries in three areas—Burma, Texas, and Egypt, and by far the most important is Egypt. From the Fayum region of the Sahara, from the upper levels of the Qatrani Formation, has come the first evidence of the emerging Anthropoidea. A number of different genera have been described from the Fayum, including *Apidium, Propliopithecus, Oligopithecus, Parapithecus, Aeolopithecus,* and *Aegyptopithecus.* The Fayum, indeed, seems to be the cradle of the Old World anthropoids. Simons has indicated that the possible ancestors of Old World monkeys (*Parapithecus, Apidium*), of gibbons (*Aeolopithecus*), of apes (*Aegyptopithecus*), and even of man (*Propliopithecus*) are all represented in the fauna. Unfortunately, apart from fragmentary scraps of limb bones, there are no postcranial bones to help. From the evidence provided by the possible forerunners and probable descendants of the Fayum primate fauna, however, it is reasonable to assume that quadrupedalism was becoming established as the typical locomotor pattern, and that vertical clinging and leaping, the characteristic gait of the Eocene forebears of the fauna, was probably retained by at least some of the genera represented at this site.

Rooneyia, an important transitional form

Of unusual interest is the recent discovery of the cranium of a North American omomyid called *Rooneyia;* it is of particular note in view of a belief that primates had disappeared from North America by the late Eocene times. *Rooneyia* is also of considerable interest in itself. The

skull possessed a mixture of primitive (prosimian) and advanced (anthropoid) features, precisely the combination that might be anticipated in a transitional form between lower and higher primates.

Miocene. The Miocene Epoch is probably the most fruitful of all for paleoprimatology. During the 19 million years of its duration (beginning about 26,000,000 years ago), dramatic changes in geomorphology, climate, and vegetation took place. The Miocene was a period of volcanism and mountain building, during which the topography of the modern world was becoming established; of particular relevance to the story of primate evolution are the vegetational changes resulting from the orogenic ones. Grasses, known since the early Tertiary Era, flourished in the new conditions and in many areas that were previously forested. Grassland is known regionally by such names as savanna, llanos, prairies, etc. A new type of primate—the ground inhabitant—came into being during this period. The generalized nature of the bodily form of primates, combined with their specialized brain, made possible this critical step.

In the last few decades considerable additions to the knowledge of ape and human evolution have accrued from Miocene fossil beds in East Africa and Europe.

In Europe further discoveries of the gibbon-like genus *Pliopithecus* and the gorilla-chimp-like genus *Dryopithecus* have been made. *Pliopithecus* remains from Neudorf-an-der-March in Czechoslovakia have provided a remarkably complete picture of the habits of the Miocene apes, which, on this evidence, appear to have possessed bodily forms of a quadruped with morphological overtones of incipient brachiating specializations. Perhaps most interestingly of all, *Pliopithecus,* with its retention of numerous platyrrhine or New World monkey characters, emphasizes the phylogenetically conservative nature of anthropoid evolution.

In East Africa the excavations of the inshore islands and of the Kenyan shores of Lake Victoria by L.S.B. Leakey and a number of colleagues have illuminated man's knowledge of the evolution of the African great apes and have set up provocative questions regarding the origins of man. The *Proconsul* group (relegated to subgeneric status within *Dryopithecus* by the revision of Simons and Pilbeam) is known from three fossil species of ape, *Proconsul major, P. nyanzae,* and *P. africanus.* All three species are represented by a large number of jaw and facial fragments, but the only complete skull, discovered in 1948, belongs to *P. africanus.* Though complete, it was somewhat distorted by pressure of the surrounding rocks during fossilization. Subsequent reconstruction revealed a skull more monkey-like than apelike in its contours, an appearance which belied the unquestionably chimpanzee-like affinities of its teeth. The form of the skull and braincase of *P. africanus,* along with the forelimb skeleton, which is known in great detail for this species, indicates a body form that most closely resembles that of living monkeys. An American anthropologist, Sherwood L. Washburn, has described these creatures as "dental apes." The monkey-like characteristics of the *Dryopithecus* (*Proconsul*) group reflect the recency of the separation of the monkey and ape stocks of the Old World.

Proconsul africanus has been suggested as a possible common ancestor for apes and monkeys, but the general consensus is that *P. africanus* was already too specialized dentally to fit such a role. Be this as it may, there is no doubt that this Miocene fossil ape provides a very reasonable "model" for such an ancestor.

In East Africa two species of fossil gibbons, *Limnopithecus,* have also been discovered, closely related, if not generically identical, to the fossil "gibbon" of European deposits termed *Pliopithecus.* The taxonomic status of these fossils as "gibbons" is somewhat uncertain.

Appearance of hominids

It is towards the end of this turbulent epoch that the first probable ancestral hominids appear in the fossil record. *Ramapithecus punjabicus* was discovered in 1934 by an American paleontologist, G.E. Lewis, in the Siwālik Hill Miocene deposits in northwest India. Lewis tentatively expressed the view that *Ramapithecus* was a hominid, but the scientific climate of the period was not propitious for claims that seemed to imply such remote antecedents for man; as a result Lewis' perceptiveness was largely ignored. Recently, Simons and Pilbeam have resurrected Lewis' specimen and have demonstrated its hominid characteristics. At Fort Ternan in East Africa, meanwhile, Leakey found what at present appears to be another specimen of *Ramapithecus,* but of a different species and called by him *Kenyapithecus wickeri.*

First ground-living monkey

It was also from the late Miocene that the first unequivocally ground-living monkey came to light. Attributed to the living subfamily Cercopithecinae of Old World monkeys, these specimens of *Mesopithecus pentelici* demonstrate the transition between the ancestral monkey type (today represented by the leaf-eating monkeys or Colobinae of Africa, India, Ceylon, Malaya, Burma, and Southeast Asia) and the newly evolved type of ground-living monkey epitomized by *Macaca* or *Papio.* In its known features of skull and postcranial skeleton, *Mesopithecus* is broadly equivalent to the macaque monkeys.

Pliocene. The Pliocene Epoch (from about 7,000,000 to 2,500,000 years ago) was very similar to the present in terms of its geomorphology and climate. Discounting the effects of the recent influence of man on the distribution of forest and savanna in the tropics, the face of the land cannot have differed much from the aspect it presents today. Thus, one would expect that during the Pliocene (bearing in mind the effectiveness of environmental selection) essentially modern forms of primates would have made their appearance, but perhaps the most famous of the Pliocene fossils was far from "modern." *Oreopithecus* was first discovered, fragmentarily, in brown coal deposits in Italy in the middle of the 19th century. The most recent discovery of a complete, though paper-thin, skeleton (so grossly has it been compressed) in Tuscany has demonstrated the power of the environment acting upon the genetic structure, in once again reproducing the trade marks of two of the "inevitable" trends of primate evolution, brachiation, and bipedalism. *Oreopithecus* possessed a number of dental and bony characters that are typically hominid; it also possessed limb characteristics that are typically pongid. The canines were relatively short, the face was abbreviated, and the pelvis was broad. The foot showed characteristics associated with bipedal walking as did the vertebral column (all hominid characters), yet the arms were long and the fingers long and curved. The limb proportions are those of a brachiator (pongid characters). Finally, a number of characters of this confusing hominoid typically reflect those found in Old World monkeys. The present consensus of attitude towards this composite primate is that it is best segregated within the superfamily Hominoidea but in a family of its own, the Oreopithecidae. That the end of the *Oreopithecus* story has not yet been heard is certain; in the meantime the Tuscany fossil stands as a memorial to the primate order, the order which, of all of the class Mammalia, is most subject to environmental adaptation by virtue of its behavioral plasticity.

Few fossils referable to the ape family are known during the Pliocene, and monkey families are scarcely better known. *Libypithecus* and *Dolichopithecus,* both monkeys, were probably ancestral cercopithecines that still retained some features of their colobine ancestry; but neither genus can be placed in a precise ancestral relationship with modern members of this subfamily.

Pleistocene. The Pleistocene, which on faunal evidence is estimated to have started 2,000,000 to 3,000,000 years ago, is the epoch of hominid expansion (see EVOLUTION, HUMAN: *Classes of Hominidae*). Knowledge of nonhuman primates is surprisingly sketchy. No ape fossils are known until relatively recent times and monkeys have been identified in only a few regions in Africa and even fewer in Asia; *e.g., Cercopithecoides* (a colobine-like form), *Gorgopithecus, Dinopithecus,* from South African deposits (forms of the living genus *Papio*); and *Simopithecus* (a giant, ancestral forerunner, according to one authority, of the present-day genus, *Theropithecus,* the gelada) from Olduvai Gorge and South Africa. It is possible that the *Papio-Theropithecus* divergence can be pushed well back into the Pliocene.

One genus of the Pleistocene that is neither ape nor

The special case of *Gigantopithecus*

monkey in the sense that these taxa are interpreted today is *Gigantopithecus*. The romantic story of the discovery of the gargantuan molar teeth of *Gigantopithecus blacki* by von Koenigswald, the Dutch paleontologist, in a Chinese pharmacy has often been told. The boldness of the move that erected a new genus on such apparently slender grounds has been amply justified by the subsequent discovery of several massive jaws from Kwangsi in South China and attributed to this genus. The most recent discovery is of an enormous jaw in the Dhok Pathan deposits of the Siwālik Hills of India. Possibly dating back to the middle Pliocene, this discovery provides a respectably long period of existence for this aberrant giant-toothed hominoid genus. Clearly *Gigantopithecus bilaspurensis* was a member of the family Pongidae with divergent dental specializations that were possibly adaptive for foraging in grassland where tree products were unavailable and ground products available but hard to get. *Gigantopithecus* was a fossil ape that fell by the wayside on its way up to hominid status, a victim, possibly, of overspecialization.

CLASSIFICATION

Distinguishing taxonomic characters. The identifying characters of the primate order have already been discussed above. The point has been made that no single morphological characteristic distinguishes this group exclusively from other mammalian taxa. For instance, while it is possible to make the generalization that primates bear flat nails on the ends of their fingers and toes, it is not strictly accurate, for species such as the tree shrews, the marmosets, and the tamarins have either overt claws (tree shrews) or modified claws (marmosets and tamarins). One can also aver that primates characteristically possess two pectoral mammae, but it must be pointed out that the aye-aye (*Daubentonia madagascariensis*) has its nipples sited in the inguinal region (lower lateral region of the belly) and that mouse lemurs and tree shrews have multiple sets. Generalizations regarding opposability of the thumb, a "classic" primate character, are particularly vulnerable since some marsupials possess this trait and some primates do not. Examples of the generalized nature of primate specializations are too numerous to mention here. At first sight, big toe opposability provides a more promising diagnostic feature but, as with other features, there are exceptions to the rule, man for instance; moreover, it occurs in nonprimates.

Features and exceptions

As F. Wood Jones said in his book *Man's Place Among the Mammals:* "There is no single [characteristic] which constitutes a peculiarity of the Primates: for a primate animal may only be diagnosed by possessing an aggregate of them all." It was with this precept in mind that Sir Wilfrid Le Gros Clark introduced his list of evolutionary trends of primate evolution in 1959, which is amplified below:

1. The preservation of a generalized structure of the limbs with a primitive pentadactyly [*i.e.*, containing five digits] and the retention of certain elements of the limb skeleton (such as the clavicle) which tend to be reduced or to disappear in some groups of mammals.
2. An enhancement of the free mobility of the digits, especially the thumb and big toe (which are used for grasping purposes).
3. The replacement of sharp compressed claws by flattened nails, associated with the development of highly sensitive tactile pads on the digits.
4. The progressive abbreviation of the snout or muzzle.
5. The elaboration and perfection of visual apparatus with the development to varying degrees of binocular vision.
6. Reduction of the apparatus of smell.
7. The loss of certain elements of the primitive mammalian dentition, and the preservation of a simple cusp pattern of the molar teeth.
8. Progressive expansion and elaboration of the brain, affecting predominantly the cerebral cortex and its dependencies.
9. Progressive and increasingly efficient development of those gestational processes concerned with the nourishment of the foetus before birth. (From W.E. Le Gros Clark, *The Antecedents of Man;* © 1959 Edinburgh University Press.)

Two further evolutionary trends may be added to the list:

10. Progressive development of truncal uprightness leading to a facultative (opportunistic) bipedalism.
11. Prolongation of postnatal life periods.

From J.R. and P.H. Napier, *A Handbook of Living Primates* (1967).

A British zoologist, R.D. Martin, has criticized many of these evolutionary trends as irrelevant as determinants of primate origins. In his view, many of them are either symplesiomorph characters (primate characters inherited from an ancestral mammal and therefore nonspecific for primates) or convergent characters (characters that have been acquired by many groups of primates in response to the environmental effects of natural selection). The disputed trends of primates, nonetheless, comprise characters that are of considerable significance for primate classification. Martin's point is that such characters are of little use in distinguishing the order Primates from other placental orders. Martin argues that the inadequate nature of the current definitions of primates is the principal reason why, when so much behavioral and structural evidence points to the opposite conclusion, the tree shrews (Tupaiidae) are presently included in the order. This may well be so. But until the true synapomorph characters (characters of independent acquisition) are identified by extensive comparative studies of all of the relevant characters of primates and nonprimates, there is little justification for arbitrary disenfranchisement of the tree shrews. "Possession is nine-tenths of the law," a principle that, as in this instance, supplies the highly desirable element of stability to zoological nomenclature.

Annotated classification. It is apparent that the taxonomy of primates is in a very fluid condition; but this state of affairs is wholly preferable to the virtual stagnation that existed during the first half of the present century. In view, therefore, of the likelihood of further major revisions taking place during the next decade that may well reveal the fallacies of certain current concepts, the informational content of this section is best served by a conservative, rather than a radical, approach to primate classification. Thus, the most widely accepted current classification is here followed, while fully admitting that it is not wholly adequate. In order to avoid some of the descriptive complications, derived from an imperfect classification, the following scheme outlines the distinguishing taxonomic characters of order, suborders, and families; intermediate taxa, such as infraorders and superfamilies, are omitted.

ORDER PRIMATES

Distinguished by the possession of a hallux (first toe) and a pollex (thumb), one or both opposable; eyes frontally placed on the face, orbit ringed by bone and, in higher members of the group, closed posterolaterally by bone. Brain relatively large. Most of the fingers and toes bear nails, not claws. Number of teeth 18 to 36; canines prominent but dentition as a whole not highly specialized. Diet omnivorous; body weight from about 100 grams (a few ounces) to 180 kilograms (400 pounds), more or less. Twelve Recent families, 55 genera and 197 species; 8 extinct families. Extinct groups are indicated by daggers (†).

†Suborder Plesiadapoidea

†Families Phenacolemuridae, Carpolestidae, and Plesiadapidae

Middle Paleocene to lower Eocene; North America and (except Phenacolemuridae) Europe; collectively, more than a dozen genera.

Suborder Prosimii

Usually quadrupedal and plantigrade, but some species show vertical clinging in trees and saltation on the ground; all have long tails (except *Indri*). Very small to medium in size; substantially insectivorous, but fruit- and leaf-eating common; activity rhythm nocturnal or crepuscular. Suborder characterized by possession of long or longish snouts, naked rhinarium, and naked tethered upper lip (except *Tarsius*); facial vibrissae (whiskers) present. Specialized sublingual fold, often serrated, below tongue. Eyes large, fully frontally facing in *Tarsius,* but somewhat divergent laterally in others; bony orbits not fully closed posterolaterally. Teeth characterized by special arrangement of lower incisors and canines to form a dental "comb," except in *Tarsius* (structure absent) and in *Tupaia* (canines take no part in comb); upper incisors usually small and peglike. Hallux and pollex more or less opposable, tips of all digits bear flat or flattish nails (except in *Tupaia*), second digit of hindlimb bears toilet (grooming) claw. Placenta epitheliochorial and nondeciduate (except *Tarsius*). One or more pairs of nipples. Single births are the rule.

Family Tupaiidae (tree shrews)

Small animals with elongated body and long bushy tail (sca-

ly with terminal tuft in *Ptilocercus*). Short limbs, with fore longer than hind. Long pointed snout. Relatively large brain (compared with nonprimates) with diminished olfactory representation; turbinal bones of nasal region of skull not well developed. Tympanic ring within middle ear cavity. Primitive dental pattern, with 3 pairs of lower incisors, protuberant and arranged as modified "dental comb"; cheek teeth primitive in form. Digits of hands and feet bear claws; pollex is incipiently divergent, but nonopposable; hallux, nonopposable. Both arboreally and terrestrially adapted species occur within the family. Head and body length 10–24 cm (3.9–9.4 in.), tail 12–14 cm (4.7–5.5 in.); 5 genera, 17 species; India and Southeast Asia and Philippines.

†*Family Adapidae (Notharctidae)*
Lower to upper Eocene; Europe, Asia, and North America; about 10 genera.

Family Lemuridae (lemurs)
Size very small to medium; tail longer than head and body length combined; hindlimbs considerably longer than forelimbs; tail more developed in subfamily Lemurinae than in the other subfamily Cheirogalinae. Coat colours highly variable; one species, *Lemur macaco,* shows sexual dichromatism (males black, famales rufous). Tympanic bullae large, tympanic ring contained within middle ear cavity (contrasting with Lorisidae). Cutaneous glands on forearm and upper arm serving marking functions in *L. catta* and *Hapalemur griseus;* absent from other species. Dentition typical of Prosimii, including "dental comb" and reduced upper incisors (absent in *Lepilemur*). Diurnal (*Lemur, Hapalemur*) or nocturnal. Births seasonal; twins common in *Microcebus.* Head and body length 13–46 cm (5.1–18.1 in.), tail 17–56 cm (6.7–22 in.); 6 Recent genera; about 14 species (some possibly extinct); restricted to Madagascar.

Family Indriidae (indri, sifaka, and avahi)
Arboreal animals with exceptionally long limbs, associated with specialized locomotor habit of vertical clinging and leaping. Skull with somewhat downwardly facing foramen magnum, short facial skeleton and very large auditory bullae. Dentition atypical of Prosimii; lower canines absent, first premolar somewhat caniniform, 2 premolars only. Anatomy of stomach and cecum, particularly, specialized in association with a bulky vegetarian diet. Activity rhythm nocturnal (*Avahi*) or diurnal (*Propithecus* and *Indri*). Fossils from Pleistocene. Three living genera, 4 species; Madagascar.

Family Daubentoniidae (aye-aye)
Arboreal, nocturnal family comprising single species *Daubentonia madagascariensis.* Hand specialized by markedly elongated and attenuated third digit. Dentition anomalous: central incisors reduced to single large pair having long roots and a rodent-like function; canines absent, premolars absent in lower jaw and reduced in upper. Symphysis between right and left halves of mandible is fibrous, a condition unique among the primates. Single inguinal pair of nipples. Madagascar.

Family Lorisidae (galagos, lorises, angwantibo, and potto)
Arboreal and nocturnal. Two major locomotor types comprise family: vertical clingers and leapers (subfamily Galaginae) and slow quadrupedal climbers (Lorisinae), the former possessing excessively long hindlimbs, and the latter subequal limbs. Tails short or absent in Lorisinae and long in Galaginae. Second digit of hand reduced, extremely so in Lorisinae (particularly *Perodicticus* and *Arctocebus*) and first digit of hand and foot disproportionately large. Skull globular in form with little protrusion of facial region. Tympanic bullae large, short external auditory canal with tympanic ring sited at junction of middle and external ear. Orbits large, with little lateral divergence. Dentition typical of suborder. Omnivorous (with large insectivorous element). Family comprises 5 Recent genera with about 12 species; Africa, Asia (Lorisinae); Africa (Galaginae).

†*Family Anaptomorphidae*
Upper Paleocene to middle Eocene; North America and Europe; about 10 genera.

†*Family Omomyidae*
Upper Paleocene to lower Oligocene: North America, Europe, and Asia; about 20 genera.

Family Tarsiidae (tarsiers)
Arboreal and nocturnal. Hindlimbs almost twice the length of forelimbs. Hands with 5 digits, fingers surmounted by circular digital pads and backed by tiny triangular nails; thumb opposable, but in a fashion distinct from other Prosimii. Head of femur cylindrical (not spherical as in all other Prosimii except *Galago*), and tibia and fibula fused in lower third (unique to *Tarsius*). Skull characterized by enormous, frontally directed orbits, incompletely closed posterolaterally; nasal space much compressed. Foramen magnum (opening for spinal cord) directed downwards. Short external auditory canal; middle ear with prominent bulla and tympanic ring, sited as in Anthropoidea. Dentition unlike typical prosimian pattern, lacking dental comb; central lower incisors absent; molars sharp-cusped tuberculo-sectorial. Placentation is hemochorial, discoidal, deciduate, as in Anthropoidea. Single young usually produced. One genus, 3 species; Southeast Asia.

†*Family Microsyopidae*
Lower to upper Eocene; North America and Europe; 4 genera.

Suborder Anthropoidea
Quadrupedal, brachiating or bipedal in gait. Diurnal in habit (except night monkey, *Aotus*). Omnivorous in diet. Tail present or absent, varying in length from species to species, prehensile in some. Size ranges from very small (*Saguinus* and *Callithrix,* the tamarins and marmosets) to very large (*Gorilla* and *Homo.*) Faces show varying degrees of protrusion, most marked in the baboons and mandrills and least developed in man. Neurocranium relatively large and with or without sagittal or nuchal crests or both; foramen magnum directed downwards and backwards. Orbits frontally facing and fully closed posterolaterally; nasal cavity narrow. Dentition comprising 30–36 teeth, canines are well developed except in man. Hands and feet prehensile; thumb semi-opposable or fully opposable (except in certain platyrrhine species). Hallux widely abductable (capable of being rotated inward) and capable of prehension in all species except man; flat nails on fingers and toes (except Callitrichidae); specialized toilet claws absent. Two pectoral mammae only. Placenta hemichorial, discoidal, and deciduate. Lengthy gestation period and extended life periods (compared with Prosimii) characterize the suborder. Suborder contains two geographical groups: New World monkeys and Old World monkeys: 6 families and 35 genera; both the tropical and subtropical zones of New and Old World.

Family Callitrichidae (tamarins and marmosets)
Small animals, diurnal, arboreal, and mainly insectivorous. Quadrupedal gait, hindlimbs being 25 percent longer than forelimbs. Thumb nonopposable, but hallux fully abductable; all digits bear modified claws (tegulae) except hallux, which has a flat nail. External nostrils face forwards and sideways (platyrrhine condition); nasal septum broad. Coat colour highly variable between species including wholly black and wholly white forms; one species (*Leontopithecus rosalia*) golden; distinguishing external features include a variety of ear tufts and a diversity of mustaches. Skull long and ovoid in shape with bulging occiput; tympanic bulla present; external auditory canal absent. Dentition consists of 32 teeth (36 in *Callimico*) with 3 premolars and only 2 molars. Tamarins are distinguished from marmosets by long lower canines which project beyond incisors; in marmosets canines and incisors are the same length. Head and body length 13–37 cm (5.1–14.6 in.), tail 20–39 cm (7.9–15.4 in.); weight about 100–700 gm (3.5–24.7 oz). Five genera, 15 species; South America.

Family Cebidae (New World monkeys)
Larger in size than Callitrichidae, with ears more or less naked externally. All digits bear flattened or curved nails; thumb imperfectly opposable. Gait quadrupedal. Prehensile tail present in *Ateles, Alouatta, Brachyteles,* and *Lagothrix.* Bony ear formation as in Callitrichidae. Nose typically platyrrhine. Dentition of 36 teeth; molars quadrangular and quadricuspid, and canines moderately developed in both sexes. Head and body length about 24–70 cm (9.4–27.6 in.; tail usually of same magnitude), weight about 600 to 10,000 gm (1.3 to 22 lb); 11 genera, 29 species; southern Mexico to Brazil.

†*Family Parapithecidae*
Lower Oligocene; North Africa; 2 genera.

Family Cercopithecidae (Old World monkeys)
Medium to large in size. Arboreal or terrestrial. Characterized by naked callosities covering expanded ischial bone of pelvis. Quadrupedal, plantigrade, or digitigrade in gait; hindlimbs longer than forelimbs in most species, but in ground-adapted forms limbs are nearly equal. Thumb present in all genera (except *Colobus*) and functionally opposable. Cheek pouches present in subfamily Cercopithecinae and concerned with temporary storage of food; in subfamily Colobinae pouches are absent but stomach is sacculated in association with a predominantly leaf-eating diet. Tail length variable from very long (Colobinae) to very short or absent (some Cercopithecinae). Muzzle more or less prominent. Noses with external nasal opening facing forwards and downwards or directly downwards ("catarrhine" condition); internasal septum narrow (as contrasted with Callitrichidae and Cebidae). Skull of anthropoid type but without auditory bulla; external auditory canal present. Dentition of 32 teeth; canines large in males, moderate in females; first lower premolar sectorial, molars quadricuspid with transverse linking crests (bilophodonty). Head and body length about 32–

110 cm (12.6–43.3 in.); tail, when present, 2–100 cm (0.8–39.4 in.); weight about 0.7–30 kg (1.5–66 lb); 13 genera, 72 species; Africa, India, China, Japan, and South East Asia.

Family Hylobatidae (gibbons and siamang)

Demonstrate highly specialized form of locomotion called brachiation. Arboreal, tailless catarrhines with forelimbs twice length of hindlimbs. Hands exceptionally long and prehensile; deep cleft between first and second digits; thumb fully opposable by means of unique mechanism at carpo-metacarpal (basal thumb) joint, associated with a specialized musculature. Ischial callosities as in Cercopithecidae but developing late in life. Well developed laryngeal (vocal) sac in *Symphalangus.* Skull of typical anthropoid form, orbits large, forwardly directed and having a well-buttressed lateral margin; strong brow ridge; sagittal and nuchal crests rare; mastoid processes absent. Long sabre-like canines equally developed in both sexes; molar cusp pattern similar to that of the Pongidae and Hominidae. Chromosome diploid number: 44. Head and body length about 40–64 cm (15.7–25.2 in.); weight 4–13 kg (8.8–28.6 lb); 2 genera; *Hylobates* (6 species) and *Symphalangus* (1 species); Southeast Asia.

Family Pongidae (great apes: gorilla, chimpanzee, and orangutan)

Arboreal or ground living. Locomotion consists of brachiation and a modified form of quadrupedalism, called knuckle walking. Characterized by total lack of tail, relatively long forelimbs, elongated hands, and short but fully opposable thumbs. Brain relatively large compared with body size and showing well marked cerebral fissuration. Permanent dentition 32 teeth: incisors large and spatulate, particularly in upper jaw; canines large and conical, particularly in males; well-marked diastema present between upper lateral incisors and canines. Molars large; quadricuspid in upper jaw and quinquicuspid in lower; cusps arranged in so-called "*Dryopithecus*-pattern"; enamel much wrinkled in *Pongo;* third molar relatively small in *Pan.* Skull bears sagittal and nuchal crests in adult males (but rare in *Pan*). Ischial callosities usually absent (*cf.* Cercopithecidae, Hylobatidae). Thorax widest transversely (*cf.* Cercopithecidae); pelvis expanded in iliac region; lumbar vertebrae reduced in number (*cf.* Cercopithecidae). Life periods extended compared with New and Old World monkeys. Chromosome diploid number: 48. Family includes three genera: *Gorilla, Pan, Pongo. Distribution:* Africa and Southeast Asia.

Family Hominidae (man and fossil relatives)

Ground-living, hairless, omnivorous and bipedal. Hindlimbs longer than forelimbs (contrasting with Pongidae and Hylobatidae). Thumb fully opposable and hand capable of both precision grip and power grip; big toe aligned with other toes, foot nonprehensile. Brain relatively large and highly fissured. Neurocranium expanded particularly in parietal and frontal regions; Mastoid process of periotic bone present from birth. Facial skeleton nonprotrusive and strongly flexed relative to cranial base. Dentition of 32 teeth: canines small and blunt, lying flush with occlusal plane; premolars bicuspid; molars small, low, bearing rounded cusps; last molar in upper and lower jaw frequently unerupted. Postcranial skeleton shows adaptations to upright posture particularly in vertebral column, pelvis and lower limb. Chromosome diploid number: 46. Life periods (including growth period and life expectancy) extended as compared with Pongidae. Head and body length about 80 to more than 130 cm (31.5–51.2+ in.); total length (height) about 130 to over 220 cm (51.2–86.6 in.); weight from about 40 to more than 120 kg (88–264 lb); one Recent species, *Homo sapiens;* distribution world-wide.

Critical appraisal. The classification of the primates by George Gaylord Simpson is the most widely used list today. Since it was published in 1945 a number of revisions at family level and below have been incorporated; the higher categories have remained unchanged, although they have been the subject of considerable criticism and doubt. W.C. Osman Hill's classification, which, with modifications, follows R.I. Pocock's list published in 1918, has not enjoyed a wide popularity, in spite of the success of "Primates," Hill's monumental treatise on the order. Hill excludes the tree shrews from the primates altogether; he also removes the tarsioids from among the lemurs and lorises and places them with the monkeys and apes. A compromise view between the classifications of Simpson and Hill has been proposed by American paleontologist Alfred S. Romer. Romer also omits the tree shrews from the order and gives the tarsioids equal status with lemurs and lorises, Old World monkeys, New World monkeys, and apes in a separate subordinal group. It seems likely that before very long Romer's classification, or one closely similar to it, will become generally accepted. Reference to the annotated classification given above (based on Simpson's) will indicate how ill-fitted is the living genus *Tarsius* to share subordinal status with the prosimians.

Problem of the tarsiers

Several revisions of lower taxa (genus and species) have been proposed in recent years which are proving generally acceptable. It has been suggested that *Cynopithecus niger,* the Celebes black "ape," be included within the synonymy of *Macaca,* in which case this species therefore becomes *Macaca nigra.* Also widely accepted is inclusion of *Comopithecus* (the hamadryas baboon) in the synonymy of *Papio* (the common baboon), but opinions differ as to whether the hamadryas and the common baboon are not better considered a single species in view of the good evidence of hybridization in the wild. Synonymy of the genus *Mandrillus* with *Papio* has also been suggested, largely on the evidence of molecular data from blood protein studies, but has not found much support. Less controversial is the conclusion, based on field evidence, that the three generally accepted species of savanna monkeys (*Cercopithecus aethiops, C. sabaeus* and *C. pygerythrus*) actually comprise a single polytypic species, *C. aethiops.*

Less universally acclaimed is the suggestion by Simpson that the structural and behavioral differences between the chimpanzee and the gorilla are insufficient to justify continued generic separation. According to this view, which has some support, *Pan,* having nomenclatorial priority, should include two species, *Pan troglodytes* (chimpanzee) and *Pan gorilla* (gorilla). At a symposium on the Old World monkeys held in 1969, the synonymy of *Nasalis* (the proboscis monkey) with *Simias* (the Pagai Island langur) and *Rhinopithecus* with *Pygathrix* was recommended but, as yet, is by no means universally accepted. Field, museum, and laboratory work has led to numerous changes in classification of primates by subspecific level that need not concern us here. (J.R.N.)

Edentata (armadillos, sloths, anteaters)

Edentata, an order of mammals, includes 31 living species distributed among the armadillos, true anteaters, and tree sloths, as well as eight extinct families of ground sloths and armadillo-like animals. The living families and six of the extinct families comprise the suborder Xenarthra. A second suborder, Palaeanodonta, consists of two extinct families. The entire evolutionary history of the edentates is restricted to the Western Hemisphere and the majority of the living species occur today in South America.

GENERAL FEATURES

Edentata means lacking teeth, though in reality only the true anteaters are toothless. The majority of edentates have simple, peglike cheek teeth that lack enamel; caninelike teeth do occur in some forms. Certain armadillos may have as many as 100 teeth. Edentates possess specialized traits, such as reduced dentition, a long sticky tongue, powerful, clawed, forefeet, associated with their insect diets. They also have primitive traits, such as the possession of five toes on the hindfeet, a simple uterus, and small, uncomplicated brain, which place them close to the primitive stock that gave rise to the infraclass Eutheria, which includes all placental mammals.

Primitive traits

A persistent myth holds that any kind of armadillo, in defense against enemies, can roll itself into an impregnable armour-plated ball. In fact, only the three-banded armadillo (*Tolypeutes tricinctus*) is able to perform this feat. Most armadillos contort themselves into a reasonable facsimile of a ball, but some only slightly increase the convexity of the dorsal shell (carapace) while flattening against the ground. Individuals of some species flee to burrows or attempt to burrow right on the spot but may freeze and feign death when flight and rapid excavation are somehow prevented. Misconceptions about other edentates also exist. The toothless, weak-jawed anteaters seem to be sucking in ants and termites through a small tube, when, in fact, it is the sticky product of the greatly enlarged submaxillary glands covering the lengthy extensible tongues that facilitates the capture of these insects. The

slow-moving arboreal sloths, typically depicted as moving upside down along a branch, really spend only about 10 percent of their time moving at all.

The armadillos and anteaters are important ecologically, helping to control many kinds of abundant insects, especially ants and termites. Armadillos feed on some carrion and thus play a small part in cleaning up the environment. Although some armadillos may become agricultural pests, and the armour of several species is made into purses and other curios, edentates are generally of little economic importance. Generally, the greatest interest in edentates lies in their bizarre appearance.

Size range

Living edentates range in size from the tiny pink fairy armadillo or "lesser pichiciego" (*Chlamyphorus truncatus*) of Argentina, measuring about 16 centimetres (6.3 inches) in length and weighing little more than 100 grams (3.5 ounces), and the slightly larger two-toed or silky anteater (*Cyclopes didactylus*), just over 37 centimetres (14.6 inches) long and weighing about 325 grams (11.5 ounces), to the 60-kilogram (132-pound) giant armadillo (*Priodontes giganteus*), nearly 1.5 metres (five feet) long, and the giant anteater (*Myrmecophaga tridactyla*), which weighs up to 25 kilograms (55 pounds) and may be over two metres (6.6 feet) in length. Among extinct forms, an Oligocene armadillo, *Prozaedyus proximus,* had a skull under eight centimetres (3.1 inches) in length and was only somewhat larger than the smallest of the living armadillos. The Pleistocene ground sloth, *Megatherium americanum,* was six metres (20 feet) long and was larger than a modern elephant; a Pleistocene glyptodon (*Doedicurus clavicaudatus*) was more than four metres (13 feet) long and 1.5 metres (five feet) high. Both flourished in South America about 500,000 years ago.

Anteaters, tree sloths, and armadillos are strikingly different in appearance, but within each of these families, there is only a modest diversity of structure. The toothless anteaters vary from strictly arboreal to terrestrial forms that are characterized by progressively more elongate tubular mouths. The arboreal species have only two (*Cyclopes*) or three (*Tamandua*) clawed toes on the forefeet, whereas the terrestrial giant anteater has four. In addition, the arboreal species have prehensile tails. At a casual glance, the two genera of herbivorous (plant-eating) tree sloths (*Bradypus, Choloepus*) differ from each other in general appearance only in the number of toes on the forefeet (two in *Choloepus* and three in *Bradypus*) and in pelage coloration. The armadillos (Dasypodidae) are distinguishable on the basis of relative size and on the distribution, flexibility, and pattern of the bony shields that form their body armour.

The anteaters, essentially a tropical family, occur from northeastern Argentina and northern Uruguay to the extreme northern part of the state of San Luis Potosí, Mexico. Tree sloths are restricted to forests from northern Rio Grande do Sul, Brazil, to southeastern Honduras. Armadillos are found from the northern part of the province of Santa Cruz, Argentina, to south-central Kansas. Most living edentates are fairly common in their preferred habitats, although several species have disappeared from parts of their range largely because of habitat destruction and excessive hunting by man. These endangered species include the giant anteater, the three-toed sloth (*Bradypus torquatus*), the three-banded armadillo, the giant armadillo, the pink fairy armadillo, and Burmeister's armadillo (*Burmeisteria retusa*).

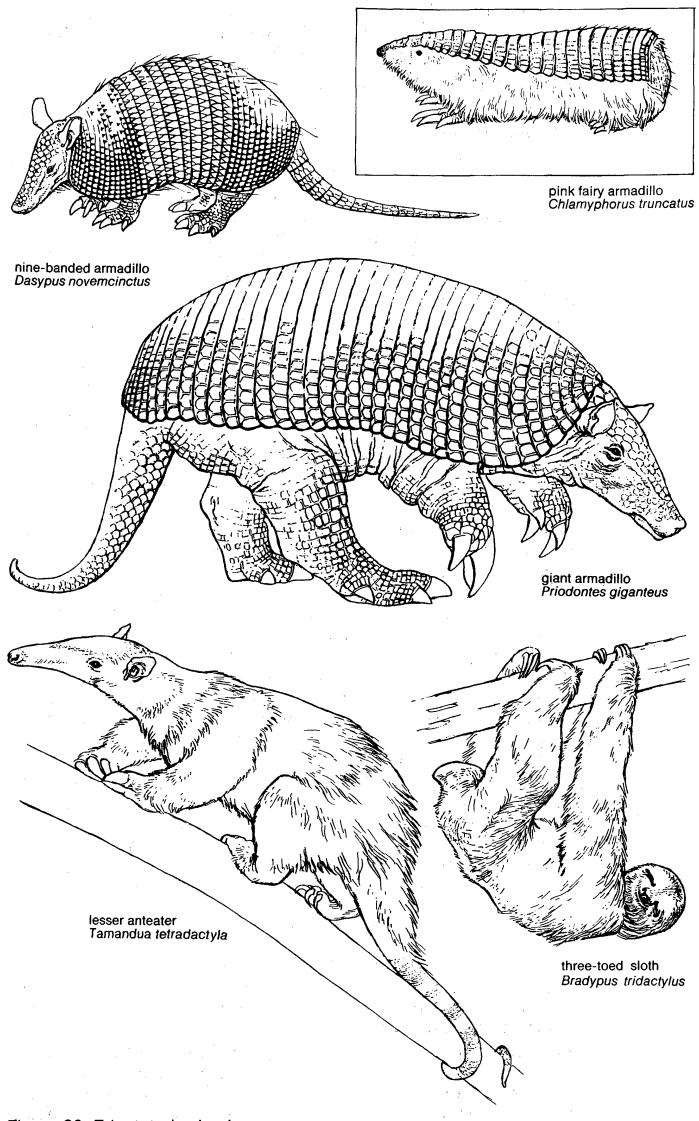


Figure 26: Edentate body plans.
Drawing by J.C. Barberis

NATURAL HISTORY

Behaviour. Anteaters eat termites, ants, and occasionally other insects. They may forage in trees (two-toed), on the ground (giant *Myrmecophaga*), or in both situations, as does the lesser anteater (*Tamandua tetradactyla*). Their prey is usually located by smell, the sense of which is better developed in these animals than are those of eyesight and hearing. With the long, powerful foreclaws, the anteaters rip open anthills and termite mounds, into which they then insert their long sticky tongues. The tongue, covered with prey and debris, is then drawn into the tubular mouth. When frightened, the giant anteater usually flees at a slow clumsy gallop, but if forced it will fight, using its claws to rip the enemy. *Cyclopes and Tamandua* also use their foreclaws for defense, anchoring themselves to a branch with their prehensile tails. When alarmed, these two emit shrill calls.

Anteaters travel solitarily or as mother and young and do not seem to have permanent resting places. The giant anteater chooses a secluded spot such as a hollow log or the abandoned burrow of some other animal and then curls up to rest, with its long bushy tail covering its head and body. The arboreal anteaters find shelter in trees. The two-toed and lesser anteaters are mainly nocturnal, but the giant anteater is both nocturnal and diurnal. In areas remote from civilization, the giant anteater is most active during the day, whereas in heavily populated regions it is primarily nocturnal.

Food habits of sloths

Tree sloths are strictly herbivorous, but the absence of enamel on their teeth suggests that the group perhaps was once insectivorous. The lack of enamel could pose serious problems relating to tooth wear for animals eating such abrasive food as vegetation. The difficulty, however, is alleviated by continuous growth of the teeth, which have an outer covering of hard cement. As is typical of herbivores, tree sloths have large multichambered stomachs in which the slowly digesting plant material steeps. Constantly filled, the stomach and its contents constitute about 30 percent of the sloth's body weight. A meal may take up to a week to digest. Every two to eight days, the sloth descends to the foot of its home tree and passes a large amount of excrement.

The diet of the two-toed sloth consists of a variety of leaves, stems, and fruits. The three-toed sloth, on the other hand, eats only the leaves of the cecropia tree. The sloth locates its food by touch and smell and then hooks a branch with its curved claws and pulls it to the mouth. Although colour vision is present, the eyesight and hearing of sloths are not very well developed, and they orient themselves mainly by touch.

In the wild, the hair of the sloth's body is inhabited by algae, which flourish in the jungle humidity. A variety of mites, beetles, and small moths live as commensals (harmless companions) in the hair, feeding on the algae.

Tree sloths are normally solitary creatures and spend the days and most of the nights sleeping. The little time that remains is spent eating, resting, and moving. This "laziness" contributes greatly to their survival, because their diurnal sleeping posture (hanging from a branch with feet bunched together and head tucked in on the chest) and their coloration give them the appearance of a bunch of dead leaves, rendering them virtually invisible. At night, when they are active, their slow movements do not attract the attention of nocturnal enemies. In addition, sloths have a thick skin and the ability to tolerate strong poisons and severe injuries.

When angered or otherwise disturbed, the three-toed sloth utters a high-pitched cry that sounds like "ai," whereas the two-toed sloth snorts and hisses. Normally, sloths are docile and in the presence of danger remain still, relying on their protective coloration to conceal them. If molested, however, they bite savagely and strike out furiously with the sharp foreclaws.

Burrowing by armadillos

Armadillos are primarily nocturnal, spending the daylight hours in their burrows, in expanded nest chambers lined with grasses and leaves. The burrow is also used for the escape from enemies and as a food trap. When using a burrow, in an emergency, some species are able to flex their bony back plates and brace with the feet in such a manner as to become firmly wedged in the tunnel and virtually impossible to pull out. When in danger away from a burrow, armadillos will draw in their feet and lie so that the edges of their armour touch the ground, or will run, burrow rapidly, or even defend themselves with their claws. The three-banded armadillo can roll itself into a ball with feet inward and the head and tail protected by special shields. If captured, an armadillo usually reacts by "playing dead," either stiffening or relaxing, but in either case remaining perfectly still for a brief period. If this course of action does not result in release, the captive armadillo begins kicking vigorously, often facilitating escape. Several kinds of sounds are reported to have been made by fleeing or otherwise agitated armadillos. The peludos (three species of *Chaetophractus*) make snarling sounds, the mulita (*Dasypus hybridus*) repeatedly utters a guttural monosyllabic sound reminiscent of the rapid fluttering of a human tongue, and other species emit a series of grunts or buzzes when alarmed.

The act of burrowing involves the forelimbs, hindlegs, and tail. The heavy claws on the forefeet are used for scraping and loosening the dirt, pushing it toward the rear. The hindlimbs throw the dirt out behind while the tail elevates the rear end of the animal so that the hindlegs may kick more freely. In this manner, burrows from 0.6 to six metres (two to 20 feet) long, from 0.3 to 1.5 metres (one to five feet) underground and with as many as 12 entrances, may be constructed in a relatively short period. Occasionally, other small animals share these burrows. In North America, rabbits (*Sylvilagus*) and cotton rats (*Sigmodon*) have been found living in the burrows of the nine-banded armadillo.

Armadillos eat insects, mollusks, frogs, small reptiles, carrion, and some plant material and will burrow underneath an animal carcass to obtain maggots and other insects. An acute sense of smell enables some armadillos to detect grubs and other insects up to 12 centimetres (nearly five inches) below the surface of the earth. The peludo is reputed to make conical holes with its snout in the earth in search of food. It first inserts its snout and then rotates its body so that it literally drills the hole. It is able to hold its breath under such circumstances for up to four minutes. Armadillos have relatively good hearing and vision.

Locomotion. The giant anteater walks with a shuffling gait bearing the weight of its foreparts on the side of the fourth (outermost) digit on the front foot. When hard pressed, this species breaks into a slow, clumsy gallop. It is also a good swimmer. The tamandua walks rather clumsily on the sides of its forefeet when on the ground, in a manner similar to that of the giant anteater. The tamandua is an agile climber, aided both by its foreclaws and hindclaws and by its prehensile tail. Two-toed anteaters are strictly arboreal and are thus very strong climbers. They also have a prehensile tail.

The claws of tree sloths, which are quite similar to those of their terrestrial forebears, have become adapted to a strictly arboreal life. On the ground, a tree sloth is nearly helpless and cannot move at all unless there is something to grasp, then it is able to drag itself along with its claws. Most of the time sloths simply hang upside down, but when they do move they use a slow hand over hand motion both on horizontal and upright branches. Interestingly enough, they are excellent swimmers.

Bipedalism in armadillos

Generally, armadillos walk on the tips of the claws of their front feet and on the soles of the hindfeet; occasionally, they rear up and walk bipedally on their hindlegs. They trot with a stiff-legged gait. At a gallop, some species can outdistance a man. Because of their heavy dermal armour, armadillos have a specific gravity so high that they would normally sink in water. To overcome this problem, they gulp quantities of air that give them the buoyancy needed for swimming. Under some circumstances, armadillos seemingly take advantage of their specific gravity and cross small streams simply by walking under water on the bottom.

Habitat. Edentates occur in a variety of habitats, in temperate, subtropical, and tropical environments. Giant anteaters are found in savannas and in wet forests, where they feed either at night or in the day, ripping the terrestrial mounds of termites and ants apart with their powerful foreclaws. Lesser and silky anteaters are both arboreal, the latter totally so. Lesser anteaters prefer tropical forests, parklands, and, to a lesser extent, savannas, in which they forage mostly in trees on ants, termites, and various other insects. They are predominantly nocturnal but may be active at twilight as well. Silky anteaters, which are nocturnal, are found in tall humid tropical forests usually high up in the trees, where they feed almost exclusively on termites. The herbivorous tree sloths occur only in dense tropical rain forests. Armadillos frequent grasslands, parklands, open and even thick forests. They may live largely in subterranean burrows (pink fairy armadillo) in grasslands or dig burrows for temporary shelter in forests (*Dasypus* species). The shape of the burrow entrance is often diagnostic of the cross-sectional shape of the owner; the nine-banded armadillo (*Dasypus novemcinctus*), for example, which has a high rounded carapace, digs a nearly circular entrance. The six-banded armadillo (*Euphractus sexcinctus*), which is compressed from top to bottom, digs a burrow with an entrance that is wider than it is high. The habitat preference of the three commonest armadillos in Uruguay illustrates the ecological diversity of this group. *Dasypus hybridus,* the smallest of the species, prefers grassland and most of its burrows are found there. The six-banded armadillo burrows near streams in open woodlands and grasslands where it occasionally overlaps with the former species and with the nine-banded, which most frequently makes its burrows in woodlands.

Reproduction. The gestation period of those edentates for which it is known ranges from 65 days in the hairy armadillo (*Chaetophractus villosus*) to 263 days in the two-toed sloth. The protracted gestation period of the sloth is attributed to delayed implantation. The very early embryo, consisting of but a few cells, lies free in the uterus without developing further for several months, before becoming imbedded in the uterine wall, at which time normal embryonic growth recommences. Anteaters and tree sloths normally bear single young. Most armadillos give birth to one or two young, although some species demonstrate polyembryony (the birth of from two to 12

Delayed implantation

identical young, which develop initially from a single fertilized ovum) and delayed implantation (attachment of the ovum to the uterine wall after a variable delay period). The nine-banded armadillo has both of these phenomena. The delayed implantation is seasonal, the stages occurring at fixed times of the year, related to environmental influences. In the Northern Hemisphere, young of this species are born in March and April. Breeding occurs between June and September, but, regardless of when breeding occurs, implantation is delayed until November and the postimplantation pregnancy is between November and March. Delayed implantation is presumably beneficial because birth does not take place until spring, when food is plentiful.

Little information is available on the birth and early life of many edentates. A recent study of the two-toed sloth (*Choloepus didactylus*) revealed that the young is born as the mother hangs upside down. The baby emerges head first and face upward, and, as soon as its forelimbs are free, it grasps the abdominal hair of the mother and pulls itself up on her chest. The mother may assist in birth by pulling on the young. Subsequently, the female breaks the umbilical cord with her teeth.

The young sloth's eyes and ears are open at birth. It suckles for about five months but does not leave the mother until the ninth month, when adult proportions, but not size, have been attained. Adult size is reached at from two and one half to three years. Young anteaters also cling to their mothers for several months to a year after birth. The giant anteater rides on its mother's back, the two-toed anteater on its mother's tail.

Armadillo young are born with soft leathery skins, which harden as they grow older. In the six-banded armadillo, the eyes do not open for several days after birth, but in the nine-banded armadillo, the young are born with their eyes open. Young armadillos are able to walk when only a few hours old. In the nine-banded armadillo, physical maturity is achieved at six months, but females do not have their first litter until they are two years old. In captivity, the giant anteater has lived 14 years, the two-toed sloth 19 years, the peludo 15½ years, and the three-banded armadillo 11 years.

FORM AND FUNCTION

External appearance. Anteaters (Myrmecophagidae) have elongate heads; tubular muzzles with no teeth; and long, sharp claws. Their toes vary in number from two to four on the forefeet and from four to five on the hindfeet, depending on the species. The thick coat of the giant anteater is gray with lateral white-bordered blackish stripes, and the hair is long and straight, especially on the characteristically stiff bushy tail. Tamanduas, or lesser anteaters, have short hair, which is usually cream-coloured to brownish, with a black vestlike area over the upper thorax. The entire underside of the tail, as well as the tip, is naked. The two-toed anteater (*Cyclopes*) has soft silky fur, usually golden yellow in colour. In this species, only the underside of the prehensile tail is naked. Males and females are of approximately equal size in *Cyclopes* and *Tamandua,* but males are much larger in *Myrmecophaga.* All anteaters have enormously developed salivary glands that supply the sticky saliva that covers their long extrusible tongues.

The shaggy pelage of tree sloths grows from the belly toward the back on the body and toward instead of away from the body on the limbs (*i.e.,* downward, when the animal is suspended beneath a branch), thus facilitating the shedding of rain. Tree sloths have pale faces with darker eye rings. The body hair is yellowish to brownish, but the individual hairs are fluted and contain a growth of algae that gives a greenish cast to the entire animal, especially in the rainy season. In the drier conditions of captivity, the algae die and the animal loses the natural coloration. The legs are long, with the forelimbs longer than the hindlimbs, and are designed for suspension of the body rather than for "column-like" support.

Algae on tree sloths

Armadillos are characterized by a dorsal bony carapace made up of scapula (shoulder), dorsal, and pelvic shields, separated by a series of movable bands. A bony cephalic (head) shield is also present. The tail is typically armoured as well, except in the naked-tailed armadillos of the genus *Cabassous.* The body hair is greatly reduced, but some, such as the pink fairy armadillo, have dense fur on their sides and underparts. A sparse growth of hair also occurs among the bony plates of all species.

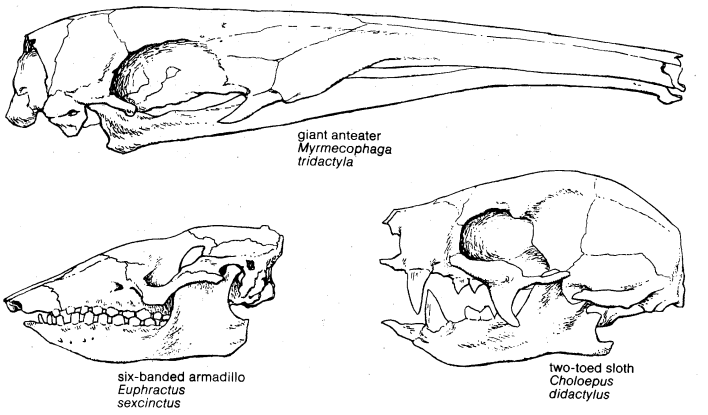


Figure 27: Edentate skulls.
Drawing by J.C. Barberis

Specializations of internal anatomy. Female edentates have a common urogenital sinus (*i.e.,* both reproductive and excretory functions involve a single chamber) and a simple uterus (a single uterine cavity rather than a branching two-chambered organ). The placenta (the uterine connection between the mother and embryo) is discoid in shape. In other words, the actual areas between which gases, food, and nitrogenous wastes are exchanged are restricted to a single disc-shaped region. The placenta is also deciduous; *i.e.,* it is shed at the time of birth. Females have one or two pairs of mammary glands that are located just lateral to the "armpits," on the chest or on the abdomen. In males, the testes are abdominal and the length of the penis is associated with the degree of difficulty encountered in copulation. In tree sloths, where the partners hang abdomen to abdomen, it is quite small, whereas in armadillos, in which the carapace prevents the partners from getting very close to one another, the penis is long and pendulous.

In general, edentates have poor temperature control. Definitely related to the conservation and manipulation of heat are the extensive anastomoses (fusions) of the veins and arteries (termed retia mirabilia) in the limbs and tails of most species. These structures conserve heat, transferring heat from the arterial components carrying warm blood outward along the limbs to the veins lying adjacent for return to the trunk. At the same time, the appendages are cooled.

Tree sloths are heterothermic (*i.e.,* have imperfect control of body temperature) and under experimental conditions their body temperature, normally ranging between 28° and 35° C, may drop to 20° C. At this temperature, the animals become torpid. Temperature control is likewise relatively poor in armadillos. The fact that it builds a subterranean leaf-lined nest has aided the nine-banded armadillo in the northward expansion of its range into the United States. Nonetheless, many individuals of this species perish during severe winters.

EVOLUTION AND PALEONTOLOGY

The ancestry of the order Edentata is believed to be rooted in some unknown but probably generalized insectivore (order Insectivora). The oldest known members of the two presently recognized suborders of edentates were already too specialized for either group to have given rise to the other. Members of the extinct suborder Palaeanodonta, a group restricted to North America, by the time of their first appearance in the fossil record had already developed specializations that eliminate them as possible direct ancestors of the suborder Xenarthra, which flourished throughout the Tertiary in the then isolated South American landmass.

The palaeanodonts, which lived from the late Pale-

The palaeanodonts

ocene until the beginning of the Oligocene, comprised two families, Metacheiromyidae and Epoicotheriidae. The most completely preserved form in the former family is *Metacheiromys,* an archaic mammal about 45 centimetres (17.7 inches) in length, reminding one somewhat of an armadillo lacking armour. This animal possessed an assemblage of clearly edentate traits, including the near absence of teeth (except for the enlarged bladelike canines), long compressed foreclaws, and a slight inward twist of the wrist, which is typical of virtually all members of the order. Other metacheiromyids did not show the strong development of the canines. *Palaeanodon,* known from the upper Paleocene and probably ancestral to *Metacheiromys,* had a series of four or five peglike cheek teeth that lacked enamel. The Epoicotheriidae, which had a skull that in outline is reminiscent of that of a small armadillo, were apparently small aberrant forms, highly specialized for a fossorial (digging) livelihood. When the palaeanodonts passed from the scene at the end of the Oligocene, they left no descendants.

The second suborder, Xenarthra, by far the more successful group of edentates, enjoyed an evolutionary history independent from that of the palaeanodonts because North and South America were separated throughout most of the Tertiary, not becoming rejoined until the Pliocene. Contemporaneous with the earliest palaeanodonts of North America was the first South American xenarthran, *Utaetus,* already recognizable as a primitive armadillo and thus sufficiently advanced to eliminate any possibility that known palaeanodonts might be directly ancestral to xenarthrans. The fact that primitive armadillos are the first known members of the suborder strongly suggests that the other xenarthran families are in some way derived from the Dasypodidae.

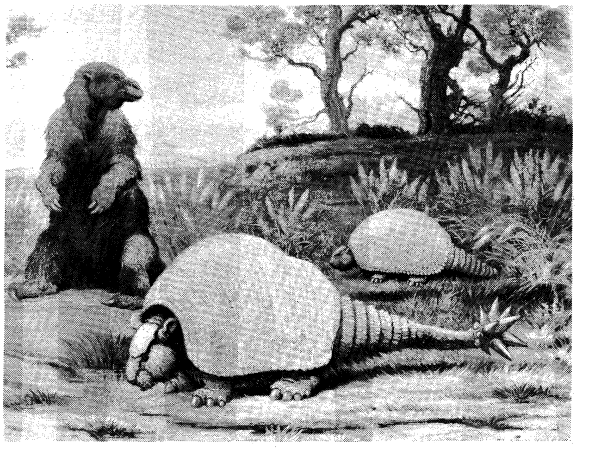
The Tertiary radiation of edentates in South America resulted in several more or less separate lineages, which are conveniently arranged in three infraorders. The first of these infraorders, Loricata, includes the armadillos (Dasypodidae) and their relatives, the extinct glyptodonts (Glyptodontidae) or so-called turtle armadillos, distinguished from the former especially by their possession of a rigid carapace. Armadillos flourished in South America in the Tertiary and were abundant by the Miocene. In the Pleistocene, some species had moved northward as far as what is today the central United States.

Fossil armadillos

Four distinctive lines developed in the Dasypodidae, comprising the subfamilies Dasypodinae, the Stegotheriinae, the Pampatheriinae, and the Chlamyphorinae (including only the living genera *Chlamyphorus* and *Burmeisteria*). The Dasypodinae includes most of the living genera of armadillos as well as *Utaetus.* The fossil members of this group varied in size from that of a small contemporary armadillo to animals as large as the South American Pliocene form *Macroeuphractus* that had a skull nearly 28 centimetres (11 inches) long. The head and body of the latter must have been almost two metres (6.6 feet) long. The fossil members of this subgroup all seemed to favour invertebrates in their diets, but they probably ate carrion and some plant material as well. This distinguished them from the extinct Stegotheriinae, a subfamily whose members were highly specialized for a strictly insectivorous diet. *Stegotherium* of the lower Miocene of South America had an elongate anteater-like skull with much reduced cylindrical cheek teeth.

The Pampatheriinae, known from the late Miocene to the Pleistocene, was a group characterized by great size. At least one form, *Pampatherium,* a nearly complete skeleton of which was recovered from a Pleistocene deposit in Texas, was nearly as large as a rhinoceros. This animal and its closest relatives were apparently adapted to an herbivorous diet. Both insect- and plant-eating groups are thus included among the true armadillos.

The genus *Peltephilus* and several other closely related genera of aberrant armadillo-like creatures have been placed in their own family, Peltephilidae, because, among other characteristics, they possessed at least one pair of hornlike protuberances on the muzzle. These animals, which appeared in the Oligocene and disappeared in the Pliocene, were apparently specialized for flesh eating.



(Left) Ground sloth (*Megatherium*) and (right) glyptodonts (*Doedicurus*). Painting from fossil skeletons by Charles R. Knight.

By courtesy of Field Museum of Natural History, Chicago

The glyptodonts (Glyptodontidae) evolved from primitive armadillos, first appearing in the fossil record in the Eocene with the genera *Glyptatelus* and *Lomaphorelus* but remaining relatively uncommon until the beginning of the Miocene. Subsequently, they experienced a radiation of their own and by the Pleistocene several kinds had spread into North America. Throughout their history, they demonstrated an increase in size culminating in the Pleistocene giant, *Doedicurus.* They became extinct before the end of that epoch.

The giant glyptodont

The second infraorder, Pilosa (hairy) includes the sloths, both the living arboreal forms (Bradypodidae) and the fossil ground dwellers of the families Megatheriidae, Megalonychidae, Mylodontidae, and Orophodontidae. There are no fossils of tree sloths, but they appear to be structurally related to the ground sloth families Megatheriidae and Megalonychidae. The oldest fossils of ground sloths belong to the Megatheriidae and come from Oligocene deposits in South America. By the early Miocene, primitive members of this group were rather common. One form, *Hapalops,* although typical of that time, was of relatively small size for a ground sloth. It had a slender body about 1.3 metres (4.5 feet) in length and had five toes with developed claws on both its forefeet and hindfeet. *Hapalops* had a moderately elongated skull with slender premaxillae (bones of the upper jaw), and a process resembling a spout that protruded from the lower jaw and may have supported horny plates for cropping. This animal had five upper teeth and four lower teeth in each jaw, the first upper pair of which resembled canines. The small Pleistocene ground sloth *Nothrotherium* apparently was a descendent of *Hapalops* and the two comprise a branch of the Megatheriidae, which throughout its history was given to small size. This stock differed greatly from the main stem of the family, which had no specialized "canines" and had the two inner toes reduced, represented by the Pleistocene *Megatherium,* a massive ground sloth larger than an elephant.

The megalonychids, although related to the megatheriids, were distinguished from them by their definite caninelike teeth and lack of the spoutlike process on the lower jaw. The megalonychids ranged in size from one island-dwelling animal no bigger than a domestic cat to the mainland ox-sized *Megalonyx.* Both the Megalonychidae and the Megatheriidae reached southern North America by the Pleistocene, but they became extinct on the Continent at the end of the epoch. Some megalonychids invaded the West Indies, however, and experienced a minor radiation in Hispaniola, Puerto Rico, and Cuba. One or more species may have flourished there to within a few hundred years of the present.

Distinct from the megatheriid–megalonychid line of ground sloths were the Mylodontidae, which first appeared in the Oligocene; from the Pliocene until their extinction at the end of the Pleistocene, they were the dominant South American group of ground sloths. The mylodonts

Mylodonts differed from the other ground sloths in their partially developed upper canines, reduced first toe of the hindfoot and rounded ossicles (bony tubercles) in their skin. The mylodonts followed the trend of other ground sloth families toward great size and awkwardness, which in this group culminated in the Pleistocene giant, *Glossotherium,* which was the size of a large ox. As with other xenarthran groups in the Pleistocene, the mylodonts extended as far north as the southwestern United States.

The other family of ground sloths, the Orophodontidae, is known only from the Oligocene of South America. This briefly-lived side group is represented by *Orophodon* and *Octodontotherium,* both of which were rather small sloths that do not seem to have left any descendants.

The final infraorder, Vermilingua, comprising the anteaters (Myrmecophagidae), has not been found in the fossil record any earlier than the Miocene. The fossil myrmecophagids, as represented by the Pliocene form *Palaeomyrmedon,* were already completely toothless and presumably well adapted to an insectivorous diet. The Myrmecophagidae first appeared in South America in the early Miocene and apparently only fairly recently have spread to Central America.

CLASSIFICATION

Distinguishing taxonomic features. The degree of development of certain sets of structural characteristics taken in concert defines the Edentata and its subgroups. None of these features is univerally distributed throughout the order. The most important are as follows: (1) in addition to the normal zygapophyses (protuberances), the occurrence of supplementary articular surfaces called xenarthrous articulations between the arches of sequential posterior trunk vertebrae; (2) development of certain projections of the scapula bone, or shoulder blade, called the acromion and coracoid processes; (3) articulation of the ischium bone of the pelvis with the proximal caudal (tail) vertebrae, forming an elongate sacrum; (4) length and massiveness of limb bones; (5) fusion of distal limb bones; (6) hyperdevelopment of claws; (7) development of forefeet and hindfeet; (8) a skull characterized by a small brain housed in a long tubelike brain case, typically with a reduced premaxilla (*i.e.,* elongate palate), often an incomplete zygomatic arch, and a reduction in number of teeth and their enamel covering.

Annotated classification. The classification presented here combines those proposed by the two paleontologists R. Hoffsteter and Alfred S. Romer. The features given below are only those that distinguish each group from the others. Groups indicated by a dagger (†) are known only as fossils.

ORDER EDENTATA

Placental mammals with reduced dentition, often lacking enamel; heavily clawed forelimbs for burrowing. Coracoid process of the scapula more strongly developed than in other eutherian mammals. About 31 recent species.

†Suborder Palaeanodonta

Fossil only. Bony carapace and the xenarthrous articulations on the vertebrae lacking; no fusion between the ischium and caudal vertebrae; simple teeth, usually single-rooted and lacking enamel; slight inward twist to the forelimbs.

†*Family Metacheiromyidae.* Upper Paleocene to lower Eocene. Small, lightly built edentates with sharply pointed somewhat laterally compressed canines; postcanine teeth reduced or lacking; armadillo-like feet with 4 to 5 digits; the lateral digits reduced, the inner greatly so. North America.

†*Family Epoicotheriidae.* Lower Eocene to lower Oligocene of North America. Very small burrowing edentates, with a domed skull having a depressed snout and possessing a zygomatic arch; having 6 slightly differentiated cheek teeth, nearly cylindrical, lacking enamel, and single-rooted.

Suborder Xenarthra

Some extinct and all living edentates; all with xenarthrous posterior trunk vertebrae. Lack teeth or have only poorly differentiated cheek teeth, lacking enamel and having single roots. Fusion (symphysis) between the ischium and the anterior caudal vertebrae in all except the 2-toed anteater (*Cyclopes*).

Infraorder Loricata

Bony plates covered with horny plates, forming a protective carapace over the trunk.

Family Dasypodidae (armadillos). Lower Eocene to Recent in South America; Pleistocene to Recent in North America; forests, savannas, and plains; primarily in the tropics and subtropics.Teeth more than $\frac{6}{6}$ (6 above and below on each side); top and sides of the body covered with horny scutes over a bony, flexible carapace.

†*Family Peltephilidae.* Upper Oligocene to lower Pliocene (questionably) in South America. Teeth $\frac{7}{7}$; short, broad skull; hornlike structures on the head armour.

†*Family Glyptodontidae* (glyptodonts). Mid-Eocene to Pleistocene in South America; upper Pliocene to Pleistocene in North America. Fusion of bony plates on the back into a solid turtle-like carapace; teeth $\frac{8}{8}$ with 3-lobed pattern.

Infraorder Pilosa

Incomplete zygomatic arch (bony ridge below eye); 4 or 5 simple cheek teeth; expanded acromion, which unites with the coracoid region.

†*Family Orophodontidae.* Oligocene in South America. Bilophodont teeth, $\frac{5}{4}$, made of a mass of compacted dentine; highly specialized astragalus (ankle bone). Small to medium body size.

†*Family Megalonychidae* (ground sloths). Upper Pliocene to Pleistocene in South America; middle Pliocene to Pleistocene in North America; Pleistocene in West Indies. Ground sloths. Large canine-like teeth in the upper and lower jaws and no spoutlike terminus on the lower jaw.

†*Family Megatheriidae* (ground sloths). Upper Oligocene to Pleistocene in South America; Pleistocene in North America. Partial development of upper "canines"; spoutlike terminus on the lower jaw.

†*Family Mylodontidae* (ground sloths). Upper Oligocene to Pleistocene in South America; Pleistocene in North America. Distinguished by a partial development of upper "canines," triangular-shaped cheek teeth, and a hindfoot with the 1st toe reduced.

Family Bradypodidae (tree sloths). South and Central America; tropical forests; no fossil record. Dental formula of $\frac{5}{4.5}$; 6 to 9 cervical vertebrae.

Infraorder Vermilingua

Toothless and hairy edentates.

Family Myrmecophagidae (anteaters). Lower Miocene to Recent in South America; Recent in Central America; tropical forests and savannas. Teeth lacking; tubular mouth with a small terminal opening.

Critical appraisal. The edentates are a natural assemblage of several seemingly rather diverse groups. Given that the oldest known representatives of the palaeanodonts and the xenarthrans are too advanced for either to be ancestral to the other, it would be of great interest to find the common ancestor of both groups, as well as the insectivore precursor of this ancestor. To date, no good fossil candidates have been unearthed in either category.

Authorities disagree on the relationships and degree of distinctness of some groups of fossil edentates. The palaeanodonts are believed by some to be closer to the Old World pangolins (order Pholidota) than to the New World edentates. The armadillo-like *Pseudophorodon* from the upper Oligocene of South America is sometimes placed in its own family but is considered here to belong in the Peltiphilidae. The palaeopeltids, known especially from polygonal plaques of bone, are considered by some to have been aberrant glyptodonts but are considered by Romer to have been the distinct superfamily Palaeopeltoidea. The unusual Miocene ground sloth *Entelops,* due to the presence of teeth on the premaxillary bone, is sometimes separated from the Mylodontidae as sole member of the family Entelopsidae.

Relationship of sloths and anteaters

The degree of relatedness of sloths and anteaters has been the subject of disagreement. It has been shown that despite certain adaptations of the front limbs and digestive tract for insect eating, anteaters are close to the most primitive xenarthran edentates in many characteristics of the forefeet and hindfeet. Hofstetter has suggested that the evolutionary lines that produced the modern Myrmecophagidae and Bradypodidae separated early and that, therefore, the two families are not closely related but should be placed in separate suborders.

The origin of the tree sloths and the relationship of the currently recognized two genera is a question of some

interest. As mentioned above, the two genera may be independently derived from two extinct families of ground sloths. *Choloepus* shows affinities to the megalonychids and *Bradypus* is more like the megatheriids.

At one time, the Old World scaly anteaters or pangolins (order Pholidota) and the South African aardvarks (order Tubulidentata) were included in the Edentata. The adaptations for an insectivorous diet that these two groups seem to have in common with the New World anteaters and the armadillos are most assuredly a result of convergent evolution, and not a reflection of any close phylogenetic relationship. (J.C.B.)

Lagomorpha (rabbits, hares, pikas)

The terrestrial mammals of the order Lagomorpha include the relatively well known rabbits and hares and also the less frequently encountered pikas, or mouse-hares. Rabbits and hares characteristically have long ears, a short tail, and strong hindlimbs that provide a bounding locomotion. In contrast, pikas have shorter, rounded ears, no external tail, and less well developed hindlimbs associated with scampering locomotion.

GENERAL FEATURES

Wild lagomorphs are small to small-medium sized, ranging from the smallest pikas, about 150 millimetres (5.9 inches) in length and 100 grams (3.5 ounces) in weight, to the largest hares, 700 millimetres (27.6 inches) and 4.5 kilograms (10 pounds). Wild rabbits range between pikas and hares in size, while some varieties of domestic rabbit may reach up to seven kilograms in weight.

Nearly worldwide in distribution, lagomorphs are absent only from most of Southeast Asian islands, Australia, New Zealand, Madagascar, southern South America, and Antarctica. Man has introduced rabbits and hares into areas outside their original ranges for purposes of sport and to provide a readily available food supply. The natural range of the Old World or European rabbit, *Oryctolagus cuniculus,* appears to have been southwestern Europe and North Africa, but in Roman times this species was often introduced to islands, where it flourished, and spread north and east, partly following man's agricultural activities, reaching Britain in Norman times. More recently *Oryctolagus* was introduced for sport and food into New Zealand and Australia, where the absence of native predators and other factors soon led to its becoming an agricultural pest and a threat to some of the native fauna. In North America several species of the cottontail rabbit, *Sylvilagus,* have shown an increase in abundance and a broadening of range in areas disturbed by man.

Economic importance of lagomorphs

Wild lagomorphs are popular with hunters for sport as well as for food and fur. Domestic rabbits, all descendants of the Old World rabbit, *O. cuniculus,* are raised for meat and skins, the latter of which are used both as pelts and for making felt. In addition, these rabbits are important laboratory animals for biological and medical research. They make good and relatively undemanding pets. Their attractive appearance and quiet manners have made them favourite characters with children's storytellers. On the other hand, wild rabbits and hares locally may become pests, depleting vegetation available to domestic grazers or damaging young trees and orchards, especially in winter. The problem becomes especially acute in areas where man has removed the natural predators of lagomorphs. Attempts have been made to control Old World rabbit populations in western Europe and Australia by introducing the virus disease myxomatosis, which exists naturally in populations of certain South American rabbits of the genus *Sylvilagus,* but in some areas the rabbits have developed resistance to the disease. Use of this virus in controlling rabbit numbers has been fought by some humane organizations because of the distressing symptoms of the disease. Pikas, usually found in areas remote from human activity, have little economic importance for man.

NATURAL HISTORY

Rabbits and hares occupy a wide variety of habitats, including grassland, desert, forest, marsh, brushland, and tundra. They are found in mountains up to 4,900 metres above sea level. Rabbits frequent areas where cover is readily accessible and do not venture far from it. Some rabbits dig their own burrows; others use burrows dug by other mammals. Although most rabbits are solitary, some live in pairs. The Old World rabbit, *O. cuniculus,* usually is gregarious, forming breeding colonies called warrens, with extensive burrow systems. The Mexican volcano rabbit, *Romerolagus,* maintains runways through dense vegetation. Although rabbits are typically terrestrial, two North American species of *Sylvilagus,* the marsh rabbit and the swamp rabbit, take readily to water when necessary. Hares tend to live in more open areas than do rabbits. They are solitary and do not use burrows but rest and take shelter in shallow depressions made in soil or vegetation. Pikas occur in northern steppes, semideserts, some forests and scrub thickets of Asia, and in rocky terrain, especially on rock debris on slopes; they range in mountainous areas to 6,000 metres (19,700 feet) in the southeastern U.S.S.R., northern and central Asia, and western North America. Steppe-dwelling pikas construct and inhabit burrows in flat areas. Rock-dwelling pikas usually seek shelter among rocks and boulders. Some pikas are colonial, whereas others are solitary and may exhibit some degree of territorial behaviour.

Vocalizations

Rabbits and hares are customarily nonvocal, making cries or screams only when frightened or injured. One rabbit, the South African red "hare" *Pronolagus,* is known to utter a warning call. Pikas are highly vocal, having a whistle or bark and a chattering call, reflected in such common names as whistling hare and piping hare. The voice is used in giving alarm signals and in maintaining territorial boundaries. Mountain hikers or climbers frequently hear their calls and may see these attractive little creatures sunning themselves on rocks and going about their daily activities.

Rabbits and hares are active mainly from dusk to after dawn, whereas most pikas are active during the day. All lagomorphs are plant eaters, grasses and soft vegetation being their dietary staples, but they may also eat twigs and bark, especially in winter. Arctic hares are known to break through crusted snow with their forelimbs to feed on the buried vegetation. Rabbits and hares are not known to store food, but pikas store and cure piles of vegetation for use during the winter. Some pikas cut grasses, carry them to the pile, and allow the vegetation to dry before adding more, whereas other pikas actually turn the contents of a pile to aid in the curing. Pikas that live among rocks first pile the vegetation in the sun and open air before moving it under the shelter of rocks for the winter. Steppe-dwelling pikas pile their hay in the open. By contrast, however, other pikas may store hay not in piles in the open but in crevices in trees, stumps, and among rocks. Some of the Asian pikas unwittingly supply hay for domestic livestock during the winter.

Some lagomorphs regularly reingest fecal pellets (coprophagy). Two kinds of pellets are produced: dry and hard, and soft and moist. The latter kind, which appear to contain vitamins and metabolic products, are eaten, often directly from the anus. The nutritional effect of this practice has been compared to that of rumination among cows.

Rabbits and hares differ from one another in the condition of their young at birth. Young rabbits are born and cared for in a nest; they are naked, blind, and relatively helpless at birth. Hares are born in the open; they are furred, have open eyes, and can run shortly after birth. Confusion of the terms "rabbit" and "hare" in common parlance is shown by such popular names as "jack rabbit" for a hare and "Belgian hare" for a variety of domestic rabbit.

Reproductive potential

The reproductive abilities of rabbits are proverbial. Rabbits and hares usually produce several litters during each breeding season, two or three litters being common among hares and three to six among rabbits. The gestation period ranges from about 28 days in rabbits to 47 in hares, and litter size is usually between two and eight. A female cottontail rabbit from an early spring litter can produce offspring by the end of the summer at age 10 weeks. In contrast to most mammals, female rabbits and hares are

usually larger than males. Reproductive cycles of pikas are less well known than those of rabbits and hares. Their breeding season occurs in late spring and summer, with a gestation period of approximately 30 days. Litter size normally ranges from two to six, and two or three litters may be born each year. In some pikas, the newborn young are only lightly furred, have eyes and ears closed, but can move about after about eight days.

Lagomorphs, especially hares and rabbits, are important elements in many food chains. They are preyed upon by a variety of carnivores, both mammals and birds, who rely upon them as dietary staples. Wolves, foxes, bobcats, weasels, predatory hawks, and owls all take their toll. More northerly species of rabbits and hares frequently exhibit cyclic fluctuations in numbers, which are reflected in the cycles of the carnivores dependent upon them. The snowshoe hare–snowy owl cycle is one of the prominent ones. Pikas are preyed upon by various carnivores, among which weasels and birds of prey are probably most important.

Externally most rabbits and hares are not dissimilar from one another, the main variation occurring in length of ear and strength of hindlimb. The coat is usually brownish or reddish brown above and lighter to white below. An exception to this coloration is found in the Arctic hare, in which the winter coat is pure white with only the eartips dark. In more northerly representatives of the Arctic hare, individuals retain this colour pattern all year, whereas more southerly forms of it may become grayish or buff-coloured. In other northern hares, the varying hare (*Lepus timidus*) of Eurasia and the snowshoe hare (*L. americanus*) of North America, the winter coat is white and the summer coat brown. Some varying hares in Ireland and snowshoe hares from the Pacific Coast do not become white in winter. The most unusual coloration in a modern rabbit occurs in the Sumatran rabbit *Nesolagus,* which has brown stripes on its buff-gray pelage and a bright red rump and tail. Pikas are usually brownish to reddish above and somewhat lighter below.

Lagomorphs illustrate adaptation to particular habitats in an interesting gradational series of features for detection of, and escape from, enemies. At one end of the series are the long-eared, wary, and alert hares, inhabitants of open country, which detect enemies at considerable distance and rely for escape on their strengthened hindlimbs and bounding locomotion that takes them up to 80 kilometres (50 miles) per hour. Rabbits, shorter eared and with weaker hindlimbs, do not react to such distant threats and are scamperers or bounders who do not venture far from cover. They are quick over short distances but lack the endurance of hares. With still shorter ears and relatively weaker hindlimbs, pikas live in areas in which a quick scamper takes them the short distance that usually separates them from secure cover.

FORM AND FUNCTION

Dentition

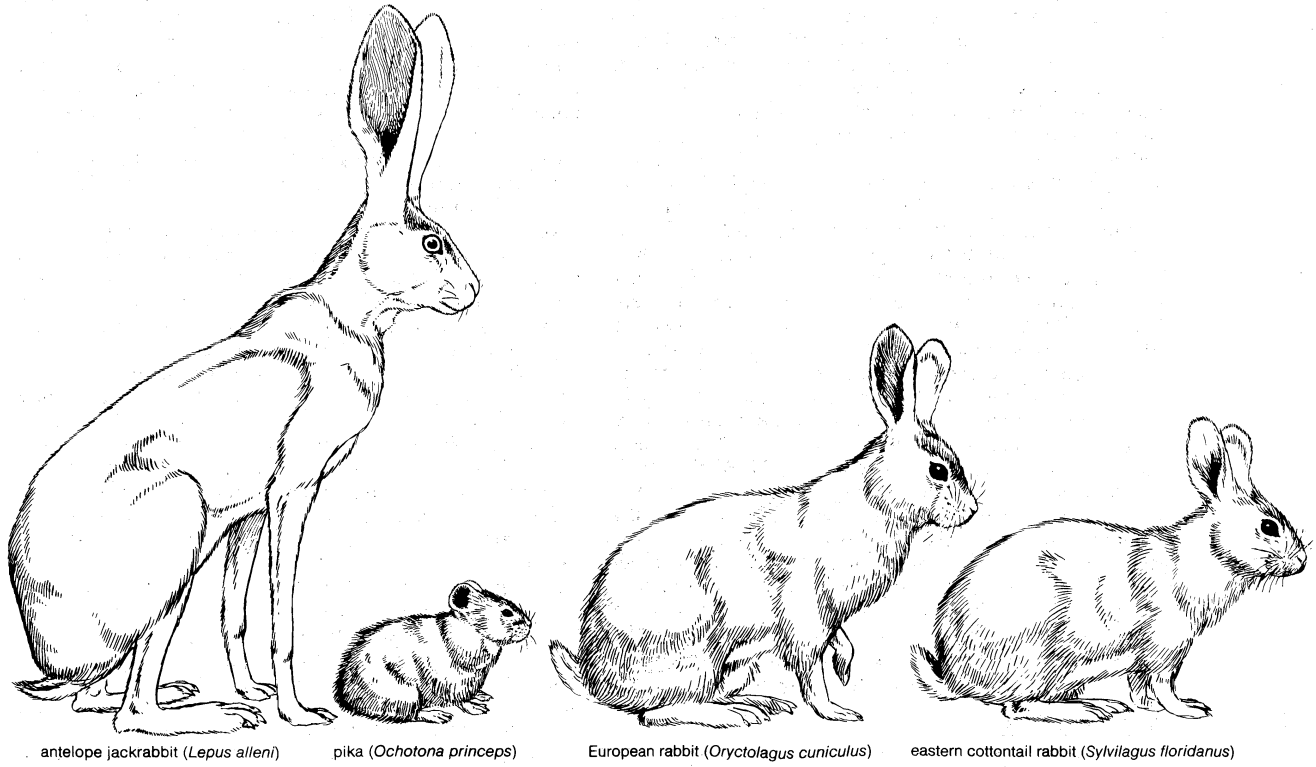
Lagomorphs are well adapted to a herbivorous diet. Like rodents, they have well-developed incisors that grow continuously from the roots, while they are worn down at the cutting edges. These teeth are extremely effective for severing plant stems and for gnawing on bark. The very mobile lips each consist of two lobes, which meet behind the chisel-like incisors when the mouth is closed. There are no canine teeth, a gap (diastema) showing where they might have been. Cheek teeth farther back in the jaw are also ever growing, wearing away as they grind abrasive vegetation. The upper tooth rows are more widely separated than the lower rows, and chewing is done with a transverse movement. The chewing muscles of the jaws, though strong, are less well developed than in rodents.

Vegetation passes through the long small intestine, which has a spiral valve (like a wood screw), providing a much greater surface area for digestion of food and absorption of nutriment. A large pouch, or cecum, located at the point of attachment of the large intestine contains bacteria that aid digestion and that produce the nutritional soft fecal pellets.

Skeletal adaptations for speed and agility are evident and are discussed as taxonomic characters in the *Annotated classifications* section, below. The bones of the hindlimb are fused where they move against the anklebone (calcaneum), affording strength and leverage employed in bounding or scampering locomotion. Lagomorphs move about as if they walked on their toes (digitigrade locomotion). There are five digits on the forefoot and four or five on the hindfoot.

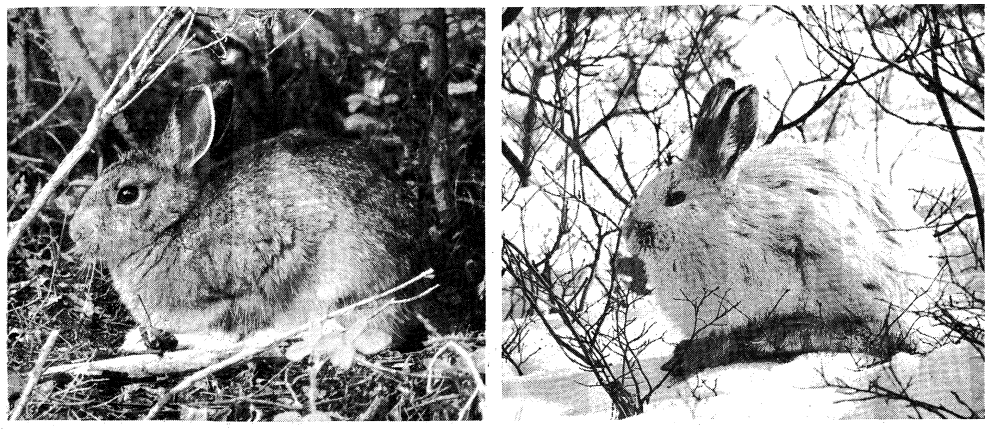
EVOLUTION, PALEONTOLOGY, AND CLASSIFICATION

The order Lagomorpha appears to have originated in northern Asia, probably by the end of the Paleocene (about 55,000,000 years ago), and has been relatively



antelope jackrabbit (*Lepus alleni*) pika (*Ochotona princeps*) European rabbit (*Oryctolagus cuniculus*) eastern cottontail rabbit (*Sylvilagus floridanus*)

Figure 28: Diversity among lagomorphs.



The summer fur (left) of the snowshoe hare (*Lepus americanus*) blends with the browns of woodland and prairie. A mottled effect, created by white hairs gradually replacing the brown, is similar to patches of new snow, while the white winter coat (right) is completely camouflaged in a snowy landscape.

The National Audubon Society Collection/Photo Researchers, (left) Charlie J. Ott, (right) Leonard Lee Rue

stable morphologically throughout the approximately 40,000,000 years since the end of the Eocene, when its fossil record first becomes well documented. Lagomorphs seem to have established early in their history the patterns of dental and skeletal development that were to change, for the most part, only by gradual advances in the basic pattern. Within the order several trends can be observed from more primitive to more advanced lagomorphs: development of ever-growing cheek teeth, simplification of the pattern of the occlusal surfaces of the upper cheek teeth, and shortening of the shaft of the lower incisor where it passes back into the jaw. When dealing with fossil forms, often known only from fragments of jaws and teeth, the most useful characters that differentiate taxa are ordinarily found in the structure of the anterior two upper, and anterior lower, premolars. Presence or absence of the last upper molar may also be important, as may the pattern of the upper cheek teeth. When preserved in fossils, structure of palate and of lower jaw also provide characters that assist in determining affinities.

Evolutionary trends

The family of rabbits and hares (Leporidae) entered North America by the end of the Eocene and underwent most of its Middle Tertiary development there. By the Pliocene (about 7,000,000 years ago) it had become reestablished in Asia and had moved also into Europe. The leporids now extend throughout those ranges and down to the tip of South Africa and as far south as northern Argentina in South America.

The pika family (Ochotonidae) spread from Asia to Europe, where they developed into a number of types in the Middle Tertiary. One of these lines persisted until the late Pleistocene or early modern times (about 1,000,000 years ago) on Corsica and Sardinia. Other ochotonid branches reached Africa and North America in the Middle Tertiary. The extant genus *Ochotona* appeared in Asia in the Pliocene and spread from there, reaching western Europe and eastern North America in the Pleistocene. The current range of the genus represents a considerable reduction from that during the Pleistocene.

Distinguishing taxonomic features. Within the order Lagomorpha living members of its two extant families are clearly separable on morphological grounds, the differences being in skull, jaw, and dental structures, as well as in postcranial morphology.

Morphological trends in the rabbits and hares include increase in arching of the skull, correlated with the development of bounding locomotion and of a relatively upright posture of the head; reduction in the ability to perform lateral movements, caused partly by fusion of bones and partly by modifications related to specialization for anteroposterior movements; strengthening of the hindlimbs and pelvic girdle; and elongation of the limbs distally. Pikas lack these skeletal modifications and are adapted rather for a scampering locomotion.

Annotated classification. In older classifications lagomorphs were usually placed in the order Rodentia. Both rodents and lagomorphs are gliriform mammals; that is, they have enlarged, ever-growing incisor teeth that function as effective tools for gnawing. Under this scheme, the lagomorphs were the suborder Duplicidentata, and the true rodents the suborder Simplicidentata. Early workers, however, recognized many features separating these suborders. For one, there are two pairs of upper incisors in lagomorphs (thus the name Duplicidentata) as opposed to one pair of upper incisors in rodents (Simplicidentata). It is now known that lagomorphs and rodents have long separate histories and that even their earliest representatives did not closely resemble one another. Accordingly, it is more accurate and currently generally accepted to classify the two groups as distinct orders, Lagomorpha and Rodentia. The following classification is the one in general use.

ORDER LAGOMORPHA

Small to medium-sized mammals with two pairs of upper incisor teeth, one pair large, ever-growing, and grooved, in front of a pair of small, peglike incisors, and one pair evergrowing lower incisors. In skull the jawbone, or maxilla, pierced or with open lacework laterally; palatal bone with large, deep-cut openings and relatively short bridge. Of muscles used in chewing, masseter muscles not greatly enlarged, temporalis muscle reduced, pterygoid muscles well developed. Twelve extant and approximately 35 known extinct genera are included in three families, one extinct (†).

†Family Eurymylidae

Late Paleocene and (?) Eocene. This extinct family of two genera known only from northern Asia may or may not be closely allied to the other two families of lagomorph. Several anatomical features, such as the presence of only two pairs of upper premolars, show that it was not directly ancestral to later lagomorphs, but it may have been near a group from which later lagomorphs originated.

Family Leporidae (rabbits and hares)

Late Eocene to Recent. Almost worldwide. Distinguished by elongated ears, hindlegs much longer and stronger than forelimbs, often greatly strengthened, and short external tail. Most anterior premolars unlike others in pattern, the following two premolars similar; usually three upper and three lower molars in each jaw, reduced to two upper molars in *Pentalagus;* shafts of upper cheek teeth extend upward into socket in bony capsule (orbit) surrounding eye. Skull with jawbone having open lacework of bone laterally; bony process present over orbit. The palate has one pair of large deep-cut holes; bony palate with longer maxillary and shorter palatine components. Tendency toward arching of skull in profile, related to increasingly upright posture of head. Clavicle rudimentary. Strong, well-developed lumbar region of back, and strong pelvic girdle for articulation with hindlimb and attachment of limb muscles. Five digits on forefoot and on hindfoot.

The subfamily Palaeolaginae is an extinct group that includes primitive Old and New World forms. Descendants of the Palaeolaginae were the extinct Archaeolaginae, including most North American Middle Tertiary forms as well as a group that reached Eurasia in the Pliocene. The most advanced leporids are in the subfamily Leporinae, which includes all living leporids (11 genera), including the American *Sylvilagus,* the Old World *Oryctolagus, Lepus* of the Northern Hemisphere,

Romerolagus of central Mexico, *Pentalagus* of the Ryukyu Islands south of Japan, *Bunolagus* and *Pronolagus* of South Africa. Other leporids showing some relatively primitive characters are *Nesolagus* of Sumatra and *Brachylagus* of the western United States. The subfamilial divisions are based partly on the pattern of the most anterior upper and lower premolars, which seem to be valid indicators of leporid relationships. Size 25 to 70 centimetres (9.8–27.6 inches).

Family Ochotonidae (pikas)

Oligocene to Recent. Asia and western North America. Distinguished by short, rounded ears, hindlegs slightly longer than forelimbs, no external tail. Three upper premolars exhibit three different patterns; two upper and three lower molars in each jaw in modern forms, ranges from three upper and lower molars in each jaw to two in each jaw in fossil forms. Shafts of upper teeth curve outward into arch of the zygomatic bone beneath eye socket. Skull with single large opening in each jawbone, other holes present in some fossil forms. Bony process over orbit absent in modern ochotonids but present in some fossils. On palate one or two pairs of deep-cut holes; bony palate with larger palatine component and smaller maxillary component. Skull typically more flattened than in leporids. Well-developed clavicle. Five digits on forefoot and four on hindfoot. One extant genus, *Ochotona,* and 14 or 15 extinct genera. Size 15 to 28.5 centimetres (5.9–11.2 inches).

Critical appraisal. Perhaps the most important unanswered question in lagomorph classification is that of affinities to other mammals. Part of the difficulty is that the fossil record from the Early Tertiary of Asia is incomplete for large parts of the Paleocene and Eocene. The Eurymylidae, known from some Paleocene and possibly Eocene deposits in Asia, have some characters similar to those later lagomorphs but other characters showing that they were not the direct ancestors of later forms. Other fossil mammals now known from the Paleocene of Asia may represent forms related to eurymylids. The evidence is incomplete, but perhaps eurymylids and these possibly allied forms were in a generally ancestral position to leporids and ochotonids. The wider affinities of this possibly ancestral group to other mammals are not yet clear. No transitional forms are known between the eurymylids and later lagomorphs, which had attained their distinctive characters by the late Eocene. Earlier attempts to arrive at the solution to the problem of lagomorph affinities often emphasized similarities between various primitive herbivorous animals, such as the condylarths, and primitive lagomorphs. These similarities, however, are probably due to the presence, in both groups, of morphological characters common to primitive mammals in general.

Relationships with other mammals

A further point causing some difficulty in studies on fossil leporids and ochotonids is the assignment of various primitive forms to one family or the other. The problem is especially acute when the fossil record is based only on a few teeth or jaw fragments. Morphologically these primitive forms seem to be intermediate between the two families.

Some differences of opinion exist in regard to the subfamilial classification of leporids. Some students of these animals have taken the step of subdividing leporids into several families. Other problems in the classification of modern leporids are matters of specific assignment of various populations. A sound subfamilial arrangement for the ochotonids has not yet been established. (M.R.D.)

Rodentia (rats, mice, beavers, squirrels, guinea pigs, capybaras)

The rodents, or Rodentia, are the most abundant order of mammals. At present, over a quarter of the families, 35 percent of the genera, and 50 percent of the species of living mammals are rodents. Probably an even higher percentage of individuals are rodents, for they tend to be small animals with dense populations. They are one of the few groups of animals that flourish in close association with men. Some, such as squirrels, live independently but fairly successfully near humans. Others, such as the house mouse (*Mus musculus*) and black and Norway rats (*Rattus rattus* and *R. norvegicus*), have adapted themselves to human civilization, and live everywhere that man does. These two rats (and the Polynesian rat, *Rattus exulans,* of Australia and Oceania) have travelled in ships and boats of all sizes, and have populated the entire habitable world, especially near human habitations.

GENERAL FEATURES

All rodents possess one pair of upper and one of lower incisors, growing throughout life, with the enamel restricted to a band on the front side of the teeth. Behind this is a large gap (diastema) followed by two to five cheek teeth. The jaw articulation is so arranged that when the cheek teeth are in use, the incisors do not meet, and vice versa. The incisors grow continuously, and must be worn off equally fast, or the whole gnawing mechanism is ruined. Because of the necessity to abrade these incisors, rodents spend a considerable amount of time gnawing hard objects.

Generally rodents are small. Some mice and dormice are among the smallest of living mammals, adults being as small as 75 millimetres (three inches) long, including the tail, and weighing as little as 20 grams (0.7 ounce). The largest living rodent is the South American capybara (*Hydrochoerus hydrochoeris*), reaching over 1.3 metres (four feet) in length and as much as 50 kilograms (about 110 pounds) in weight. A fossil rodent recently described from Uruguay is reported to have had a skull as large as that of a bull and a body bulk as large as that of a wild boar.

Size range

Rodents are of major economic importance, primarily as consumers of the grains that are the basic foodstuff for

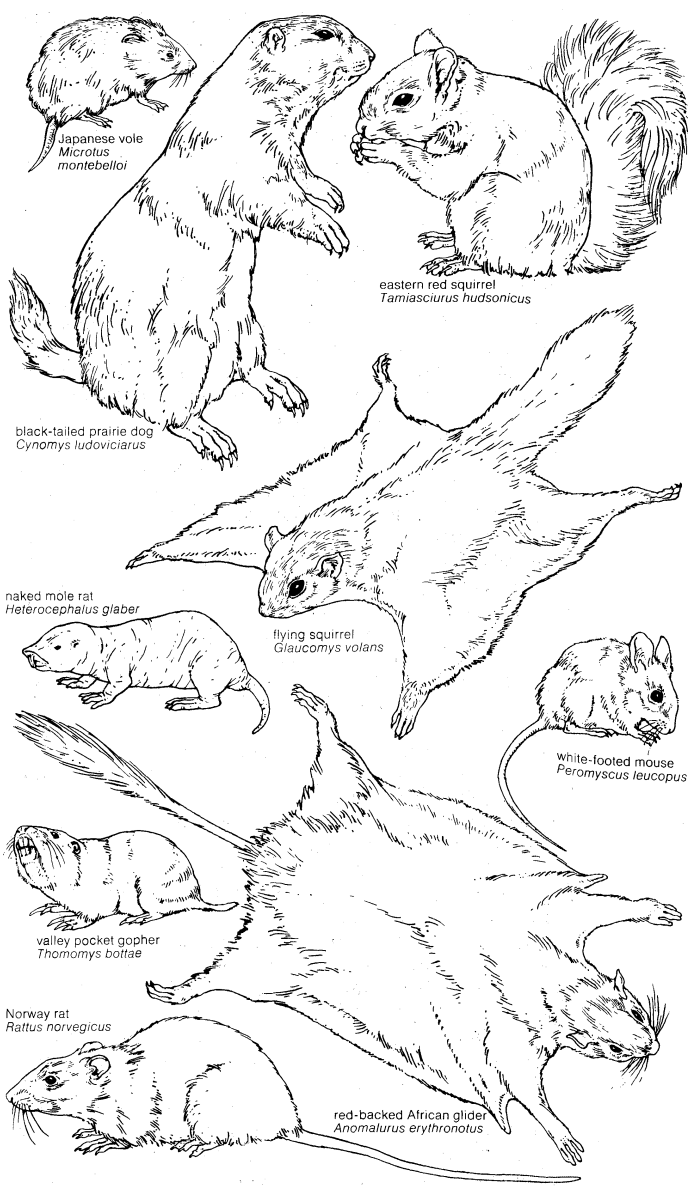


Figure 29: Range of body-plan variation of Rodentia.

man. It has been estimated that rats and mice destroy up to one-third of grain crops under conditions of heavy infestation. Burrowing rodents may damage root crops. The muskrat (*Ondatra zibethica*) and nutria (*Myocastor coypus*), introduced into Europe as fur sources, have escaped and spread over much of Europe between the Baltic and the Alps. Their burrows, particularly in canal banks, have been a major source of damage to the drainage system, most especially in The Netherlands. A number of rodents serve as reservoirs for human diseases, such as bubonic plague, tularemia, scrub typhus, and others. The plague that ravaged Europe during the mid-14th century was transmitted by fleas from rats to humans.

Several rodents (beaver, muskrat, chinchilla, nutria, squirrel) produce fur useful to man. All but beaver and squirrel have been domesticated for this purpose. Albino mice and rats, hamsters, and guinea pigs are widely used as laboratory animals for biological and medical purposes. Guinea pigs were domesticated by the Incas for food; a few kinds of rodent have been raised as pets.

Distribution of rodents

Rodents occur naturally in all parts of the land where there is an adequate food supply and are found in essentially all terrestrial habitats. They range from well above the Arctic Circle to the southern tips of Africa and South America, and were the only terrestrial placental mammals, other than bats, to reach Australia before the arrival of man. Many rodents have successfully adapted to difficult environments such as deserts. Many rodents have broad climatic tolerances, an example being the North American porcupine, which is found from the Arctic Circle to central Mexico and from the Atlantic to the Pacific. Most are quadrupedal scamperers, but they generally have much freedom of use of their forefeet in manipulating food; many are burrowers, spending most of their life underground; some are ricochetal, leaping on their hind legs; flying squirrels use skin membranes to glide from one tree to another; a few (beaver, muskrat, water vole, nutria) have become amphibious in habits, living in freshwater streams and ponds; and a number of South American rodents are cursorial (running) animals.

IMPORTANCE TO MAN

Destruction of crops and foodstuffs. The most important rodents, from the point of view of economic damage, are the Norway and black rats, with the house mouse close behind them. It has been frequently estimated that the rat population of the United States is approximately equal to the human population. A population of over 1,000 rats per acre (2,470 per hectare) on an Iowa farm has been reported. Population explosions of house mice occurred in the Central Valley of California in 1926–27 and 1941–42. During the former, the mouse population was estimated to have reached over 80,000 per acre (198,000 per hectare).

Remarkable population explosions of voles (*Microtus*) and wood mice (*Apodemus*) are well known in western Europe, recurring every few years. In France, from 1790 to 1935, there were at least 20 mouse plagues, some lasting several years. Estimates of abundance are highly inaccurate, but there are reports of 8,000 voles per acre (19,800 per hectare) and 15 to 20 vole burrows per square metre (13-17 per square yard), in peak conditions. There was a population explosion of voles and wood mice over much of Germany beginning in late summer of 1917 and lasting through 1918. Damage to crops was serious; clover was the favourite food; winter rye and wheat, sugar beets, and potatoes also were destroyed in many areas. During the winter the voles invaded barns and farm buildings, destroying all kinds of stored food. In 1932–35 voles and lemmings were in epidemic proportions in over 104,000 square kilometres (40,000 square miles) of the southern U.S.S.R. This plague resulted in extensive damage to wheat and other field crops and to orchards.

Rats will eat almost anything that humans eat. Perhaps the most serious damage is to the seeds of grain plants, both before and after harvesting. Grain stored on farms is often not only eaten by rats but also rendered unsuitable

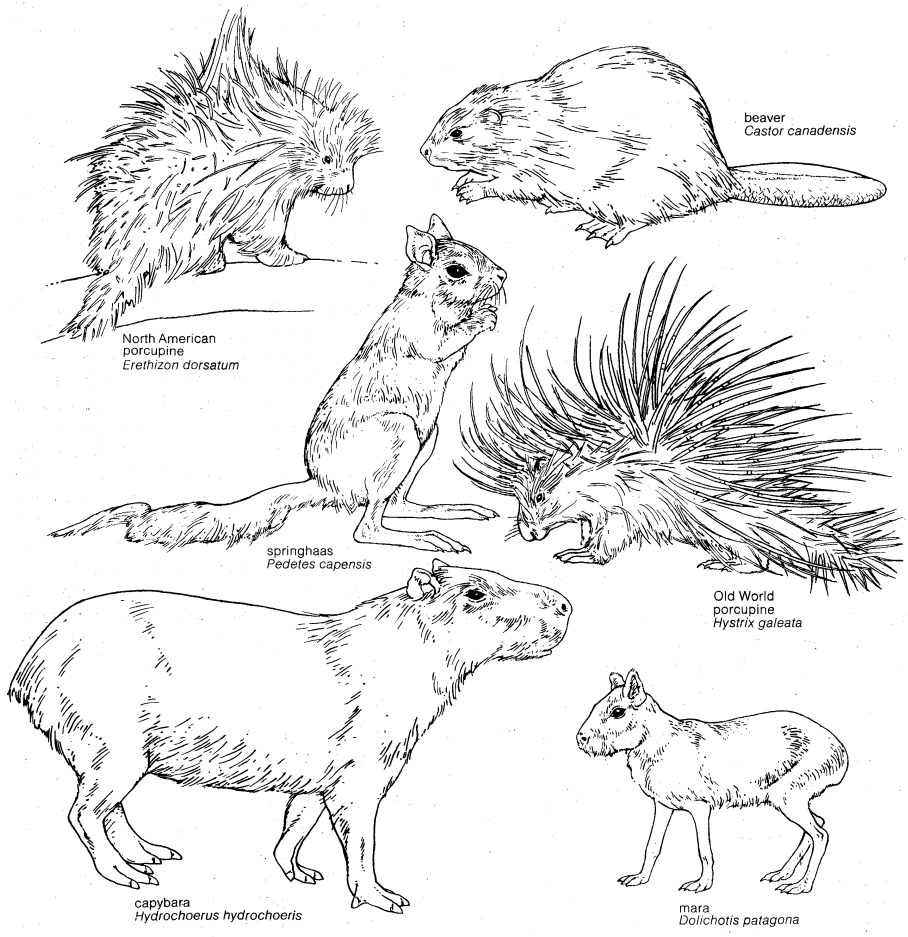


Figure 30: Range of body-plan variation of larger Rodentia.

for human consumption by being mixed with rat droppings. Food that has reached warehouses in cities is also eaten by rats, and here the excess of damage over the amount actually consumed is even greater. It is estimated that rats damage about twice as much grain as they eat. About 23 kilograms (50 pounds) of grain are required to support a rat for a year, so the total cost is about 70 kilograms (150 pounds) per rat per year. Rats sometimes demonstrate a fondness for animal food and have been known to kill several hundred baby chickens in a single night. Eggs are frequently eaten, and even full grown hens, baby pigs, and lambs may be killed. Rats will also follow a farmer in the planting season and dig up newly planted seeds.

Depredation by rats

Rats also do extensive damage in searching for food or in making nests. They gnaw holes to gain entrance to barns, warehouses, or houses, or through walls once they are inside. Clothing, upholstered furniture, and other textiles are gnawed to provide nest-building materials. In areas of cities where extensive quantities of garbage and other refuse are available, the rat populations may exceed the human populations, and rat infestation is one of the commonest complaints of slum dwellers. Such rats are very apt to bite sleeping humans, especially children, and fatal attacks on babies have been known to occur.

The house mouse has food preferences similar to those of the Norway rat but is not as abundant nor as large. Other rodents are much less likely to attack stored supplies of human food, but they do eat considerable quantities before harvesting. This is true of such forms as voles, field mice, and squirrels. The seeds of both wild and domesticated grasses are a major food source for large numbers of wild rodents. Because of the quantities of these that are eaten, such animals are often considered a major threat by grazing interests, but the benefit that rodents give by collecting seeds into underground store houses, where they may later sprout and grow into new plants, certainly comes close to balancing any damage they cause to man's interest.

The role of rats in the "Black Death"

Transmission of diseases. *Plague.* Rodents serve as reservoirs for a number of diseases that may be transmitted to humans by arthropod agents. The most devastating of these is bubonic plague. This disease is fundamentally a disease of rodents, especially rats, transmitted from one rodent to another by an intermediate host, the flea, which also serves to transmit the disease to humans. An epidemic (called a "pandemic" because of the totality of infection in the human population) that seems to have been plague spread over Europe in the 6th century AD. If this epidemic was indeed plague, it must have involved an unusual rodent host because it antedated the arrival of *Rattus* in Europe. The epidemic known as the Black Death originated in Mesopotamia about the middle of the 11th century, and spread to Europe, particularly in the 14th century, being accompanied by the spread of rats throughout the Continent. It has been estimated that 25,000,000 people died of that pandemic of the plague in Europe. The latest pandemic originated in southwestern China in the late 19th century and was spread all over the world by the rat populations of oceangoing ships. Plague is controlled by controlling rat populations, particularly those on ships, and preventing shipborne rats from reaching land. Although plague epidemics have been brought under control, there still remain numerous foci of infection.

The plague bacillus (*Pasteurella pestis*) can infect a variety of other rodents, and foci have been established among native rodents other than *Rattus* in many parts of the world. This disease is referred to as "sylvatic plague" to distinguish it from the basically urban occurrence of ratborne plague. Over 80 species of ground-living and burrowing rodents are known to be involved in sylvatic plague, including ground squirrels, some cricetids (*e.g.*, voles, lemmings, muskrats), several murids (*e.g.*, Old World mice), and a few others, including guinea pigs.

Tularemia. Tularemia is primarily a disease of lagomorphs (rabbits and hares) and secondarily of rodents. It has been reported from all parts of the United States, and seems to have spread from there to many other parts of the world. Among rodents, it is carried by ground squirrels, tree squirrels, prairie dogs, chipmunks, muskrats, and beavers, and is transmitted largely by ticks.

Rickettsial diseases. Three rickettsial diseases—murine typhus, Rocky Mountain spotted fever, and tsutsugamushi disease—involve rodents as reservoirs. Murine typhus, which is much less severe than epidemic typhus, is transmitted to man by fleas, primarily from *Rattus*. It is probably almost universal in tropical and subtropical areas. Rocky Mountain spotted fever, which apparently originated in the northwestern United States, has spread over much of that country, and very similar if not identical diseases are found in Mexico, South America, and Africa. It is transmitted to humans by tick bites, and is presumably endemic in a considerable number of rodents as well as other animals. The cottontail rabbit is believed to be a major reservoir animal. Tsutsugamushi disease, or scrub typhus, is transmitted to humans by the bite of a rat mite. It is found in the East Indies and Southeast Asia. The reservoir seems to be primarily rodents of the genus *Rattus*. A variety of other diseases of lesser importance are transmitted to humans from a reservoir in rodents, largely *Rattus*, by rat bites, ticks, or in other ways.

Damage by burrowing and gnawing. Rodents have been accused, in many parts of the world, of damage to agricultural interests because of their burrows. Cattlemen in the western United States supported campaigns to poison prairie dogs (*Cynomys*) because horses or cattle might (and occasionally did) break their legs in prairie-dog holes. But perhaps the most striking case of damage from rodent burrows has been caused by the American muskrat, introduced into European fur farms, from which it escaped as early as 1905, spreading widely over much of Europe, from Great Britain and Brittany to the Ukraine. It burrows in the banks of streams, and has caused damage especially by making burrows in the banks of drainage ditches, canals, and in dikes.

The ever-growing incisors of rodents need to be used regularly on hard substances to wear them off. This need results in the gnawing of lead pipes or telephone transmission cables, or the making of holes in boxes, walls, and stored inedible items.

Rodents as valuable furbearers

Benefits derived from rodents. *Trapping for furs.* The beaver (*Castor canadensis*) was extremely important in the development of the western United States and Canada, the trappers of the first two-thirds of the 19th century being primarily interested in beaver skins. Most of the initial exploration of the West was performed by beaver trappers. The beaver skin became the basic unit of currency over much of the western United States. The trapping was so extensive and effective that beavers were exterminated over much of the area, but they are making a slow recovery under current conditions of protection. The fur of the chinchilla (*Chinchilla laniger*) was of such value that these rodents were almost exterminated in Argentina. Muskrat and nutria have also been extensively hunted for their fur.

Rats in the laboratory. Among the best laboratory mammals, for all types of biological, medical, and psychological investigation and for testing of new drugs, are the laboratory rats and mice, normally albino strains of wild species of the Norway rat and the house mouse. Guinea pigs (*Cavia cobaya*) are also widely used. There are a number of other laboratory rodents of lesser importance, of which the Syrian or golden hamster (*Mesocricetus auratus*) is the most widely used. Some strains of rats are nearly unique among nonhuman animals in being susceptible to dental caries, and most of the experimental work on tooth decay has been performed on them.

ECOLOGY AND NICHE RELATIONSHIPS

Habitat and locomotion. The most typical members of the order are the small, ground-living rodents, of the type generally called rats or mice, similar to what was probably the ancestral condition for the order. While these generally stay on the ground, most can climb shrubs and bushes with ease and many climb trees; among the smallest rodents, harvest mice climb stalks of wheat to reach the seeds on which they feed. This central generalized adaptive type of rodent, the scamperer with limited burrowing or arboreal adaptations, includes the family Muridae (Old World rats

The basic scampering rodent

and mice, now worldwide), the family Cricetidae (field mice, wood rats, voles, lemmings, muskrats, to name a few; found in most parts of the world other than Australia), some Sciuridae (the ground squirrels and chipmunks), the nonleaping members of the New World family Heteromyidae (pocket mice); some South American rodents (degus, spiny rats, and chinchilla rats), and the South African rock rats (Petromuridae). Although there are many adaptive variants among the scampering rodents, there is a strong tendency for them to inhabit either open country or limited areas of woodland. Rodents of this type are found over essentially the entire land surface of the world. In most cases individual or family territories are set up, and intruders are successfully kept at bay by a variety of threatening and defiant postures. Normally there is no actual combat. The territories are, of course, of varying size, related to the available food supplies and the size of the animals. Individual ground squirrels have been observed over a territory of 2,500 square metres (about 0.6 acre). Changes with time in the territories of chipmunks (*Eutamias*) indicate rather striking seasonal changes, with an increase in the area toward the end of the foraging season (perhaps to allow greater accumulation of food), and very little shift of the territory from one year to another. With few exceptions, these terrestrial rodents make burrows near the centre of the territory where they may spend as much as half of the time, where the young are born, and where the adults hibernate, in forms where hibernation occurs. Normally more young are produced than there is room for and as a result there is extensive migration of subadult or young adult individuals. For most rodents, the lengths of these migrations are unknown, but it has been shown that muskrats will migrate as much as 30 kilometers (19 miles), including several kilometres cross-country, over a period of one to two weeks, looking for suitable new stream or swamp sites.

From occasional tree climbing it is a relatively short evolutionary step to an arboreal habitat, such as that occupied by the tree squirrels, or, with an increase in size and decrease in rapidity of movement, by the New World porcupines. Tree squirrels are, essentially, scamperers that have developed the capability of holding onto bark with their claws. They are as much at home on the ground as in trees, and show no special anatomical modifications for tree life. Their nests may be either in hollow trees or other sheltered places (*e.g.,* barns) or built of twigs and leaves among the smaller branches of trees. In Europe and adjacent parts of North Africa and Asia, the dormice (Gliridae) have similar habitat but spend an even larger part of their lives in the trees. Many murids and cricetids (rats and mice, in the broad sense) are about as fully arboreal in their habits as are the tree squirrels.

Gliding rodents

Scampering arboreal animals are faced with the necessity, from time to time, of moving rapidly from one tree to another, and do so by leaping. It is generally agreed that gliding forms have evolved their skin fold (patagium), which assists them in gliding, as a direct adaptation to this type of habitat. Patagia have evolved, probably several times independently, among the Sciuridae (squirrels) of both the Old World (*Pteromys* and others) and North America (*Glaucomys*) and in the quite distinct and unrelated African family of scaly-tailed squirrels (Anomaluridae). The patagium of the anomalurids is supported at the anterior end by a long, slender cartilage, which has apparently evolved from the olecranon process of the ulna (the posterior bone of the forearm) at the elbow.

Larger, heavier arboreal forms such as the New World porcupines move slowly along the trunks and main branches of trees. One of these, in South and Central America, has developed a prehensile (grasping) tail that assists in climbing.

Burrowing rodents

A large proportion of the small, ground-living rodents construct underground nests, where the young are born and reared. Often there are adjacent chambers for food storage. In many cases two or more exits provide an opportunity for escape from predators. From ancestors such as this, many lines of rodents have evolved into habitual burrowers that make extensive subterranean tunnels and rarely or never come to the surface, foraging, feeding, mating, and raising their families underground. Among the numerous varieties of burrowing rodents are the prairie dogs, colonies of which were formerly abundant in the western United States. The mouth of the burrow is surrounded by a volcano-shaped mound of dirt, on which a prairie dog frequently sits upright, alert for potential enemies. The animal forages on the ground surface, returning to the mound or to the burrow to eat. Voles (*Microtus*) may make lengthy runways, 50 to 75 millimetres (two or three inches) below the surface, arching up the ground surface over them. The pocket gophers (*Geomys* and *Thomomys*) are even more effective burrowers, throwing up mounds 30 centimetres (12 inches) or more across and spending most of their life beneath the ground. They use the incisors as well as the front legs in digging. During times of heavy snow, they may burrow through snow, feeding on plants above ground. Similar types of burrowing occur in the Old World mole rats (Spalacidae). An extreme form of burrowing is found in the African mole rats or blesmols (Bathyergidae), in which the incisors extend forward, with a fold of skin closing the mouth behind them, permitting their use as the primary digging tools, having replaced the forelimbs in this respect. The burrows of a single blesmol form a crisscross of tunnels, at several levels below ground surface, that may occupy an area a thousand or more square metres. All of these burrowing forms have short, heavy limbs used in digging, short tails, and small eyes. In one genus of blesmol, *Heliophobius,* the eyelids have grown together so that the eyes do not form images. These animals remain permanently below ground. Many burrowing rodents are solitary, but a number of them live in colonies, unlike other members of the order.

Relatively few rodents have become large animals. Several of the South American caviomorphs have done so, however, and have evolved as quadrupedal terrestrial cursorial animals. Usually these have the claws modified into or toward hooves, reducing the number of toes on each foot to four or occasionally three, and tending to make each foot symmetrical about the median plane. Among these are the capybara (the largest living rodent), nutria, paca (*Cuniculus paca*), and agouti. The guinea pig is a smaller relative with the same adaptations. Such animals normally escape predators by running away, although both capybara and nutria will take to the water. The Old World porcupines are slow-moving terrestrial quadrupeds that are protected, as are the New World porcupines, by the quills, which are elongate, pointed hairs.

Giant rodents

Many rodents have become bipedal jumpers, escaping their enemies with long leaps. These forms include the kangaroo rats (*Dipodomys* and *Microdipodops*) of western North America, the jumping mice (*Zapus* and *Napaeozapus*) of eastern North America and eastern Asia, the jerboas (Dipodidae) and gerbils (*Gerbillus*) of North Africa and Southeast Asia, the saltatorial dormouse (*Selevinia*) of Kazakstan, U.S.S.R., the chinchilla of the southern Andes, and the springhaas or Cape jumping hare (*Pedetes*) of South Africa. All of these are small, except *Pedetes,* which stands 30 centimetres (12 inches) tall. Most of the saltatorial rodents have very highly enlarged tympanic bullae—the bony covering of the middle ear—which is not true of other rodents. It is probable that this is an adaptation to life in a desert, where an amplification of the noise made by approaching predators may be of great importance. Certainly the large middle-ear cavity greatly increases the amplification of sounds. The fact that the bullae are particularly large in leaping forms suggests further that there may be some relationship between the size of the bulla and the balancing and stabilizing activities of the ears.

Although a number of rodents are at least partially aquatic, special adaptations are rare, being found only in the beavers. Beavers have large hind legs, with strong webs between the toes and an enlarged first digit, as long as the other toes, making very efficient paddles. Swimming is carried out primarily by the feet, but the broad flat tail may be used on occasion for sculling. A number of other rodents that are semiaquatic have webbed feet and soft, dense underfur but few other aquatic adaptations. These semiaquatic forms include muskrats, water voles,

capybaras, nutrias, South American fish-eating rats, and the water rats of Australia and New Guinea.

Seed-eating rodents

Food habits. As a whole rodents are herbivorous, although most of them will eat animal food on occasion. The diet includes a wide variety of plant foods, although seeds are the favourite item. Many are particularly fond of the seeds of the various grasses, both wild and domestic. These may be eaten on the spot, but large quantities are usually carried home in the cheek pouches that are common among rodents. Seeds carried home may be eaten in the safety of the burrow or stored for later use. Such storage is an important item to the plants concerned, as it places the seeds in a favourable location for sprouting, if the rodent either forgets about them or himself becomes food for a predator. Many rodents, such as squirrels, are fond of nuts and can open even the hardest black walnuts to get at the seed within. Acorns, seeds, fruits, berries, young leaves, buds and shoots, green leaves, tubers, and bulbs are eaten by a wide variety of rodents. Burrowing rodents tend to concentrate on the roots, tubers, and bulbs that they encounter in making their tunnels, although some burrowers forage on the surface as well. The larger caviomorphs, such as the capybara, paca, mara (*Dolichotis patagona*), and agouti (*Dasyprocta aguti*), are essentially grazing animals, eating all kinds of green vegetation, young stems, and fruits. Marmots (*Marmota*) have rather similar food habits. Old World porcupines are essentially omnivorous, including carrion in their food. New World porcupines eat leaves, twigs, buds, and bark. Beavers cut aspens, willows, and poplars, using leaves, buds, and bark for food; they also eat a variety of aquatic plants. Many desert-living rodents have a diet that, in the dry season, may be almost exclusively seeds, from which they are able to obtain enough water for survival, using the water produced by metabolism. African cane rats eat coarse grasses and shrubs, one of their favourite foods being sugarcane. A number of rodents, especially during hard winters, strip the bark from trees. Because of their large numbers, rodents can seriously damage the natural plant cover.

On the other hand, many rodents will, on occasion, eat animal food. Muskrats eat freshwater mussels and crayfish, as well as a wide variety of dead or dying water animals. Many rodents, will, at least occasionally, eat birds' eggs, nestlings, or insects. Several genera of South American cricetids live mostly on aquatic animals, of which fish make up the largest part. The grasshopper mice (*Onychomys*) of western United States and adjacent parts of Mexico and Canada are almost exclusively carnivores. Their food consists primarily of insects and earthworms, but may include birds and even other rodents that they kill. In the Arctic portion of their range, ground squirrels eat any carrion available, including their own dead relatives or even walrus or whale meat. They particularly like meat with a high fat content, presumably for the energy value.

Longevity. Rodents are generally considered to have a very short life span. This is certainly true in nature, where rodents are one of the primary food sources of carnivorous birds and mammals. Particularly among the small, mouselike rodents, the average life expectation at birth must be only a few weeks or months; relatively few individuals live much more than a year; and an animal two years old has reached a ripe old age. It is clear that this limitation is largely imposed on the animals by the activities of predators. Cricetids, for example, which normally live less than two years in the wild, have been kept up

Limits of life span

to five years in captivity, chipmunks and ground squirrels up to seven or eight years, gerbils and dormice over five years, guinea pigs about eight years, chinchillas and North American porcupines about 10 years, and woodchucks, pacas, and the Old World porcupine even longer.

Predation on rodents. Because rodents are one of the most important items in their food supply, carnivorous birds (especially hawks and owls) and mammals provide a very strong selective force acting on small rodents. Many raptorial birds use their eyes for hunting and are able to detect any rodent slightly less well concealed than the average. It has been shown experimentally that owls selectively capture light-coloured mice over dark ones on a dark background and dark over light on a light background down to a light level of 10^{-7} of a footcandle. As a result of such selective pressure, rodents have evolved fur colours that closely match the environmental background where they live. For example, a late Pleistocene flow of dark lava in New Mexico is now inhabited by melanistic (dark) rats and mice, in strong contrast to the normally coloured rodents surrounding it and to the almost white rodents found in the nearby White Sands area.

Among mammalian carnivores, all but the very largest normally eat rodents, at least as an important part of their diet, and many will start teaching their young how to hunt at the expense of small rodents. The abundance of rodents, moreover, affects the numbers of their predators in at least some instances. The English ecologist Charles Elton has shown that there are striking parallels between cyclic abundance of lemmings and snowy owls and foxes in northern Labrador, the lemming maxima and minima preceding those of the predators by about six months to a year.

BEHAVIOUR

Gnawing. One of the most general habits of rodents is gnawing, much of which serves merely to wear down the incisors. These teeth grow continuously, at rates that have been found to range upward from two millimetres (0.08 inch) per week in the few species in which rates have been measured. Gnawing is performed by all rodents at frequent intervals. It involves the fore-and-aft movement of the lower jaw, the upper incisors holding a hard object and the lower incisors cutting against it. The same movements may be used with nothing held in the teeth, the lower incisors merely alternating between cutting against the rear side of the upper incisors and shifting forward so that the uppers cut against the rear of the lowers. Because the hard enamel is limited to the anterior side of the teeth, such use wears away the softer dentine at a faster rate, leaving the enamel to form a sharp chisel edge. Rodents prefer to use their incisors on hard objects, which provide part of the abrasion. These may be nuts, bark, tree trunks (as in beavers), often bones (which may also be gnawed for a supply of calcium), or even human possessions such as boards in houses and barns or even metal telephone cables.

Function of gnawing

Nesting and denning. *Food storage.* A large proportion of rodents build underground homes, with a central nest chamber, in which they sleep, raise their young, and, often, hibernate. Many others have similar nests in grass or trees. Special food storage areas often are associated with such living quarters; for burrowing forms, these food areas are special chambers off the main passageways. Ground squirrels, kangaroo rats, and other rodents bring grass seeds back to such chambers in their cheek pouches (on the inside of the mouth of ground squirrels; outside the mouth and fur lined in the Heteromyidae and Geomyidae), and will often store several times as many seeds as they will eat during the season when they rely on their stores. Excavation of one kangaroo rat den, from which a single rat weighing 149 grams (about five ounces) was removed, showed nine underground storage chambers, containing from one to nine litres (one to eight quarts) of seeds, with a total of almost 39 litres (35 quarts). Kangaroo rats, in addition, dig storage pits, about 2.5 centimetres (one inch) in diameter and the same in depth, near their burrows, which they fill with seeds. In one case, 875 such caches were located in an area of 5.1 square metres (55 square feet) adjacent to a single den. It has been estimated that seed collection in this manner may result in the loss of as much as a quarter of the grain crop in some parts of the world, and at times of rodent plagues, the destruction locally may approach totality. On the other hand, such underground stores are an important source of germinating seeds for wild grasses in steppe regions such as western North America and Central Asia.

The habit of squirrels of carrying acorns and nuts to hollow trees or barns, or digging holes in the ground in which the nuts are placed, is well known. The animal's memory of the location of these stores is poor, but if enough nuts or acorns are hidden, the animal will be able to find a sufficient number to keep him alive during the winter.

The North American pack rat or trade rat (*Neotoma cinerea*) is attracted by bright and shining objects, which it picks up to carry home to its nest, a jumble of sticks, twigs, grass, and assorted collectors items. A popular superstition is that this animal is a fair businessman, who, on seeing something he wants, always leaves a replacement that is, in his opinion at least, of equal value. The fact is that, while carrying one trophy, the rat may see another that is more attractive and so puts down the first to pick up the second, leading the human who has lost a possession to believe that the rat was carrying out a "business deal."

Constructions by beavers

Beaver dams and lodges. Among the best known activities of rodents are those of beavers, which go to great lengths to store food, cutting alder and other food trees into 0.5- to 2.5-metre (1.6- to 8-foot) lengths and floating the logs in their ponds or embedding one end of these sticks in the mud of the ponds formed by their dams. These sticks make an available food supply during the winter when the ponds are frozen. In order to provide an aquatic habitat, safe from their normal predators, beavers build dams that often run several hundred feet in length, and as much as two metres (6.6 feet) high on the downstream side. The largest dam reported, from near Three Forks, Montana, was 652 metres (2,140 feet) long, 4.3 metres (14 feet) high at the highest point, and seven metres (23 feet) thick at the base. The dams are made of twigs, branches, and logs from which the bark has been eaten, piled on the ground with alternating layers of gravel or mud. When finished, the upper surface is plastered with mud, usually dug from the bottom of the pond above the dam, to make it watertight. The pond behind the dam serves as the refuge and first defense of the beaver. Consequently, beavers work continually to maintain the dams, by repairing, raising, and lengthening them, a process that may extend over several generations. Abandoned beaver ponds gradually silt up, forming very characteristic meadows.

Beavers also build lodges in the ponds behind their dams or burrows in the banks. Lodges are built up in the same manner as dams, and they are enlarged along with the rising water level of the ponds as the dam is raised. For ventilation, however, the upper part of the lodge is much more loosely constructed than the dam. The floor is kept above the water level, and there may be two levels, the lower for feeding, the upper for sleeping. The entrance to the lodge is through a tunnel, the outer end of which is below water level. Bank lodges are burrows in the bank, with the inside arranged like that of lodges. Again, access is gained through a tunnel with the mouth below water level. Burrows may extend more than nine metres into the bank.

Beaver canals are less well known but highly remarkable. They are dug to permit the transport of birch or aspen logs for food supplies from the sources where they are cut to the ponds. The logs are floated in the water, where the beaver is much safer than on land. Such canals are dug whenever they can be made with little or no damming. They vary from 0.3 to 1.2 metres (one to four feet) in width, up to 0.6 metres (two feet) deep, and canals as long as 223 metres (732 feet) have been reported.

Beavers live on the bark of trees, that of aspen being their favourite, but birch, cottonwood, willow, and alder also are used. When these are in short supply other plants are eaten, but conifers are rarely cut except for building materials. In one night a single adult beaver can fell a seven- to 10-centimetre (three- to four-inch) poplar, cut it into sections one to three metres long, and drag the logs to the water. Normally, beavers do not cut up logs over 15 centimetres (six inches) in diameter, although they will fell trees considerably larger. The largest tree on record cut by beavers was 115 centimetres (46 inches) in diameter. The cutting of trees is apparently done at random, with all gnawing likely to be on one side, unless the tree has a large diameter, in which case it may be gnawed all the way around. Because trees normally fall downhill, hence toward a pond, this is probably the origin of the legend that beavers can make trees fall as they wish.

Types of dormancy

Hibernation. Many rodents use their nests or burrows for shelter during unfavourable seasons, especially winter. There is considerable variation, even among closely related animals, in the amount of activity that occurs at such times and how often animals wake and eat some of the stored food. In a considerable number, including chipmunks, pocket mice (*Perognathus*), some jerboas, and some field mice, there is extensive torpor in wintertime, the animals sleeping extensively, but waking two or three times a day to eat. Hamsters (*Cricetus* and *Mesocricetus*) are deep hibernators, going into a sound sleep from which they awaken only periodically to eat. The most pronounced type of hibernation involves the accumulation of extensive amounts of body fat during the late summer and early fall, and deep hibernation with few or no awakenings during the winter. This condition occurs in marmots, some ground squirrels, a large proportion of dormice (whose English name is derived from the French *dormir* meaning "to sleep"), the jerboa (*Allactaga*), and most or all zapodids (jumping mice and birch mice). Estivation, or summer sleep, is rare in rodents but occurs in woodchucks and some ground squirrels. No suggestions of hibernation occur in murids, caviomorphs, or phiomorphs, any other tropical rodents, or rodents with a recent tropical or subtropical ancestry.

Population movements. Several types of rodents, subject to periodic fluctuations in numbers, reduce the amount of overpopulation by migration. Generally, the migration is simply the movement of juveniles away from a home area that has become fully populated. Migration of such animals may be quite extensive, but, because it is usually an individual phenomenon, it is often not noticed. Such animals are subject to very high mortality from predators because they are moving in unaccustomed territory and without the shelter of a permanent home. Muskrats and beavers are examples that are more readily noted, because they move from one suitable stream to another and may travel as much as several miles across country looking for a suitable home. There are many reports of major migrations. In August 1946, gray and fox squirrels migrated out of an area in northwest Wisconsin, apparently due to a shortage of acorns. When they reached streams and rivers, they crossed them by swimming, many drowning in the wider rivers. In 1935 a mass movement of gray squirrels was noted from the area east of the Hudson River, to the west. Hundreds (one report says 2,000) of drowned squirrels were found along the west bank of the Hudson River.

Lemming migrations

The most famous migrations of rodents, however, are those of lemmings, especially in Scandinavia. There tend to be two migrations per year, one in spring and one in autumn. The exact causes of the movement are not known, but a change of habitat with season seems to be the most important single factor. It has been believed that the migrations are caused by overpopulation and a shortage of food, or by claustrophobia (a feeling of mental anguish from the crowded conditions), but these probably are not valid causes.

The lemming migrations are individual activities rather than unified movements, as is popularly supposed. Careful observations of migrations show that each individual acts alone, but that there may be groups of several individuals within a few minutes, followed by a gap of 10 minutes or more. In autumn, at least, the migrations normally occur at night. Individual lemmings generally move fairly rapidly in a generally constant direction, although they will occasionally stop to feed or groom. Rates of migration have been calculated as about one metre per second (2.25 miles per hour), but such speeds are not sustained, and it is probable that distances of eight–10 kilometres (five–six miles) per day are normal. The migrations tend to be oriented by roads or paths made by humans, trails made by reindeer, or other similar routes, although the general direction seems to be outward in all directions from the general source area and to be determined before the movement begins. Contrary to the older reports and popular belief, lemmings do not rush forward without hesitation into water when reaching a shoreline. They apparently try to avoid swimming, if possible, running up and down the shore looking for a narrow crossing or an area where they can cross on ice. Apparently, they enter water willingly only if they can see the silhouette of the far shore. This is in marked contrast to activities of their relative, the water

vole (*Arvicola*), which always heads for the water when threatened. When swimming lemmings use their hind legs almost exclusively but swim very high out of the water. Lemmings become exhausted and drown after swimming 15 to 25 minutes in water with waves about 15 centimetres (six inches) high. Observations along the Swedish-Finnish border show that the lemmings had no problem crossing a strait 200 metres (650 feet) wide on a calm night, but that considerable numbers drowned on a windy night. A lemming was observed swimming at approximately the middle of a lake two kilometres (1.2 miles) wide. Reports of lemmings swimming out into the North Sea and being found alive 16 kilometres (10 miles) or more from shore must be viewed skeptically.

Recent studies in Scandinavia indicate that the migrations are a normal part of the activity of the lemmings. They serve not only as a means of dispersal of the species but also to permit the animals to occupy different habitats at different seasons.

Rodents are often thought of as defenseless animals, but the incisors are very sharp and powerful, are activated by powerful jaw muscles, and can inflict a deep wound. Among themselves, rodents fight for territories, the possessor of a territory usually being able to drive interlopers away. They also fight very effectively in self-defense, as indicated by the popular expression, "fighting like a cornered rat." A full-grown wharf rat is capable of defending itself against a determined tomcat. Even animals as small as lemmings are able to discourage attack by weasels.

REPRODUCTION

Flexibility in breeding habits

The reproductive habits of rodents are exceedingly varied in nature and are capable of being even further modified in domesticated or partially domesticated forms. Many, especially the larger types, reproduce once a year. Others produce several litters during a single season. Some have only one or two young at a time; others, large numbers. Some are capable of reproduction at very early ages, others (particularly members of the Caviomorpha) not until after a considerable period of growth. Most are born naked and helpless, and are cared for in nests; a few are able to run and keep up with their mothers almost at once. Most rodents are polygamous; some mate for the duration of a single breeding season; and a few (beavers, for example) have permanent mates.

Beavers do not breed until the second January or February after birth, producing a litter of two or four young after about 12 weeks of gestation. The capybara has a litter that is, on the average, slightly larger, after 15 to 18 weeks gestation; the viscacha, a smaller animal, has two young in the spring about 22 weeks after breeding. Nutrias breed twice a year after reaching an age of one year, if the winter temperature does not fall below 16° C (60° F) for lengthy periods, and produce litters averaging about five young. The European porcupine (*Hystrix cristata*) breeds in the spring, with one to four young being born about 16 weeks later, and some of its tropical relatives (other members of the Hystricidae) produce two litters a year. The North American porcupine (*Erethizon dorsatum*) of the caviomorph family Erethizontidae, has a single young after 30 weeks gestation. Most of the smaller members of the order, however, have considerably higher reproductive rates. In many cases there may be more than one litter per year, with high numbers per litter, examples being shown in Table 2.

Table 2: Reproductive Rates of Certain Rodents

animals	litters/year	young/litter
North American gray squirrel (*Sciurus carolinensis*)	2	2–3
Eurasian gray squirrel (*S. vulgaris*)	2	5–7
Red squirrels (*Tamiasciurus*)	2	4–6
Pocket gophers (*Geomys*)	1–2	1–4
Pocket mice (*Perognathus*)	1–2	2–8
Kangaroo rats (*Dipodomys*)	1–3	2–4
Deer mice (*Peromyscus*)	up to 4	2–7
Hamsters (*Cricetus*)	several	6–12
Lemmings (*Lemmus*)	several	3–9
Muskrats (*Ondatra*)	2–5	5–7
Voles (*Microtus*)	up to 13	4–8

Source: E.P. Walker, *Mammals of the World* (1964) and S.A. Asdell, *Patterns of Mammalian Reproduction* (1964).

This rather high rate of breeding is intensified by the fact that, in many of the smaller rodents, sexual maturity is reached at an early age, normally earlier in the females than in the males. The females breed when less than a year old in many squirrels, some pocket gophers, and the pack rat. In most of the Cricetidae, however, the young appear to reach sexual maturity considerably earlier—harvest mice (*Reithrodontomys*) in five weeks; the hamster (*Cricetus*) in six weeks; and voles (*Microtus*) in six to seven weeks.

In the United States, wild populations of the house mouse reproduce throughout the year, with an average of 5.5 litters and 31 young per female per year in buildings, and 10.2 litters and 57 young per year on farms. In the laboratory mouse, the mean litter size has been reported as ranging from 4.5 to 7.4, depending on the strain. The second litter is the largest, after which there is a steady decrease. Litter size has been reported to have a range of two to 12, and as many as 19 healthy embryos have been removed just before term from a single female. There may be five or more litters per year. The gestation period is about 20 days, and the first mating usually occurs at seven to 10 weeks of age, though it has been reported that, in one strain of albino mice, the first estrus occurs, on the average, at 39 days and results in about 50 percent pregnancies. The breeding span of most mice in the laboratory is 12 to 18 months. Lactating females may become pregnant, but the gestation will be lengthened by one to two weeks.

Reproductive potential

Wild Norway rats breed throughout the year, taking advantage of the sheltered environment available from living in association with humans. The size of the litter is correlated with the size of the mother. In the laboratory female rats reach sexual maturity in 33 to 120 days. The number of young varies with the strain, two typical ones having litters averaging 6.5 and 8.9. Normally, a female's second or third litter is the largest, and the litter size decreases rapidly after the tenth. Over a number of generations of laboratory life in a colony of gray rats, however, the total number of young per female grew from an average of 23 to an average of 63.

The laboratory guinea pig has a gestation period of about 68 days, a very long period for such a small animal but characteristic of caviomorphs. There are usually three or four young per litter. The age of puberty varies from 55 to 70 days.

Although most rodents seem to increase their reproductive potential under domestication, this is not true of chinchillas. In the wild, they produce litters of one to six after about 15 weeks pregnancy; domesticated ones seem normally to have only one offspring at a time.

FORM AND FUNCTION

In most aspects rodents are relatively primitive, not highly specialized mammals. The skeleton is usually that of a quadrupedal, scampering mammal, not far from the basic placental stock. The digits are clawed and usually five in number, with little or no opposability of the first digit. The limbs are not far from the same length, although the hind legs are always somewhat longer. The animals are adapted to a broadly herbivorous diet. The main specialization of the digestive system is the possession of a large cecum (a blind pouch) at the junction of the large and small intestines.

Specializations for gnawing. As opposed to this basic primitiveness, the rodents have developed, as their basic innovation, a gnawing mechanism of an efficiency that is not approached by that of any other mammals. The glenoid fossa, the concavity where the lower jaw articulates with the skull, slopes downward, from rear to front, and there are no processes (as there are in all other mammals) at the anterior and posterior limits of the fossa that restrict the anteroposterior movement of the jaw. The jaw, therefore, is capable of functioning when pulled to the rear, separating the incisors from each other, and permitting chewing or grinding by the cheek teeth. The jaw can also be pulled forward and downward, separating the

cheek teeth and bringing the tips of the incisors together for gnawing.

Evolutionary shifts in jaw muscle attachments

These shifts in the position of the jaw, as well as the functioning of the jaw in either position, are brought about by a combination of the actions of the various jaw muscles, of which the temporal, pterygoid, and masseter are the most important. The forward and backward (anteroposterior) movements, including those of gnawing, were caused, primitively, by all three muscles, all of which have a slight anteroposterior component. But a shift of the anterior muscle, the masseter, began to occur during the Eocene Epoch (about 50,000,000 years ago). Originally, in a condition called protrogomorphous, the masseter ran from the zygomatic arch (cheekbone) to the lower jaw. Parts of the muscle, retaining the original points of insertion on the lower jaw, shifted their origins forward onto the snout, doubling the length of the muscle and greatly changing its direction of pull. In some rodents it was the lateral portion of the masseter that shifted forward onto the face, compressing the infraorbital foramen (an opening in the bone), through which pass nerves and blood vessels, between the muscle and the snout (the sciuromorphous condition); in others it was the deeper portion of the masseter that shifted forward, through the infraorbital foramen (presumably following an initial enlargement of the foramen), onto the face, resulting in further enlargement of the foramen until it may now be larger than the eye socket (the hystricomorphous condition); and in still others, a combination of the two shifts has occurred (the myomorphous condition). Although these changes are of importance in determining rodent taxonomic relationships, they are probably not as controlling as was once thought, as each structural condition may have originated more than once.

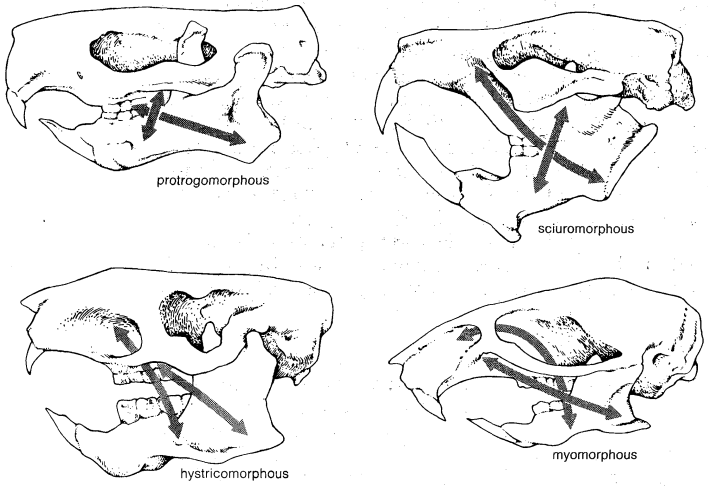


Figure 31: Skulls of the four basic types of rodents. Arrows show positions of branches of the masseter muscle.

Continuous growth of incisors

But the main structural change from the primitive mammalian condition has occurred in the incisors. In all known rodents, beginning with the very earliest, the incisors are reduced to a single pair in each of the upper and lower jaws. These teeth grow throughout life, the enamel cap being restricted to the anterior 30 to 60 percent of the perimeter of the tooth. The rate of incisor growth has been measured in several rodents, and varies from about two to three millimetres per week (0.08–0.12 inch) in nonburrowing forms to five millimetres (0.2) per week in pocket gophers, in which the incisors are used for digging as well as for gnawing. In hibernating rodents, the incisors continue to grow, but at a greatly reduced rate. If the incisors do not meet, due to deformity or breakage, the teeth will continue to grow indefinitely. Since the lower teeth are arcs of large circles, the lower incisors simply curve forward and upward, becoming completely nonfunctional. But the upper incisors are large arcs of much smaller circles. With growth, they may spiral outward, like the spirals on a handle-bar moustache, or they may grow around and upward through the throat between the lower jaws, growing through the snout and eventually locking the jaws closed. Either condition is fatal, but rodents have survived in nature for a long time with such malformed incisors.

Because of the continual growth of their incisors, rodents have an unusually heavy need for calcium, which, together with a need for abrasion of the incisors, explains the frequency with which rodents gnaw bones.

Adaptations for specialized locomotion. Many rodents have shifted their locomotion from the characteristic scampering type. In burrowing rodents, the limbs are short and massive, and the hands, particularly, are widened and strengthened with heavy claws. Many, but not all, burrowing rodents use their incisors to help in digging. In such cases, a fold of skin is likely to close the mouth behind the incisors, and the radius of curvature of both upper and lower incisors is increased, so that they extend forward from the mouth, cutting the soil an appreciable distance in front of the animal. Eyes and ears are generally reduced in size in these forms.

There is little difference, structurally, between the terrestrial scampering rodents and the arboreal scamperers, such as squirrels. The claws of the latter are sharper, but locomotion is basically the same. A few rodents, such as the New World porcupine, actually climb (as would a human) rather than running up and down the bark, as does a squirrel. Only one rodent, the tree porcupine (*Coendou*) of Central America and northern South America, has a truly prehensile tail (although several rodents have semiprehensile tails that they can wrap around branches for extra support). A number of gliding rodents, "flying squirrels" of one sort or another, some of which are not closely related to the true squirrels (Sciuridae), are supported in the air by a fur-covered membrane (patagium) extending between the fore and hind limbs and usually between the hind limbs and the tail.

A rather common development among rodents has been the reduction of the forelimbs and an increase in the size of the hind limbs and tail for locomotion by means of a series of kangaroo-like leaps, which in several forms can exceed 4.5 metres (15 feet).

Adaptations for water conservation. With their wide geographic distribution, rodents have invaded desert regions in all parts of the world. Many of these have developed the ability to function on limited supplies of water. This has been achieved in several ways. Some are nocturnal, remaining asleep during the heat of the day in burrows deep enough to reach ground moisture, thus reducing evaporation. Others seek out succulent desert plants for food. But it has been shown that a number of desert rodents (kangaroo rats and jerboas, for example) have the ability to live exclusively on seeds, with no external water source, and to obtain all of their needed water from their metabolism. Such forms also produce a greatly reduced quantity of extremely concentrated urine and a small amount of feces. Two species of spiny mouse (*Acomys*) were investigated in Israel. Both inhabit the same desert areas, but one is diurnal and the other is nocturnal. Both produce urine the urea content of which is extremely high. Neither species can survive on a diet of barley without water to drink, although gerbils and jerboas from the same area showed no ill effects after four weeks without water. The diurnal spiny mice, however, were able to survive and grow with no water except seawater, whereas the nocturnal ones could not tolerate such a salt concentration.

EVOLUTION AND PALEONTOLOGY

Rodents are relatively poorly represented in collections of fossils, in spite of their great abundance at the present time. They have often been overlooked because of their small size, but modern intensive exploration for fossils usually results in a much more abundant representation of small mammals, especially rodents.

Paleocene. The earliest known rodents come from the late Paleocene (about 57,000,000 years ago) of North America, by which time they had already acquired all of the diagnostic features of the order. The ancestral family, the Paramyidae, was also present in Europe, where it first appeared in the earliest Eocene. There was very rapid diversification of the order during the Eocene, initially involving the Paramyidae, but with other families soon

Paramyidae, the earliest rodents

appearing. Most of the Eocene rodents were protrogomorphous, the masseter muscle restricted to its primitive origin on the cheekbone. By the middle Eocene of Europe, rodents with advanced types of masseters were present, and others occurred in the late Eocene of North America. The skeleton of the paramyids was basically that of a generalized scampering animal, approximating a rat in its method of locomotion. Before the end of the Eocene it is probable that both leaping and burrowing variations had arisen, although there is some uncertainty on this point.

Oligocene. A major gap in the knowledge of rodent evolution occurs at the Eocene-Oligocene boundary (about 38,000,000 years ago). A number of modern families with advanced types of jaw muscles appeared at about the same time in North America, Europe, or Asia, including the Cricetidae, Heteromyidae, Geomyidae, Castoridae, and Ctenodactylidae, with no good indications at present as to their geographic or ancestral sources. (The Ctenodactylidae are known from the Eocene of Kazakhstan, however.) At about the same time, rodent groups appeared suddenly in South America and Africa (Caviomorpha and Phiomorpha, respectively). Both groups are hystricomorphous and hystricognathous and have often been considered to have been related, though others consider them to have acquired their similarities independently.

Miocene. The majority of the living families of rodents had appeared by the Oligocene, and most of the remaining ones occur in the Miocene (around 26,000,000 years ago), so that the later part of the evolution of the rodents was largely a diversification of the various groups. The ancestors of a wide variety of specialized modern forms can be recognized from teeth, but most of them are not known from skeletons, and the evolutionary status of the modern locomotor conditions is not well known. Highly evolved burrowing rodents were certainly present in the Oligocene in Mongolia, however, and less specialized ones occurred in North America. Ancestral kangaroo rats had attained a leaping ability by the Miocene of North America. Gliding and burrowing forms are known from the Miocene of Africa. In South America, the differentiation of the descendants of the invaders present in the early Oligocene proceeded rapidly, and the Miocene forms occupied nearly all potential rodent niches on that continent.

Richard Keane

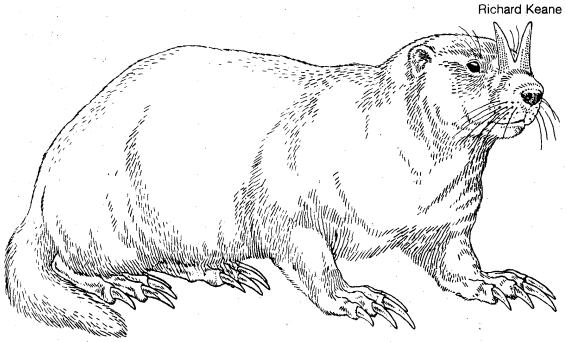


Figure 32: Reconstruction of *Epigaulus*, a primitive Pliocene rodent.

Appearance of Muridae

Later evolution of rodents. The latest rodents to differentiate apparently were the members of the family Muridae. This group appears suddenly in early Pliocene deposits of Europe, probably having invaded that continent from Asia, the southeastern part of which, together with the adjacent East Indies, is now considered the evolutionary centre of the family. The Muridae expanded rapidly all over the Old World. An isolated incisor is known from the Pliocene (approximately 4,000,000 years ago) of New Guinea, where considerable local evolution has occurred. The murids reached Australia late in the Cenozoic Era (they are unknown there before the Pleistocene), and underwent a rapid evolution, developing arboreal, burrowing, and leaping forms, paralleling the evolution of the entire order in the rest of the world.

When North and South America were united in the beginning of the Pleistocene (about 2,500,000 years ago), there was a major invasion of South America by cricetids, which expanded rapidly and became highly diverse in that area.

An interesting aspect of rodent evolution are the very rapid diversifications (evolutionary explosions) that occurred whenever highly specialized members of the order reached areas that had been previously uninhabited by gnawing mammals (South America and Africa in the Oligocene, New Guinea in the Pliocene, Australia in the Pleistocene). These, apparently, resulted in a rapid spread into most of the major available ecologic niches in a matter of, at most, 2,000,000 or 3,000,000 years. The differentiation of the South American cricetines since the beginning of the Pleistocene has not been quite as fast, though still explosive, presumably because there already were numerous rodents on that continent. The initial diversification of the rodents during the Eocene seems to have taken place at a considerably slower rate, probably because they had not yet become as thoroughly adapted for gnawing as was the case later.

Because of their small size and great abundance, rodent fossils are becoming progressively more important as guide fossils that enable accurate separation of successive geologic horizons.

CLASSIFICATION

Distinguishing taxonomic features. Many features have been used to determine the relationships of the subgroups of the rodents. The most generally used are the differentiations of the jaw muscles or the modifications of the skull and jaws resulting from muscular changes. The masseter muscle may have its primitive position, arising from the cheekbone (protrogomorphous); its lateral division may have shifted forward onto the snout, compressing the infraorbital foramen against the snout (sciuromorphous); the medial division may have moved forward through the infraorbital foramen, onto the snout, greatly enlarging the foramen (hystricomorphous); or both branches of the masseter may have shifted (myomorphous). In most rodents the angular process at the rear of the lower jaw extends downward from the ventral side of the alveolus of the incisor (sciurognathous); in some, all of which are also hystricomorphous, the angle arises from the side of the alveolus (hystricognathous). Other features that have been given important weight in delimiting the major subgroups include: crown patterns of the cheek teeth (premolars and molars); histologic structure of incisor enamel, which may have one layer (uniserial), a few (pauciserial), or many (multiserial); structure of the penis or of the baculum (os penis); structure of the male genital tract, including particularly the presence or absence of an outgrowth, the sacculus urethralis; fusion or lack thereof of two of the ear ossicles, the malleus and incus; and the pattern of development of the extra-embryonic fetal membranes.

ANNOTATED CLASSIFICATION

The classification presented here is based upon that of American paleontologist A.E. Wood. The rodents are one of the outstanding examples of a group in which closely similar structures have evolved numerous times independently, a phenomenon known as parallelism, with the result that there is very strong disagreement among taxonomists working with the group as to how the order should be subdivided.

Groups indicated by a dagger (†) are known only from fossil remains. Dental formulas indicate the number of pairs of teeth present in the upper jaw (above the line) and lower jaw (below), the figures representing, respectively, pairs of incisors, canines, premolars, and molars. When a partial formula is given the type of tooth is indicated by I, C, P, or M.

ORDER RODENTIA

Gnawing mammals with ever-growing incisors, reduced to a single pair in both upper and lower jaws and with the enamel limited to the anterior face of the incisors, so that wear maintains a sharp chisel edge; glenoid fossa slanted, with neither preglenoid nor postglenoid process, masseter the principal jaw muscle; dental formula reduced, never exceeding $\frac{1.0.2.3}{1.0.1.3} = 22$. About 350 living genera and 2,400 living species; over 400 extinct genera have been described.

Suborder Sciuromorpha

Masseter primitively either limited to zygoma, or rarely with only a slight forward displacement, but in the squirrels shifted forward onto face (sciuromorphous). Always sciurognathous. No sacculus urethralis. Malleus and incus bones of ear separate. Incisor enamel pauciserial or uniserial. Paleocene to Recent.

†*Family Paramyidae*

Paleocene to early Miocene; Northern Hemisphere. Cheek teeth cuspidate, clearly derived from basic mammalian tribosphenic type. Tympanic bullae became co-ossified with the skull several times independently, within the family. Incisor enamel pauciserial, with one possible exception. Presumably ancestral to the rest of the order. One subfamily incipiently hystricognathous. From size of a mouse to as large as a beaver. Locomotion, when known, scampering, with possible incipient leaping in one form.

†*Family Sciuravidae*

Eocene of North America; one genus from Eocene of Central Asia. Cheek teeth formed of 4 transverse crests rather than of separate cusps; locomotion probably scampering; mouse-size to rat-size. Incisor enamel pauciserial. Perhaps ancestral to several mouse- or ratlike groups.

†*Family Ischyromyidae*

Late Eocene to Oligocene of North America. Masseter muscle has begun to move forward onto the snout in some forms, in primitive position in others. Cheek teeth 4-crested. Incisor enamel uniserial. Locomotion scampering. Size of a gray squirrel or somewhat larger.

†*Family Cylindrodontidae*

Middle Eocene to Oligocene of North America, Oligocene of Central Asia. Burrowing rodents with high-crowned to ever-growing cheek teeth, based on a pattern of 4 transverse crests. The most specialized forms used the incisors for digging, as shown by forward extension of these teeth. Incisor enamel uniserial. Size from that of a chipmunk to that of a marmot. Presumably derived from North American Sciuravidae.

†*Family Protoptychidae*

Late Eocene of North America. Skull with highly inflated auditory bullae, suggesting, by analogy with other rodents, that these were leaping animals. Cheek teeth high crowned, with pattern of 4 crests. Size of a gray squirrel.

Family Aplodontidae (mountain beaver, or sewellel, and fossil relatives)

Late Eocene to Recent of North America, Oligocene and Miocene of Europe, Pliocene of Asia. Single living species restricted to wet areas from British Columbia south to San Francisco Bay. Extreme hypsodont cheek teeth, based on cuspidate rather than crested pattern. Incisor enamel uniserial. Burrowers. Head and body length of living species (*Aplodontia rufa*) 300 to 460 mm (11.8–18.1 in.), with a short tail; weight 900 to 1,800 g (2–4 lb).

†*Family Mylagaulidae*

Miocene to Pliocene of North America. Strongly hypsodont cheek teeth, with greatly enlarged premolars, whose continual growth forced 1 or more of the anterior molars out of the jaws. Incisor enamel uniserial. Some individuals had paired horns on the snout. Whether these are taxonomic or sexual characters has not yet been demonstrated. Powerful burrowers. About the same size as *Aplodontia*.

Family Sciuridae (squirrels, chipmunks, and marmots)

Oligocene to Recent of Northern Hemisphere, Miocene to Recent of Africa, Recent of South America. Highly developed sciuromorphous pattern, with masseter muscle extending forward along the side of the snout and compressing the infraorbital foramen. Postorbital process of frontal separates the eye from the temporal muscle. Cheek teeth typically very similar to those of the Paramyidae, but some forms have rather complicated tooth patterns. Incisor enamel uniserial. About 70 living genera.

Suborder Myomorpha

Lateral branch of masseter muscle usually displaced forward alongside of snout, forcing infraorbital foramen against side of snout; in most forms, deep masseter penetrates a variable distance through upper part of infraorbital foramen. Sciurognathous. Incisor enamel uniserial. Locomotion scampering, jumping, arboreal scampering, or fossorial. A few forms are partially adapted to an aquatic life. Malleus and incus never fused. No sacculus urethralis. Cheek teeth usually reduced in number, with only a single fossil genus known that retains the primitive rodent dental formula; except for the Geomyoidea, the premolars are either lost or greatly reduced; the last molars are occasionally lost as well. Most members of the group are small, from the size of a mouse to that of a rat (head and body length ranges from 50 to 300 mm (2–11.8 in.), except in the Rhizomyidae, which may be somewhat larger).

Family Cricetidae (field mice, deer mice, voles, lemmings, muskrats)

Early Oligocene to Recent of Europe and North America, middle Oligocene to Recent of Asia, late Pliocene to Recent of South America, Pleistocene of Madagascar, Recent of Africa. Deep masseter expanded through upper part of infraorbital foramen to face; no premolars, molars $\frac{3}{3}$; cheek-tooth pattern based on cusps arranged into 5 transverse crests in both upper and lower teeth; teeth low crowned to very high crowned. Scampering, fossorial, arboreal scampering, jumping, or partly aquatic.

Family Muridae (Old World rats and mice)

Early Pliocene to Recent of Europe and Asia, late Pliocene to Recent of East Indies, Pleistocene to Recent of Africa and Australia. Introduced throughout the world by man. Deep masseter as in Cricetidae; cheek teeth normally the 3 molars, but one subfamily, the Hydromurinae of Australia and New Guinea, has lost the third molars, reducing the cheek teeth to $M\frac{2}{2}$; tooth pattern based on rounded cusps, arranged in transverse rows, but derivable by modification of a pattern like that of primitive cricetids; low to medium high crowned teeth. Scampering, arboreal scampering, occasionally fossorial or semiaquatic.

Famiy Heteromyidae (pocket mice, kangaroo rats and mice)

Oligocene to Recent of North America, Recent of northern South America. These and the next two families are sciuromorphous, with no penetration of the infraorbital foramen by the masseter, and seem to be closely related. Fur-lined cheek pouches, opening beside the mouth and reaching back to the shoulders, are used for the transportation of food to underground storage areas near the nests. Central stock of family scampering; specialized ones leap with their hind legs. Generally small, mouselike; length of head and body from 55 to 180 mm (2.2–7.1 in.). Cheek teeth, which include $P\frac{1}{1}$ $M\frac{3}{3}$, are low crowned to very high crowned. Pattern of teeth based on 2 transverse rows of 3 cusps each. Derivable from late Eocene members of the Eomyidae.

Family Geomyidae (pocket gophers)

Early Miocene to Recent of North America. Similar to heteromyids in regard to jaw muscles, dental formula, and tooth pattern, and in possession of fur-lined cheek pouches. Highly adapted burrowing animals. Cheek teeth have already become high crowned in the Miocene, and those of the living forms grow throughout the animal's life, with the enamel reduced to a plate on the anterior side of the upper teeth and the posterior side of the lowers.

†*Family Eomyidae*

Late Eocene to late Pliocene of North America; late Eocene to Pleistocene of Europe. Sciuromorphous jaw muscles; cranial anatomy and jaw structure very similar to heteromyids. Cheek teeth usually low crowned, with pattern of 5 crests, rather similar to that of the cricetids, although some members of the family developed teeth suggesting that they might be ancestral to heteromyids. Cheek teeth usually $P\frac{1}{1}$ $M\frac{3}{3}$, but one genus had $P\frac{2}{1}$. Skeleton and habits unknown.

Family Zapodidae (birch and jumping mice)

Late Oligocene to Recent of Eurasia, early Miocene to Recent of North America. A late Eocene genus from California is often (perhaps incorrectly) placed here. The Old World birch mice are arboreal scampering forms, as were probably all the earlier fossil members of the family. The jumping mice are as saltatorial as the kangaroo rats. Infraorbital foramen of medium size, with the deep masseter passing through it but with the superficial masseter remaining on the zygomatic arch. Cheek teeth $P\frac{1}{0}$ $M\frac{3}{3}$, with pattern very similar to that of cricetids, with which the fossils have frequently been confused.

Family Dipodidae (jerboas)

Late Oligocene to Recent of Asia, late Miocene to Recent of Europe, Pleistocene to Recent of North Africa. Masseter muscle like that of zapodids but with larger deep division. Highly saltatorial rodents of the steppes and deserts of the Old World, with extremely inflated auditory bullae. Cheek teeth $P\frac{1-0}{0}$ $M\frac{3}{3}$, high crowned in all living and most fossil forms. Tooth pattern suggestive of that of zapodids, from which the family is probably descended. General tendency to elongate the hind foot and to fuse the 3 median metatarsals to form a cannon bone, as part of the leaping adaptation.

Family Spalacidae (mole rats)

Early Pliocene to Recent of Europe, Pleistocene to Recent of Western Asia and North Africa. Highly specialized burrowing rodents with short, powerful legs and no external tail. Eyelids permanently closed. Masseter muscle a short distance forward

on face. Cheek teeth reduced to $M\frac{3}{3}$, (or possibly $P\frac{1}{1}$ $M\frac{2}{2}$). Ancestry unknown.

Family Rhizomyidae (African mole rats, bamboo rats)

Oligocene of Europe, early Miocene to Recent of Asia, and Pleistocene to Recent of Africa. Myomorphous, with considerable forward extension of the masseter. Burrowing animals with powerful legs and short tails, superficially resembling pocket gophers (Geomyidad). Cheek teeth high crowned to ever-growing, consisting of $P\frac{1-0}{1-0}$ $M\frac{3}{3}$. The European Oligocene *Rhizospalax* has features suggesting that this family and the Spalacidae are related. Otherwise, relationships unknown.

Suborder Caviomorpha

Deep branch to masseter has shifted its origin forward onto the face, passing through the very large infraorbital foramen. Angle of jaw hystricognathous. Malleus and incus fused; usually a sacculus urethalis in male genital tract. Incisor enamel multiserial. Cheek teeth usually $P\frac{1}{1}$ $M\frac{3}{3}$, but occasionally the milk premolar is retained throughout life, the formula being $dP\frac{1}{1}$, $M\frac{3}{3}$. Early Oligocene to Recent of South America, Pleistocene to Recent of North America and West Indies.

Family Octodontidae (octodonts, degus)

Early Oligocene to Recent of South America, Pleistocene to Recent of the West Indies. Small for caviomorphs, generally somewhat ratlike in appearance. Scampering to fossorial in habits. Occur from sea level to elevations of more than 3,000 m (10,000 ft) in the southern half of South America. The teeth range from low crowned to ever-growing, the enamel of the grinding surface arranged either in the shape of a kidney or figure 8. The burrowing forms use the incisors in digging. The genus *Platypittamys* from the early Oligocene of Patagonia is very close to being a common ancestor of all the Caviomorpha.

Family Echimyidae (spiny rats)

Oligocene to Recent of South America. Pleistocene to Recent of West Indies and Central America. Ratlike in appearance. Usually with spiny fur, although there are a few exceptions. Cheek teeth rooted, with a pattern formed by transverse folds of enamel. In all but the earliest known members of the family, the milk or deciduous premolars are retained throughout life, and permanent premolars never make an appearance. Inhabit moist forested regions; climbing or scampering; one genus is burrowing.

Family Ctenomyidae (tuco-tucos)

Pliocene to Recent of South America. The single living genus inhabits all of the southern half of South America, from sea level to elevations over 4,000 m (13,000 ft). They are fossorial and presumably derived from octodontids. Body form and size are very similar to those of the North American pocket gophers, with powerful digging muscles in the forelimbs and long, powerful claws.

Family Abrocomidae (chinchilla rats or abrocomes)

The common name comes from the soft underfur and ratlike appearance. Pliocene to Recent of South America. The living species inhabit mountainous areas of Peru, Bolivia, Argentina, and Chile. Presumably evolved from octodontids. The cheek teeth grow throughout life.

Family Chinchillidae (chinchillas and viscachas)

Early Oligocene to Recent of South America. One living genus (*Lagostomus*) lives in lowlands of Argentina; the other two (*Lagidium* and *Chinchilla*) live at elevations of about 800 to 6,500 m in the southern half of the Andes. The fur of *Chinchilla* is very soft and dense—that of the other genera less so. All gregarious, but especially the lowland viscachas (*Lagostomus*), which occupy colonies resembling those of prairie dogs. The lowland viscachas are large, head and body length being 470 to 660 mm (18.5–26 in.); the mountain genera are 230 to 320 mm (9–12.6 in.; *Lagidium*) and 225 to 380 mm (8.9–15 in.; *Chinchilla*). Cheek teeth ever-growing, but details of pattern lost early in life.

Family Capromyidae (hutias, coypus)

Middle Pliocene to Recent of South America, Pleistocene to Recent of West Indies, introduced in southern United States and parts of Europe. Differ from the Chinchillidae, from which they are probably derived, in having high-crowned to ever-growing cheek teeth, but with the details of the pattern persistent. The hutias of the West Indies resemble large rats. The coypus of South America are considerably larger, looking rather like a large muskrat. The fur is excellent, and the flesh is widely eaten in South America.

Family Dasyproctidae (pacas and agoutis)

Early Oligocene to Recent of South America, Pleistocene to Recent of West Indies, Recent of Central America. Large rodents (head and body length 320 to 800 mm [12.6–31.5 in.]), with weight up to 10 kg (22 lb). Limbs modified for cursorial locomotion, with the lateral toes, especially on the hind foot, reduced in size. Flesh is very palatable. Pacas in moist forested areas from Mexico to southern Brazil; agoutis in the same regions (plus the Lesser Antilles), but not restricted to forested areas.

Family Dinomyidae (pacaranas)

Early Miocene to Recent of South America; Pleistocene of West Indies. There is only a single living genus inhabiting the lower elevations of the northern half of the Andes. The animal is among the largest of living rodents. A considerable number of fossil forms, once thought to represent a separate family (Heptaxodontidae) probably belong here.

†*Family Elasmodontomyidae*

Pleistocene to Recent of Puerto Rico and northern Lesser Antilles. These forms are all extinct, but survived until about the time the area was colonized by the Spaniards. Medium-sized ground-living rodents, with ever-growing cheek teeth formed of a series of enamel plates inclined at about 45° to the long axis of the jaws.

†*Family Eocardiidae*

Early Oligocene to middle Miocene of South America. Cheek teeth medium to high crowned. Size about that of a guinea pig.

Family Caviidae (guinea pigs, cavies, and maras)

Late Miocene to Recent of South America. Cheek teeth ever-growing, with a simplified crown pattern of alternating V's. Digits reduced to 4 on the front and 3 on the hind foot. Cavies with short legs, ears and tails; maras superficially resembling hares.

Family Hydrochoeridae (capybaras)

Early Pliocene to Recent of South America, Pleistocene to Recent of Central America, Pleistocene of West Indies and southern United States. Most members of this family are extinct, there being but a single living genus, with 1 species in eastern Panama and the other in South America east of the Andes and north of the Rio Paraná. Forest dwellers, capable of escape by running, but tend to retreat to water when closely pursued. Derived from Caviidae and have the same reduction of digits. Largest living rodents, head and body length reaching 1.2 m (3.9 ft); they may weigh over 45 kg (100 lb).

Family Erethizontidae (New World porcupines)

Early Oligocene to Recent of South America, Pleistocene to Recent of North America, recent of Central America. Large, slow-moving, heavy-bodied rodents, with some hairs modified into sharp, barbed spines that are easily detached when they make contact with an enemy. Cheek teeth rooted, with a simple pattern of reentrant folds, essentially unchanged since the Oligocene. One South American genus has a prehensile tail.

Suborder Phiomorpha

A group of families of African rodents the ancestry of which can be traced back to the early Oligocene of Egypt. Most of them are hystricomorphous and hystricognathous, and they are often considered very closely related to the Caviomorpha. Multiserial incisor enamel; malleus and incus fused and sacculus urethralis present in living forms. Cheek teeth $\frac{4}{4}$, consisting (in all but a few of the earliest types) of the last deciduous premolar, retained throughout life, and $M\frac{3}{3}$.

†*Family Phiomyidae*

Oligocene to Miocene of Africa. These rodents reached Africa when it was isolated from the rest of the world and differentiated to become highly diverse, including scampering, burrowing, and perhaps arboreal and leaping forms. In the Miocene other rodents reached Africa, and only a few lines of phiomyids were able to survive the resulting competition. Several Oligocene genera retained permanent premolars, although in 1, at least, they apparently never were functional.

Family Petromuridae (rock rats or dassie rats)

Pleistocene to Recent of southern South Africa. Ratlike in body form, the single species inhabits rocky hills, living in narrow crevices in rocks.

Family Thryonomyidae (cane rats)

Miocene to Recent of Africa and early Pliocene of India. Body form and habitat generally similar to a muskrat, inhabiting marshes and the borders of streams and lakes. There is only a single living genus, spread over Africa south of the Sahara. Head and body 350 to 610 mm.

Family Bathyergidae (blesmols or African mole rats)

Miocene to Recent of Africa. The most completely fossorial of all rodents, with short, powerful limbs and heavy claws. The incisors are highly procumbent, and in some forms are used in burrowing. In some genera the growing base of the upper incisors has shifted backward to a level behind the rear of the upper cheek teeth. Only slight penetration of in-

fraorbital foramen by masseter (probably secondary reduction); angle hystricognathous but peculiar. One genus (*Heterocephalus*) has the pelage reduced to scattered short hairs. External ears are greatly reduced; eyes are reduced and the eyelids usually kept closed. The number of cheek teeth reaches $\frac{6}{6}$; it is by no means clear which teeth are involved. These animals spend essentially all their lives in their burrows; some are colonial.

Families of uncertain relationships

The remaining 10 families of rodents are of completely uncertain relationships, and are best treated as individual families.

Family Hystricidae (Old World porcupines)

Pliocene to Recent of Asia, Africa, and Europe, Pleistocene to Recent of East Indies. Spined forms, resembling the New World porcupines in appearance, but less arboreal in habits; quills without barbs. The ease with which even a slight touch loosens the quills, leaving them inserted in the attacker, is the origin of the belief that these animals are capable of shooting their quills like arrows, as reported, for example, by Marco Polo. Hystricognathous and hystricomorphous; malleus and incus fused; incisor enamel multiserial; sacculus urethalis present; cheek teeth consist of $P\frac{1}{1}$, $M\frac{3}{3}$, all exhibiting a crown pattern broken up into numerous small cuspules. The origin and relationships of the group are not clear, but they may have originated in southern Asia.

Family Castoridae (beavers)

Early Oligocene to Recent of Europe and North America, late Oligocene to Recent of Asia. Sciuromorphous and sciurognathous rodents, with cheek teeth ($P\frac{(2-1)}{0}$, $M\frac{3}{3}$) progressively becoming high crowned and ever-growing. Restricted at present to a single genus (*Castor*), formerly widespread in Eurasia and North America, but greatly reduced as a result of very intensive trapping for the very fine, soft fur. Aquatic animals, with webbed feet and broad, flattened tail. Numerous fossils from the Tertiary do not seem to have been aquatic but rather to have been fossorial. In western Nebraska an early Miocene beaver seems to have constructed spiral burrows ("Devils' corkscrews").

†*Family Eutypomyidae*

Rodents from the Oligocene of North America sometimes (but probably erroneously) thought to have been related to the beavers, but with very complexly folded enamel on the cheek teeth, and with a peculiar foot structure of uncertain function (very slender inner digits and heavy outer ones).

Family Anomaluridae (scaly-tailed "squirrels")

Miocene to Recent of Africa. Sciurognathous but hystricomorphous rodents; arboreal, and most living forms with a membrane extending from front to hind legs to tail, used in gliding from one tree to another. A long cartilaginous support, derived from the ulna in the region of the elbow, supports the anterior end of the membrane. Head and body length about 60 to 430 mm (2.4–16.9 in.).

Family Ctenodactylidae (gundis)

Oligocene of Mongolia, Miocene of India, and Miocene to Recent of Africa. Hystricomorphous and sciurognathous rodents; malleus and incus fused; perhaps a sacculus urethalis; incisor enamel multiserial. Quite diverse in Oligocene of Mongolia, but the fossils do not suggest close relationship to any possible ancestors. Scamperers, with body form similar to that of a guinea pig; 4 toes on each foot. Head and body length 160 to 240 mm (6.3–9.4 in.).

Family Pedetidae (springhaas or Cape jumping hare)

Miocene to Recent of Africa. Hystricomorphous and sciurognathous; malleus and incus not fused; incisor enamel multiserial. The single living genus lives in eastern and southern Africa. The Cape jumping hare is a large rodent, with elongate hind limbs, jumping like a kangaroo. The cheek teeth ($P\frac{1}{1}$, $M\frac{3}{3}$) are ever-growing, with the pattern preserved only in completely unworn teeth. Head and body length about 350 to 430 mm (13.8–16.9 in.).

Family Gliridae (dormice)

Middle Eocene to Recent of Europe, Recent of Asia, Miocene to Recent of northern Africa. Sciuromorphous and sciurognathous; uniserial incisor enamel. Small arboreal rodents, similar in appearance to the smaller squirrels. Generally have partial hibernation. It has recently been demonstrated that they can be traced back to an ancestry among small European paramyids of early to middle Eocene; as a result, although they are similar to sciurids in being sciuromorphous, they are clearly of independent origin. Head and body length about 60 to 190 mm (2.4–7.5 in.).

Family Seleveniidae (jumping dormice)

Probably related to dormice; a single genus recently discovered in the deserts of Central Asia. Elongate hind legs and a long tail. Head and body length 72 to 96 mm (2.8–3.8 in.).

†*Family Pseudosciuridae*

European, middle Eocene to Oligocene. Hystricomorphous and sciurognathous; incisor enamel pauciserial. Cheek teeth low crowned. They can be derived from European early Eocene paramyids, and gave rise to the Theridomyidae. The gradual development of the hystricomorphous condition can be traced in this group.

†*Family Theridomyidae*

Abundant European late Eocene and Oligocene rodents, with low-crowned to ever-growing cheek teeth. Incisor enamel pauciserial or uniserial. They have had a central position in many theories of the origin and relationships of the various hystricomorphous groups (Caviomorpha, Phiomorpha, Hystricidae, Anomaluridae, Ctenodactylidae, and Pedetidae), but there is no evidence to show that they were actually related to any of these.

Critical appraisal. The rodents are one of the most clearly demarcated of all mammalian orders. No animals are known, either living or fossil, where there is question as to whether or not they belong to this order. Within the order, however, there is great disagreement as to the interrelationships of various families or groups of families. There is no question that the Caviomorpha, as listed above, are related, and there seems to be little question that the Phiomorpha are a natural group. There is disagreement as to whether the two suborders should be united with each other, or with the Hystricidae. If the three are united, the combination should be called the suborder Hystricomorpha (the oldest name for such a group). The Anomaluridae, Ctenodactylidae, Pedetidae, Pseudosciuridae, and Theridomyidae are included in the Hystricomorpha by some authors. The relationships of all five families are very uncertain (except that pseudosciurids clearly gave rise to theridomyids). Much uncertainty exists as to whether the families here included in the Myomorpha are a natural unit. There is considerable diversity within the group. Many authors unite the Muridae and Cricetidae into a single family; still others separate the microtines from the cricetids as a third family. The two families are recognized here largely as a matter of convenience—about 100 genera each of murids and cricetids are known; combining them makes an unwieldy family including over half the order. The microtines were quite obviously derived in late Pliocene times from cricetids; if they are recognized as a distinct family, it is the only family of either animals or plants known to have originated so recently.

Uncertain relationships

Currently, attempts are still being made to find satisfactory bases on which to subdivide the order. The incompleteness of knowledge of fossil rodents is a major handicap that will certainly be overcome in the future. The classification presented here is an attempt to give what seems to be a reasonable interpretation of the present knowledge of the subject. (A.E.W.)

Carnivora (cats, dogs, bears, skunks, seals, walruses)

The order Carnivora includes 10 families of living mammals: Canidae (dogs, wolves, jackals, and foxes), Ursidae (bears), Procyonidae (raccoons), Mustelidae (skunks, mink, weasels, badgers, and otters), Viverridae (civets and mongooses), Hyaenidae (hyenas), Felidae (cats), Otariidae (eared seals), Odobenidae (walrus), and Phocidae (earless seals). The term carnivore is frequently applied by mammalogists to members of this order and is employed in that sense in the present article. In a more general sense, a carnivore is any animal (or even, occasionally, a plant) that eats the flesh of other animals, as opposed to a "herbivore," which eats plants. Although the Carnivora are basically meat eaters, a substantial number, especially among bears and procyonids, feed extensively on vegetable material.

GENERAL FEATURES

Importance of carnivora. There is probably no other group of animals more familiar to man than mammals belonging to the order Carnivora. The more popular do-

mesticated pets of man, the dog and the cat, are both derived from wild members of this order. The majority of luxurious natural furs (ermine, mink, sable, otter) worn by women of fashion, as well as the furs of utility used by primitive man, come from members of the Carnivora. Many of the animals that attract the largest crowds at circuses and zoos are members of this order. Most of the large, dangerous carnivores are the objects of hunters, who wish to obtain a trophy grizzly bear, polar bear, or tiger. The man stranded in the wild fears the wolf, dhole, bear, tiger, leopard, or jaguar, virtually the only predators capable of injuring him. The stockman is concerned about possible depredations upon his herd by nature's large predators.

The carnivores, the meat eaters, form the apex of the pyramid of life, the food pyramid, and thus are basic to the existence of a "balance of nature" in a world unmolested by man. In today's world, man's world, this precarious balance was first upset by the extermination of many carnivores. Formerly considered undesirable because of their predation on game animals, carnivores are now recognized as necessary elements in natural systems. Far from depressing game numbers, carnivores improve the stability of game populations by keeping numbers within the carrying capacity of the food supply. As a result, individual animals are better fed and less subject to disease. Many of these predators dig dens and provide burrows in which other forms of wildlife can take refuge. Those carnivores best known for their burrow building are the badgers and the skunks. Digging results in the mixing of soils and the reduction of the runoff of water during rains.

Intelligence and training

Carnivores rank high on the scale of intelligence among mammals. The large size of the brain, compared to that of the animal, is an indication of their superior mental powers. For this reason, these animals are among the easiest to train for entertainment purposes, as pets, or as hunting companions. Their highly developed sense of smell supplements the sharper vision of man. Dogs are the most common carnivores trained for hunting, but the cheetah, caracal, and ferret have also been used to some extent. In China the otter is trained to drive fish under a large net, which is then dropped and pulled in. Carnivores, dependent for survival upon their ability to prey upon living animals in a variety of situations, have evolved a relatively high degree of learning ability (see LEARNING, ANIMAL).

Carnivorous mammals tend to establish territories, areas defended against others of the same species; herbivorous ones, which eat vegetation, are less apt to do so. Territories are often exclusive, defended by the residents against other animals of their own kind. Such areas may sometimes be marked by secretions produced by anal or scent glands.

There is a wide range of social patterns among carnivores. Some (bears, raccoon, red and gray foxes, genets, most cats, and most mustelids) are strictly solitary, except during the breeding season. Some remain paired throughout the year (black-backed jackal and lesser panda) or occasionally hunt in pairs (gray fox, crab-eating fox, and kinkajou). Other carnivores, such as the wolf, African hunting dog, dhole, and coatimundi, normally hunt in packs or bands. Still others form sedentary colonies during the breeding season (sea lions, fur seals, and elephant seals), during a somewhat larger part of the year (sea otters), or all year round (meerkats).

Form and function. The smallest living member of the Carnivora is the least weasel (*Mustela nivalis*), which weighs about two ounces. The largest terrestrial form is the Alaskan grizzly (*Ursus arctos*), weighing up to 780 kilograms (1700 pounds). The largest aquatic form is the elephant seal (*Mirounga leonina*), which may weigh 3,640 kilograms (four tons). Most carnivores weigh between four and eight kilograms (nine and 18 pounds).

Most members of the order are terrestrial. Some, such as the sea otter, river otter, and polar bear, spend most of their lives in water. The pinnipeds, or seals, are more aquatic than other members of the order. Aquatic or semiaquatic forms tend to have body specializations such as a streamlined body and webbed feet for this mode of life.

Dentition

Carnivores, like other mammals, have a number of different kinds of teeth: incisors in front, followed by canines, premolars, and molars in the rear. Most carnivores, especially those that feed exclusively on meat, have carnassial, or shearing, teeth that function in slicing meat and cutting tough sinews. The carnassials are usually formed by the fourth upper premolar and the first lower molar, working one against the other with a scissorlike action. Cats, hyenas, and weasels, all highly carnivorous, have well-developed carnassials; while the bear and procyonids, which tend to be omnivorous (eating both plants and animals), and the seals, which eat fish or marine invertebrates, have little or no modification of these teeth for shearing. The teeth behind the carnassials tend to be lost or reduced in size in highly carnivorous species. Most members of the order have six prominent incisors on both the upper and lower jaw, two canines on each jaw, six to eight premolars, and four molars above and four to six molars below. Incisors are adapted for nipping off flesh. The outer most incisors are usually larger than the inner ones. The strong canines are usually large, pointed, and adapted to aid in the stabbing of prey. The premolars always have sharply pointed cusps, and in some forms (*e.g.*, seals) all the cheek teeth (premolars and molars) have this shape. Except for the carnassials, molars tend to be flat teeth utilized for crushing. Terrestrial carnivores that depend largely on meat, such as weasels, cats, and hyenas, tend to have fewer teeth (30–34), the flat molars having been lost. Omnivorous carnivores, such as raccoon and bear, have more teeth (40–42). Seals have fewer teeth than terrestrial carnivores. In addition, seals exhibit little stability in the numbers of teeth; for example, a walrus may have from 18 to 24 teeth.

Several features of the skull are characteristic of the order Carnivora. The articulating surfaces (condyles) on the lower jaw are transverse, their axis at right angles to that of the head, forming a half-cylindrical hinge that allows the jaw to move only in a vertical plane but with considerable strength. The clavicles (collarbones) are either reduced or absent entirely and, if present, are usually embedded in muscles without articulation with other bones. This allows for a greater flexibility in the shoulder area and prevents breakage of the clavicles when the animal springs on its prey.

The brain is large in relation to the weight of the body and contains complex convolutions characteristic of highly intelligent animals. The stomach is simple, and a blind pouch (cecum) attached to the intestine is usually reduced or absent. Since animal tissues are in general simpler to digest than plant tissues, the carnivore's dependence on a diet with a high proportion of meat has led to less-complex compartmentalization of the stomach and a decrease in the length and folding (surface area) of the intestine. The teats are located on the abdomen along two primitive lines (milk ridges), a characteristic of mammals that lie down when nursing.

Distribution and abundance. Carnivores are found worldwide. Terrestrial forms are absent from most oceanic islands, though the coastlines are usually visited by seals (pinnipeds). Except for the dingo (which probably was introduced by aboriginal man), Australia has no native terrestrial members of the Carnivora. Man has taken his pets, as well as a number of wild species, to most islands. For example, a large population of the red fox now inhabits Australia, having been introduced there by fox hunters.

Since carnivores are large and depend on meat, there must be fewer carnivores in the environment than those animals that form their diet. The maintenance of established territories limits the number of predators to the carrying capacity of prey populations. In general, carnivores have a population density of approximately 0.4 per square kilometre (one per square mile). By comparison, omnivorous mammals average about eight per square kilometre (20 per square mile), and herbivorous rodents attain densities of up to 40,000 per square kilometre (100,000 per square mile) at peak population. Existing at relatively low densities, carnivores are vulnerable to prey, population fluctuations, habitat disturbance, and predation by man. In some cases the mobility and adaptability of carnivores has enabled species to shift ecological roles and survive the changes brought about by human activities; in other instances less-flexible species have become extinct.

World carnivore population

The world population of aquatic carnivores (pinnipeds) has been estimated to be between 13,000,000 and 27,000,000 animals. Estimates of the total numbers of terrestrial carnivores have not been made. The yearly sale of furs indicates a harvest of at least 6,000,000 pelts of wild carnivores; this is equalled only by the number of mink raised and harvested on fur farms (approximately 8,000,000 annually in the mid-1960s). There are probably around 50,000,000 to 60,000,000 terrestrial members of the Carnivora present today. Some species (many cats, pandas, and bear, some seals) are becoming quite rare and near extinction.

SURVEY OF CARNIVORE FAMILIES

Dogs and allies (family Canidae). *Natural history.* Canids are basically meat eaters, although some vegetable matter is taken. The gray or timber wolf (*Canis lupus*), the African hunting dog (*Lycaon pictus*), and the dhole, or wild dog of India (*Cuon alpinus*), are strictly carnivorous. The various foxes and jackals, coyotes, and the raccoon dog eat whatever food is abundant—small mammals, birds, insects, crustaceans, mollusks, fruits, or berries. The canids that are strictly carnivorous tend to hunt in packs; those that are omnivorous tend to be solitary in their hunting habits. Since the carnivorous forms depend primarily on one or a few prey species, they usually follow the moving herds of caribou or antelope, on which they feed, or move into areas where the prey species are more numerous. African hunting dogs are extremely social, always hunting in packs so intricately organized that some researchers doubt that an individual dog could survive alone. The varied diet of the omnivore reduces the necessity for organized attack on the prey species and for extensive movement.

Canids have the ability to endure a continuous chase pattern with exceptional stamina but are not capable of great bursts of speed. Their senses of smell and hearing are highly developed. Smell is used largely to track prey, and hearing to warn of impending danger. Smell is also used in association with the demarcation of territory, which tends to be outlined by frequent urination at established scent posts. Sight is less well developed, but movement is noted quickly. Canids are highly intelligent, can be trained easily, and form a basis for such sayings of man as "sly as a fox." From a behavioural standpoint wild canids are cunning or crafty, even vicious or treacherous by human standards, but usually cowardly or furtive unless running in a pack.

Reproductive behaviour of canids

Canids give birth in a den in the ground, in a hollow log or tree, in a hidden brushy area, among boulders, or in a crevice of rock. The African hunting dog often dens in old diggings of the aardvark. Usually one litter of from four to six young is produced each year after a gestation period of from 51 to 80 days, depending on the species. The Arctic fox, *Alopex lagopus,* may produce two litters in a season. Most breeding takes place late in the winter, the young being born in middle or late spring. Canids living in the northern regions breed a month or so later than those in the southern regions. Carnivorous animals tend to produce young soon after the peak in the prey reproduction, a time of abundant food availability. Thus the timing of the breeding season and the duration of the gestation period must be coordinated with the season of available food in order that the young be produced when the most food is available for their rapid growth. The eyes of the young usually open in about two weeks, and they nurse for from four to six weeks. Canids of the smaller species can begin production of their own young when only one year old, but the larger forms, such as the wolf, reach sexual maturity at two or three years of age.

Canids are basically adapted to running and do so on their toes (a mode of walking or running called digitigrade). They live in a variety of habitats but generally tend to be animals of the open or grassland areas where their prey species are more abundant. Only the rare bush dog (*Speothos venaticus*) confines itself to forested areas. The red fox (*Vulpes vulpes*) tends to be an animal of the forest edge, and the gray fox (*Urocyon cineroargenteus*) an animal inhabiting wooded areas. Thus, in North America, where both forms exist, these foxes live in slightly different niches. Although basically terrestrial in habits, the gray fox is not averse to climbing trees. The raccoon dog (*Nyctereutes procyonoides*) often lives in tree hollows if the entrance is close to the ground. The family Canidae is worldwide, being absent only from New Zealand, Antarctica, and most oceanic islands. The Arctic fox is found farther north than any other strictly terrestrial mammal. Every major habitat has some type of canid, from the Arctic tundra, through the desert and grassland, to the tropical forest.

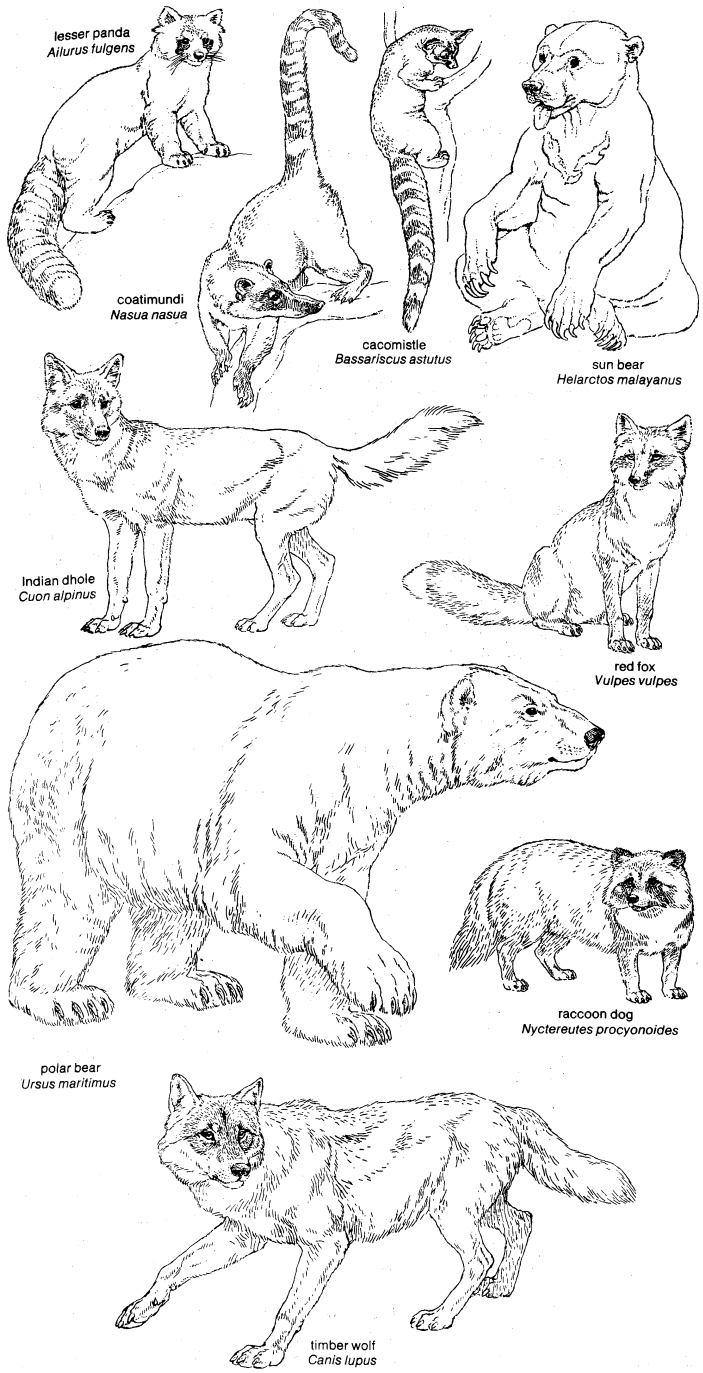


Figure 33: Representative carnivores: Procyonidae, Ursidae, Canidae.
Drawing by R. Keane; Indian dhole based on Zoological Society of London photograph

Canid communication

Canids communicate with a variety of sounds. The vocal repertoire, most highly developed in social species, includes howls, yelps, snarls, barks, and growls. These sounds are frequently associated with specialized visual signals involving movements of the ears and tail, raising of certain areas of the fur, and baring of the teeth. Within the social group or pack there is a complex dominance hierarchy,

involving age levels, pair bonds, physical condition, and sexual state. Vocal and visual signals serve to minimize aggressive interactions, such as quarrels over food, that might prove injurious to the group as a whole and to the individual members. In solitary species vocalizations serve to advertise the territory, to ward off aggressors, and to communicate with the mate and young.

Importance to man. The domestic dog (*Canis familiaris*) is undoubtedly the canid of the greatest importance economically. Numerous humans are actually employed in raising, selling, doctoring, grooming, training, or providing and manufacturing food for "man's best friend." Dogs probably were the first wild animals to be domesticated; they have been found associated with Neolithic sites dating back some 8,000 years.

Some canids are important to the fur trade. At one time a mutant form of the red fox, called the "silver fox," formed a significant part of the fur-farming industry, with a breeding pair bringing as high as $40,000. The raccoon dog is still raised in some regions of the Soviet Union. The more important canids whose fur is utilized are the raccoon dog, red fox, grey fox, Arctic fox, and timber wolf. Coyotes, kit and swift foxes (*Vulpes macrotis* and *V. velox*), jackals (*Canis aureus, C. adustus, C. mesomelas*), Corsac fox (*Vulpes corsac*), hoary fox (*Vulpes cana*), dhole, South African silver fox (*Vulpes chama*), Cape fox (*Octocyon megalotis*), maned wolf (*Chrysocyon brachyurus*), crab-eating fox (*Cerdocyon thous*), and the South American foxes (*Dusicyon* species) are occasionally used by the fur trade. Other canids, especially the red fox, are hunted for sport. Some hunters call or lure foxes up to them by using an imitation of the sound of an injured rabbit to attract the attention of the fox.

Canids also are basically important as a natural control over rodent populations. Livestock is occasionally endangered when not properly protected from the forays of wild canids. Individual foxes learn where an easy meal is located, if the farmer does not have his chickens properly housed. In many cases all individuals of the species are condemned because of the damaging activity of a few. Proper selective control of those individuals actually doing the damage is the only sane conservation practice. Some species also form important natural reservoirs for parasites, causing a number of diseases, especially the virus producing rabies. Not many canids are utilized by man for food, but the American Indian once regarded dog meat as a delicacy to be eaten on special occasions.

Form and function. Canids have unspecialized incisor teeth, and large fanglike canines used to kill their prey by slashing. The premolars are narrow and sharp and the carnassials well developed; the molars form broad surfaces that can crush substantial bones. Most canids have the basic dental formula of terrestrial carnivores, with a total of 42 teeth.

Distinguishing physical features of canids

The possession of a long face or muzzle is characteristic of wild canids. All have a relatively long and obviously bushy tail. The ears are pointed, held erect, and often quite large in desert species. As well as functioning for sound detection, these large ears are believed to act as heat regulators, allowing a greater amount of heat to be dissipated in hot climates. Arctic foxes tend to have small ears, providing less loss of heat in a region where heat conservation is important to survival. Most canids have relatively long legs (especially long in the maned wolf, *Chrysocyon brachyurus,* of South America). The long legs and digitigrade gait are special adaptations for ease of running. Canids have four well-developed toes, plus a dew claw (reduced hind toe) on the front foot (except in the African hunting dog, which lacks the dew claw), and four toes on the rear foot. Each toe is capped by a blunt, nonretractile claw (*i.e.,* with no sheath into which it can be withdrawn). The short claws are not adapted for use as a weapon but do aid the animal during the digging of a den or during pursuit of small prey animals in their burrows. Scent glands are often present at the base of the tail. Most canids have a uniform coloration, although there are some contrasting colours on the gray fox, a raccoon-like mask on the raccoon dog, a blotching of black, yellow, and white colours on the African hunting dog, and a lighter coloured belly in most canids.

Bears (family Ursidae). *Natural history.* Members of the Ursidae, except for the carnivorous polar bear (*Ursus maritimus*), are omnivorous, consuming many items that seem small for an animal the size of a bear. Ants, bees, seeds of trees, roots of the skunk-cabbage, nuts, berries, insect larvae such as grubs, and even the dainty dog-toothed violet are eaten. Many bears relish honey, and the sun bear (*Helarctos malayanus*) is sometimes called the "honey bear" because of its preference for this food. Meat items taken by bears include rodents, fish, deer, pigs, and lambs. Grizzlies and Alaskan brown bears (North American subspecies of the widespread *Ursus arctos*) are known for their skillful fishing abilities during the spawning runs of the salmon. The polar bear feeds almost exclusively on seals and is the most carnivorous of all bears, but little vegetation grows within its range, so the carnivorous habit is dictated by the Arctic environment. The sloth bear (*Melursus ursinus*) delights especially in raiding and destroying termite nests, sucking up termites and larvae with its funnel-like lips. Bamboo shoots form the major food of the giant panda (*Ailuropoda melanoleuca*), which has a special bone formation of the forefoot that functions as a sixth digit, opposable to the other five and thus useful in handling bamboo.

Moving slowly about, bears seem to be sluggish and oblivious to much that is going on around them, but once they are disturbed their reactions may be violent and highly unpredictable. Although clumsy in appearance, a grizzly can move surprisingly fast, even through dense cover that would seriously impede a man or a horse. Their senses of sight and hearing are poorly developed, and most hunting is done by the sense of smell. Thus the smell of bacon kept in camp is an invitation to a bear in search of food. Bears, like many of their canid relatives, are solitary, except during the mating season.

Bears tend to congregate during the mating season and then pair off and mate in seclusion. Males play no role in raising the young, leaving the female soon after mating. The gestation period may vary, the fertilized egg remaining dormant in the uterus (delayed implantation), insuring the birth of young when food is abundant. Ursids breed only once a year. With a breeding season usually in late spring or early summer, and delayed implantation, most young are born in January or February when the female is in the winter den. Newborn grizzlies weigh about one pound and are about nine inches long from the tip of the nose to the tip of the short tail. The Himalayan brown bear (a race of *U. arctos*) have their young later (in April and May) than the North American populations, but the polar bear has the greatest variation of any bear population in the dates when young are born (January to April). Twins are most common in bears, but up to five young may be produced. The cubs nurse for about two months and stay with the female until the next breeding (about a year and a half after birth). Most young, however, can get along on their own when about six months old. Bears reach breeding condition at three and one-half to four years of age, males usually a few months later than females. Litters are produced every second year after the initial litter.

Winter sleep of bears

Most bears eat large amounts of food before entering a den for a period of deep sleep during the winter. The polar bear digs a den in the snow. Grizzlies build large mounds of dirt in front of their dens. Bears are not true hibernators, since they lack the physiological characteristics (lower heart rate, body temperatures, breathing rate, and blood pressure) exhibited by those animals that do hibernate. A true hibernator is slow to be awakened from its deep sleep, and those who believe bears hibernate need only kick a sleeping bear to become convinced to the contrary. Bears usually are quiet, but they do growl at times when feeding or when being challenged by another bear.

Ursids do not roam over great distances even though they are rather large carnivores. Black bears (*Ursus americanus*) tend to stay within an area of about 36 square kilometres (14 square miles), grizzlies within an area 16 kilometres (10 miles) in diameter, or about 200 square kilometres (80 square miles). A grizzly marks the boundary of its

territory each spring by rubbing trees, scratching bark, or even biting large pieces from the trunks of trees. All bears, except the grizzly, Eurasian brown, and probably the polar bear, climb trees. Bears inhabit forested areas and areas that have a minimum of disturbance by man. At one time the grizzly was an animal of the plains and was one of the mammals reported by Lewis and Clark in their journey through eastern Montana in 1804. Polar bears have been noted in the open sea 65 kilometres (40 miles) from the nearest shore and on ice 320 kilometres (200 miles) from the nearest land.

Ursids are largely animals of north temperate regions, being found farther north than any other mammal (only the Arctic fox is found as far north on land). Most of the southern continents, Central and South America, Africa, and Australia, lack bears. The bear that formerly occurred in the Atlas Mountains of Morocco (*Ursus crowtheri*) became extinct in the early 19th century. The spectacled bear (*Tremarctos ornatus*) of the Andes is the southernmost form.

Importance to man. If taken when young, the grizzly can be tamed quite easily and is not uncommon in circus acts. Unfortunately, man has come to picture the bear as being quite tame and harmless, with the result that the bear is viewed without the respect that this potentially dangerous creature deserves. The grizzly is the most dangerous of the bears to man, of the most interest economically because of the damage that it can do to livestock, and of the most interest to the big-game hunter as a prized trophy. Some ursids, such as the Asiatic black bear (*Selenarctos thibetanus*), destroy fruits. At times the North American black bear does considerable damage to fruit crops and to corn fields when the corn is in the milk stage, pulling the stalks into a pile from as far as the bear can reach.

Use of bear pelts

The pelts of bears have been used for a number of purposes. Perhaps most popular has been the bearskin rug. Skins also have been used for lap rugs or sledge rugs, hats for the English Guard regiments, trimmings of coats, and for muffs. Skins of the rare giant panda commanded a high price in the early 1900s, but this animal apparently is close to extinction today. Black bears are prized as a food source in China, as is the flesh of the polar bear in the far north. The liver of the polar bear, however, is highly poisonous, due to high levels of vitamin A. The teeth and claws of bears have been favourite ornaments with the Indians and Eskimos for years, and the fat furnishes the "bear grease" of commerce.

Form and function. The teeth of the omnivorous bears are unspecialized. The first three premolars are usually either missing or extremely small. The carnassials are poorly developed, and the molars have broad, flat crowns. Except for variability as to the presence of premolars, the ursid dental formula is that of the Carnivora generally. The sloth bear lacks one pair of upper incisors. The jaw of the bear is controlled at the hinge by a powerful set of muscles.

Bear claws

Bears have an elongate skull that is especially heavy in the back portion. These plantigrade animals (plantigrades walk on the sole and heel of the foot) are powerful in build and have relatively short, massive legs with five toes on each foot. The toes end in enlarged, nonretractile claws that are especially well developed in the sloth bear and to some extent in the grizzly. The claws on the front feet are usually better developed than those on the rear and are especially adapted for digging out marmots or other small rodents. Bears, however, seldom use the claws to construct burrows. They generally have naked soles, but those of the polar bear are covered with hair, enabling the animal to walk on ice with a firm footing. In swimming, the polar bear uses only its front limbs, an aquatic adaptation found in no other four-legged mammal. Bears lack a clavicle but have a baculum (penis bone). Their lips, which are not attached to the gums, are protrusible and mobile. Ursids have a short (7 to 13 centimetres; 2.8 to 5.1 inches), stubby tail and are the largest living terrestrial members of the Carnivora. The Alaskan grizzly may weigh as much as 780 kilograms (1,700 pounds), and the smallest bear, the sun bear, weighs approximately 27 kilograms (60 pounds). In most species the two sexes are about the same size, but the male tends to be larger in the grizzly and polar bear. Bears of the genus *Ursus* tend to be of a uniform colour, ranging from white to brown, blue gray, and black. Unlike the majority of carnivores, ursids are often black or mostly black. The North American black bear occurs in a number of colour variants or phases; in addition to the typical black phase, there is a tan-brown phase (called the cinnamon bear) and a bluish-black phase (called the glacier bear). The spectacled bear and the several specialized Asian bears are black with white or tan markings. The giant panda is strikingly black and white, and the remaining genera largely black, with areas of white or orange on the chest, neck, or face. The fur is coarse, long, and shaggy in most northern forms.

Raccoons and allies (family Procyonidae). *Natural history.* The family Procyonidae includes raccoons, coatis or coatimundis, olingos, and the ringtailed cat, kinkajou, and lesser panda. Procyonids, like bears, are omnivorous, eating numerous types of insects, crayfish, crabs, fishes, amphibians, reptiles, birds, small mammals, and a variety of fruits, nuts, roots, and young plants. A characteristic of most omnivorous animals is that they take whatever foods are available, varying their food habits with the season, the locality, and the abundance or population levels of the food items from one year to the next. The ringtailed cat (*Bassariscus astutus*) tends to be more carnivorous than the raccoon (*Procyon lotor*). The lesser panda (*Ailurus fulgens*) is the most vegetarian of the group.

The various senses of procyonids do not seem to be especially well developed, but the raccoons and lesser panda appear to be above average in intelligence and are often described as being cunning. Raccoons are well known for their insatiable curiosity and ability to use their "hands" in performing feats such as opening doors and getting into mischief when kept as pets. This curiosity of raccoons is well known by trappers, who have used bright objects placed on the treadle of a trap to attract them. The sense of touch probably is of the greatest importance to the raccoon in its food-gathering activities. Most procyonids, except perhaps the coatis (*Nasua* species), are more active in the twilight and evening hours than in the daylight. Procyonids tend to be solitary, although families often move and feed as a group. Coatis travel in bands, some individuals moving through the trees as others follow on the ground. Olingos (*Bassaricyon* species) also move about in small bands. All procyonids are arboreal and move around in trees with agility. On the ground, the raccoon resembles a small bear as it ambles along.

Procyonid breeding habits

Procyonids breed in late winter or early spring. During copulation, the female raccoon makes sharp rattling cries. After a gestation period of about two months (for the raccoons 54 to 66 days, for lesser panda 90 days), two to six young are produced. One litter a year is usual. Young raccoons "twitter" like young birds when disturbed in the nest, and the young of other procyonids also produce high-pitched calls. Most adult procyonids produce a variety of snarls, growls, whines, screams, and barks. Most of these calls have little carrying power. Young raccoons are dependent on the female for the first 10 weeks but may nurse until approximately 20 weeks of age (ringtails to about four months) and may stay in the family group into winter. Being mainly arboreal, young procyonids are born in tree dens. Raccoons use old buildings or dens in the ground if tree dens are scarce. Ringtails use crevices in rocks.

Except for the lesser panda, which is found in the southeastern Himalayas of Asia at elevations from 1,980 to 3,650 metres (6,500 to 12,000 feet), procyonids are found naturally only in the New World. They are largely mammals of the lower elevations of temperate and tropical regions, in areas well supplied with water and trees. Raccoons have been introduced successfully into many parts of Europe (Germany, White Russia, the Ukraine) and Asia (Lake Balkhash, Siberia).

Importance to man. The raccoon is the only procyonid of major economic significance. Raccoon hunting has been a sport of substantial importance in the United States, and in some parts of the United States many people own dogs trained to hunt raccoons. The coatimundi also is hunted with dogs and its flesh is also eaten. Meat of the raccoon is

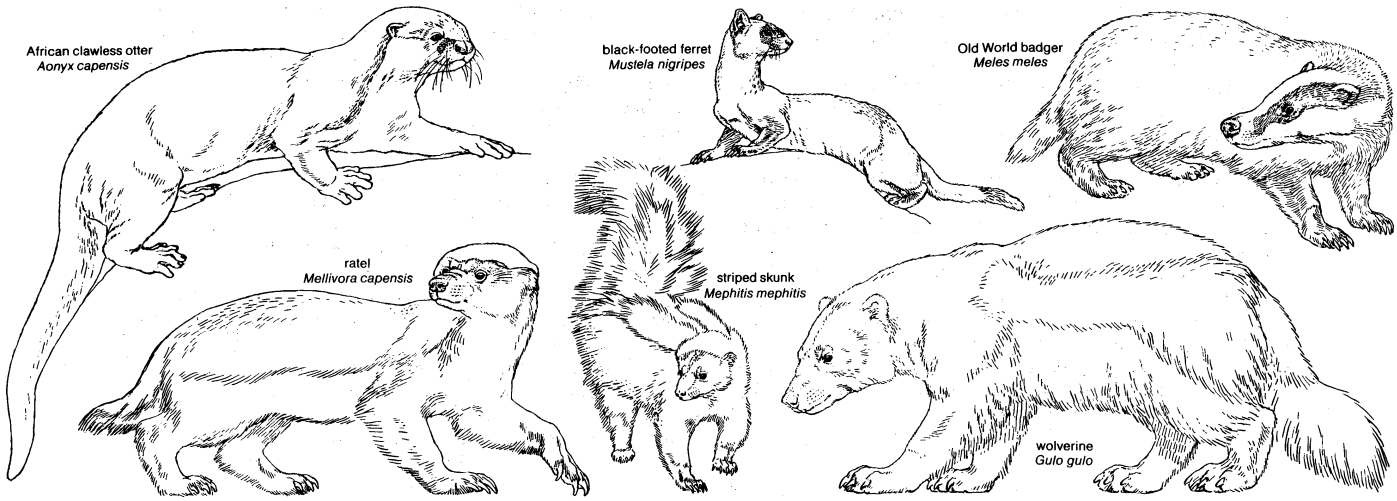


Figure 34: Representative carnivores: Mustelidae.
Drawing by R. Keane; striped skunk based on E.P. Walker, *Mammals of the World*

in more demand as food than that of any other carnivore. The meat is roasted, fried, stewed, barbecued, fricasseed, and made into patties. The "coonskin" cap has become part of American legend. The fur of the raccoon is one of the staple items of the fur trade in the United States today, as it has been over many of the past 100 years. Also used are the furs of the ringtailed cat, kinkajou (*Potos flavus*), and the lesser panda.

Raccoons at times do damage to fields of corn and to abandoned buildings, which they frequently inhabit when natural dens are scarce. In certain regions, raccoons become a problem to nesting waterfowl, especially wood ducks and goldeneye, which, like the raccoons, raise their young in hollow trees.

Most procyonids make interesting pets. Raccoons, kinkajous, and coatis are frequently sold as house pets, but their inquisitive nature and climbing ability make them troublesome if allowed to run loose indoors.

Form and function. The unspecialized teeth of procyonids are adapted for an omnivorous diet. The elongate canines are oblong in cross section, and the premolars are small, narrow, and pointed. As in bears, the carnassials are poorly developed and the molars are broad. Procyonids have the basic dental formula of the Carnivora, except that there is one less lower molar.

Procyonids are plantigrade or semiplantigrade, with legs of moderate length and five flexible toes with nonretractile or (in the ringtail cat and lesser panda) semiretractile claws. The tail is well developed (20 to 70 centimetres [7.9 to 27.6 inches] long), bushy, and in all forms except the kinkajou ringed by light and dark bands. In the kinkajou, it is prehensile, gripping a branch like a fifth limb. The anal scent glands of the kinkajou and lesser panda produce a musky odour when the animal becomes excited. The snout is extremely flexible in the coati and somewhat movable in other procyonids. The members of this family are medium in size, ranging from 0.8 to 22 kilograms (1.75 to 48.4 pounds) in weight (the record weight of a raccoon is 30.2 kilograms [66 pounds, 6 ounces]). There is no difference in size between males and females. In coloration the procyonids range from gray or brown to the rich, reddish brown of the lesser panda.

Mustelids (family Mustelidae). *Natural history.* The family Mustelidae contains a variety of animals unmatched by any other family in the Carnivora except the civets (Viverridae). The family includes the weasels, ferrets, mink, marten, fisher, skunks, wolverine, otters, badgers, and a number of less well-known animals, a total of about 70 species in 25 genera. The weasels, mink, and polecats (*Mustela* species) are fierce predators, feeding exclusively on small mammals and birds, hunting and trailing their prey with a keen sense of smell. At times, bird eggs, salamanders, snakes, and insects are taken. The smallest mustelid, the least weasel (*M. nivalis*), consumes a third to half its body weight per day, killing the prey by biting at the base of the skull. Weasels often store food in underground caches. The Asiatic polecat (*M. eversmanni*) and the North American black-footed ferret (*M. nigripes*) feed almost exclusively on the animals whose colonies they inhabit, the polecat on marmots and ground squirrels, the ferret on prairie dogs. Both of these animals occasionally feed on insects. Mink (*M. vison* and *M. lutreola*) feed largely on aquatic animals (muskrats, crayfish, mollusks, and fish) as well as on organisms inhabiting the banks of bodies of water.

Feeding habits of mustelids

The fisher (*Martes pennanti*) feeds mainly on mammals, but a few nuts, birds and eggs, fish, and seeds are taken. This animal is noted for its ability to kill and eat porcupines. Small terrestrial mammals make up over 90 percent of the food of the American marten (*M. americana*). The Eurasian pine marten (*M. martes*) and the stone or beech marten (*M. foina*) feed largely on small mammals but in the autumn may turn to wild berries and fruit. The yellow-throated marten (*M. flavigula*) of Asia is one of the largest martens and feeds on the young of larger mammals such as musk deer, wild boars, roe deer, spotted deer, and raccoon dogs, as well as on squirrels, rabbits, and the smaller rodents. These large mustelids also take a variety of other food items: mollusks, grasshoppers, spawning fish, birds and their eggs, cedar nuts, grapes, and berries.

Badgers are quite variable in their diet. About 50 percent of the diet of the American badger (*Taxidea taxus*) is composed of ground squirrels, which are captured in their burrows. Badgers can dig with remarkable speed and can overtake the rodent when the latter has taken refuge in a blind tunnel. The remainder of their food is made up of small rodents, rabbits, birds and their eggs, snakes, turtle eggs, and insects. Insects, in fact, form a staple food of the young badger. The Eurasian badger (*Meles meles*) and the ferret badgers (*Melogale* species) are more omnivorous than the American badger and prefer insects (especially bees) and their larvae (grubs). Earthworms, mollusks, crustaceans, frogs, lizards, birds, eggs, as well as vegetable matter, such as berries, nuts, grains, fruits, bulbs, and green vegetation, are also taken. Ratels or honey badgers (*Mellivora capensis*) eat small mammals, lizards and snakes, insects, fruit, and especially honey. An interesting association exists between the African honey guide (a piciform bird) and the honey badger. When the bird finds honey it gives a special call, which attracts the badger. The latter breaks open the hive with its long, sharp claws, eats the honey, and leaves the wax and bee larvae for the bird. If the bird does not succeed initially in attracting a honey badger, it seeks one out and, using the "honey" call, which the badger recognizes, leads it to the bee tree.

Skunks (*Mephitis, Spilogale, Conepatus*) are omnivorous mustelids, with small mammals forming an important part of the diet. Skunks may be becoming scarce in some parts of North America due to poisons gradually accumulated through extensive feeding on insects poisoned in agricultural control programs.

Otter feed largely on the invertebrates and vertebrates (especially fish) found in their aquatic environment. The oriental small-clawed otter (*Aonyx cinerea*) and the clawless otter (*Aonyx capensis*) feed to a greater extent on invertebrates (clams, snails, crabs) than do the river otter (*Lutra* species). *Lutra canadensis* at times take large numbers of crayfish. The African small-clawed otter (*Aonyx* species) are believed to feed on small mammals, the eggs of birds, and amphibians more than on fish. The sea otter (*Enhydra lutris*) is perhaps known best for its habit of breaking open sea urchins and clams on a rock held on its chest while it floats on its back. Seaweed, cuttlefish, and small fishes also are eaten.

Sharp senses of mustelids

Smell apparently is the most important sense to mustelids hunting on land. Black-footed ferrets have been noted to stop repeatedly and sniff the air while hunting. Hearing is also well developed and seems to be utilized as a means of detecting danger. Sight is less developed and, in the ferret at least, of little importance beyond about 90 metres (300 feet). Of most importance to otters, living in an aquatic environment, is an acute sense of touch, through their very sensitive whiskers, which they use as an aid to guide them through the water. Most mustelids are silent, seldom making more than a squeak, whistle, or bark. Many will snarl or growl when annoyed. The giant otter (*Pteronura brasiliensis*) of South America is perhaps the most vocal, with its high-pitched whistle. Yellow-throated martens give a harsh cry when excited and often chuckle in a low tone.

Almost all mustelids are active both day and night, although most of their activity is nocturnal. Only the Old World badgers (*Meles meles*), ferret badgers (*Melogale* species), spotted skunks (*Spilogale* species), and hog-nosed skunks (*Conepatus* species) seem to be almost strictly nocturnal. The giant otter and sea otter seem to be more diurnal than other mustelids.

Except when travelling with young in a family group, most adult mustelids are solitary in habits. Sea otters seem to be more or less gregarious, as are Old World badgers. Badgers build a labyrinth of underground passages that some call "badger cities," because of the extent of the burrows and number of badgers that inhabit them. Female skunks (*Mephitis* species) often den together in large groups during the winter. Grisons (*Galictis* species) and yellow-throated martens often travel in groups of five or six individuals. Tayras (*Eira barbara*) and ratels tend to travel in pairs.

Since a great variation exists in the habits of mustelids, a great diversity would be expected in the habitats in which they live. Most mustelids, however, are terrestrial, living in forested or brushy areas. Many species can survive in a variety of habitats, from forest to desert. Some, like the marten and fisher, are strongly arboreal. Others, like the otter, are largely aquatic. The New World skunks and the North African spotted weasel (*Poecilictis libyca*) tend to inhabit areas farmed by man. The ferrets, American badger, and Eurasian polecats live mainly in grassland or semi-arid areas. Wolverines live in the more northern areas of taiga and forested tundras.

The locomotion of mustelids also varies. Most small mustelids scamper across the ground, making intermittent bounds as they move along. The larger forms sometimes lumber slowly. During their travels, mustelids tend to stop, sit up on their haunches, and look over the area. Some are extremely agile in trees; others are excellent swimmers.

Delayed implantation is perhaps more characteristic of the mustelids than any other group of the Carnivora. Members of the genus *Mustela* have gestation periods (determined from the date of fertilization) ranging from as much as 220 to 337 days (long-tailed weasel, *M. longicauda*) or as little as 36–42 days (ferrets). In the genus *Martes* the period is highly variable, between 220 and 297 days. The gestation period of badgers (*Taxidea, Mellivora, Meles*) ranges from six to eight months and that of otter from eight months to a little more than a year. Skunks (*Mephitis* species) do not have delayed implantation and apparently have the least variation in the period of gestation (62 to 72 days). Mink (*Mustela vison*) usually produce young in 45 to 50 days, with extremes from 38 to 76 days. On mink ranches, the later in the year that mink are bred, the shorter the gestation period.

Mustelid reproduction

The breeding season usually is in the late winter or early spring (mink, skunks, polecats, otters) or in the summer (weasels, marten, wolverine, sables, European and American badgers). Some mustelids breed at almost any time of the year (least weasel, river and sea otters). Those with exceptionally long gestation periods breed soon after the litter is born or in the following year. Some forms have two litters a year (striped skunk, least weasel, ratel), but in most only one litter is produced. Average litter size varies from one young in the sea otter to between eight and 12 young (up to 18) in the Asiatic polecat. Most mustelids have between three and five young in a litter. The young of mustelids are born in ground dens, in rock crevices, under the roots of trees, under haystacks or buildings, or in hollow trees (the pine and American martens). The young of most species are weaned when between six and 10 weeks of age, but young river otters nurse for about four months. Wolverine young are allowed to remain with the female for two years before establishing their own territories. In most species, however, the family disperses near the end of the summer or early fall. Mustelids are found worldwide, except for most oceanic islands, Australia, and Madagascar. Introductions have been made on some of the larger islands, such as New Zealand. Mustelids, like bears, generally are more abundant in the more northern continents.

Importance to man. Of all mammalian families, the Mustelidae contains the greatest variety of valuable mammals utilized by the fur trade. Fur coats that command the highest prices are usually those made from these luxuriant furbearers, especially mink, ermine (short-tailed weasel), and sable (a name applied by the fur trade to several species of *Martes,* especially the Eurasian *M. zibellina,* but also the American *M. americana*). Excellent skins are produced by the marten, fisher, skunk, and otter. In fact, the durability of all furs, including those from other mammalian orders, is based on the river otter, which has been given the top rank of 100. Pelts of the ermine and sea otter have been the furs of royalty for hundreds of years. More than $100,000,000 worth of ranch-mink skins alone are produced yearly, and some breeds (like the Kojah mink) brought $2,700 per skin in the late 1960s. During the 18th and 19th centuries a collar made from the sea otter was regarded by some as the highest of status symbols, equivalent to owning a Rolls Royce or private yacht today. In 1968, at the first sale of sea-otter skins in 55 years, the better quality skins brought $2,300 apiece.

The pelt of the wolverine is prized highly by the Eskimos and others in Arctic areas because, when used for the lining of a parka, it can be cleared of frost from breath by a sweep of the hand. Moisture does not freeze to wolverine hair as it does to other furs. Badger hair has been utilized in the past to manufacture shaving brushes. Hair from the tails of many species is in high demand for paint brushes. Among the Chinese, the tail of the kolinsky or China mink (*Mustela sibirica*) is used for making delicate paint and artists' brushes and is in more demand than is the rest of the pelt. The tails are removed from the skins and sold by the pound. Hairs from the tail of the sable also make excellent brushes, since the taper of the individual hairs causes the brush to form a sharp point when wet (see also FURS, LEATHERS, AND HIDES).

Importance in the food chain

Being largely carnivorous, the mustelids form an important link in the biological food chain. Rodents are kept at a reasonable level by this natural control. The spotted skunk, ferret badger, long-tailed weasel, Patagonian weasel, zorille (*Ictonyx striatus*), and sometimes the striped skunk (*Mephitis mephitis*) have been encouraged to live in or near the habitations of man because of their ability to keep the rodent and pest insect populations under control. Systematic destruction of these valuable predators can be a costly mistake to farmers.

Being carnivorous, mustelids sometimes become pests, especially around poultry—particularly when the poultry farmer is lax in his sanitation. Weasels, marbled polecats, and striped skunks are the common offenders. Considerable exaggeration of the prowess and deliberate maliciousness of the wolverine exists in literature. The behaviour of this animal has led many trappers to regard it as a

true "demon of the north." A wolverine travels a trapline because it learns that a ready source of food is being held captive. In many cases the wolverine will follow the trapline back to the trapper's cabin and ransack the area in its search for food. The largest of the mustelids (around 11.5 kilograms, or 25 pounds), this animal can do extensive damage.

Those mustelids that are proficient diggers (*e.g.,* badgers) sometimes become a hazard to horses (and riders), which stumble when stepping into a burrow. Other mustelids, such as the North American skunk, tend to be one of the more common wild mammals that serve as a reservoir for rabies. Meat of the Malayan stink badger (*Mydaus javanensis*), and sometimes the skunk (*Mephitis mephitis*), is eaten in some areas of the world. The obnoxious scent glands of mustelids discourage man from eating most of these animals, however. On the other hand, the scent from these glands was used in the past as a base for perfumes. Grisons and the ferret (*Mustela putorius*) have been trained by man to enter the burrows of other animals and drive them out; the grison to flush chinchillas, and the ferret, rabbits and rats. In China Oriental clawed otter and river otter are trained to catch fish or to chase them under nets which are then dropped, securing the catch. In some areas the bones or various parts of the tayra, African striped weasel, Malayan stink badger, hog-nosed skunks, and the African small-clawed otter are believed to possess medicinal value.

Form and function. Being highly carnivorous, mustelids have well-developed carnassials and a reduced number of premolar and molar teeth, the total number of teeth being between 30 and 38 (average, 34). With such a reduction in tooth number, the facial area of these animals is shortened. In the wolverine and American badger the bone of the lower jaw has a projection, the postglenoid process, that curves over the jaw articulation (glenoid fossa), resulting in a condition in which the mandible can be locked in place.

Mustelids are either digitigrade or plantigrade, and they have five toes on each foot. Most mustelids have short legs and many tend to have an elongate, slender body. They are quiet, agile, and graceful in their movements. There is a great variation in the tails of mustelids: flattened and elongate in the otter; stubby in the stink badgers; short in some of the weasels and badgers, long and thin in other weasels; long and bushy in marten; and especially elongate, with long hairs, in the skunks. The ears are small and rounded. Colours are usually brown or black. Some weasels change colour, becoming white in the winter, thus protecting the animal or, perhaps more probably, hiding it from its prey and making capture easier. Such white weasels are called ermine in the fur trade. Most mustelids have well-developed anal scent glands. The skunks, zorille, and the marbled polecat (*Vormela peregusna*), and to some extent the stink badger and ratel, can forcibly eject the vile-smelling material as a spray or fluid. All mustelids that possess powerful, controllable scent glands (except the stink badger) are marked in obvious contrasting colours, usually black and white. Such a striking colour combination may serve as a warning to potential predators, minimizing the likelihood that the animal will be attacked. After a single encounter with an adult mustelid, a predator avoids all animals so marked, including the less formidable young of the mustelid.

Repellent scent of mustelids

The smallest member of the Carnivora, the least weasel, belongs to the family Mustelidae and weighs 30 to 70 grams (one to 2.5 ounces). The largest mustelid is the sea otter, which may weigh 41 kilograms (90 pounds). In most mustelids the males are larger than the females.

Civets and allies (family Viverridae). *Natural history.* Viverrids are more numerous than the mustelids in terms of living genera (about 36) and species (about 75).

They are largely omnivorous in food habits, eating a variety of small mammals, birds, reptiles, eggs of birds and reptiles, amphibians, fishes, crustaceans, insects and their grubs, and earthworms, as well as vegetation such as fruits, nuts, bulbs or roots, and other plant material. A few, such as the palm civets (*Nandinia, Arctogalidia, Paradoxurus, Macrogalidia*), tend to be mostly vegetarian but at times take small mammals, birds, and insects. The binturong (*Arctictis binturong*), the largest member of the family, feeds mainly on fruit, although carrion is sometimes eaten. Although the suricates or slender-tailed meerkats (*Suricata suricatta*) tend toward an omnivorous diet, bulbous roots constitute most of their food. Approximately half of the genera of viverrids are composed of species that tend to be more carnivorous than herbivorous; the remaining genera are truly omnivorous. Some species tend to feed on one type of food: the ruddy mongoose (*Herpestes smithi*) on large snails, the Indian mongoose (*H. edwardsi*) on carrion remaining from the feasts of the large predators, the water mongoose (*Atilax paludinosus*) on crocodile eggs, the striped-necked mongoose (*H. vitticollis*) and the Congo water civet (*Osbornictis piscivora*) on fish. The banded palm civet (*Hemigalus derbyanus*) and Owston's civet are sometimes known to eat significant amounts of earthworms.

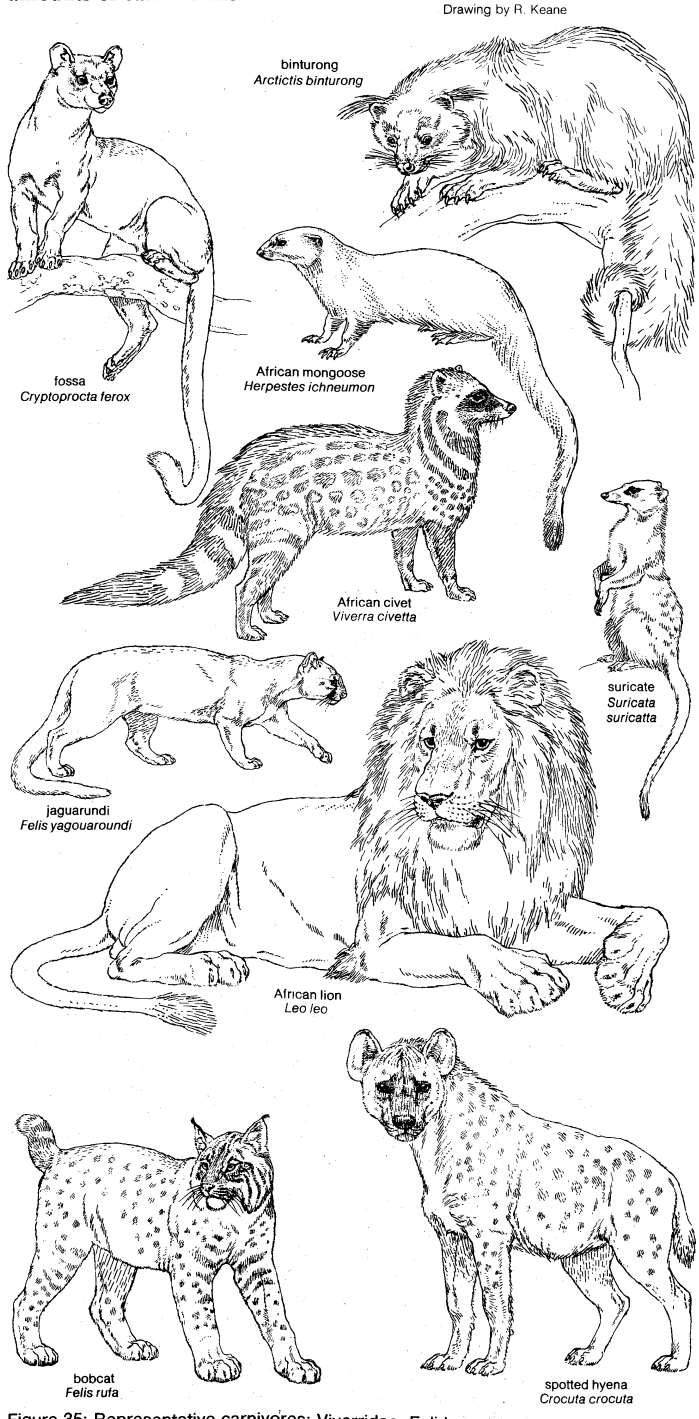


Figure 35: Representative carnivores: Viverridae, Felidae, Hyaenidae.

A number of viverrids (*Herpestes, Atilax, Mungos, Helogale*) are noted for their peculiar habit of opening eggs, as well as other food items with hard shells (crabs, mollusks, nuts). The animal stands on its hindlegs and hits the egg against the ground. Sometimes it carries the egg to a rock and, standing with its back to the rock, throws the egg between its legs and against the rock until the shell is broken. (Early reports of this behaviour met with skepticism but have been verified by other observers.) The Madagascar narrow-striped mongoose (*Mungotictis lineatus*) exhibits the same behaviour but lies on its side and uses all four feet to toss the egg.

Viverrids as snake-killers

Viverrids are perhaps best known for their ability to kill some kinds of poisonous snakes, especially cobras, which are slow in striking and apparently unable to strike when at close quarters. The mongoose can dart within the outer striking range of the snake and grab and break the lower jaw. The mongoose is much less successful against a viper such as a rattle-snake, which has a very rapid strike and return to the pre-strike defensive position. The white-tailed mongoose (*Ichneumia albicauda*) has the unusual habit of feeding on the tree hyrax, which it follows and kills in the tops of trees.

Viverrids are highly variable in hunting and social behaviour. Some species are solitary; others hunt in groups, ranging from a single pair to a family group or a wandering band of more than 20 individuals. Two species, the meerkats (*Suricata suricatta* and *Cynictis penicillata*), live in colonies. Colonial organization is unusual for animals that are largely carnivorous but is found also in the European badger. An interesting relationship exists between the type of association and the time of day that viverrids are active. The viverrids that travel singly or hunt in pairs comprise the majority of species, and all are nocturnal. All those viverrids that hunt in bands or live in colonies are diurnal. The only exceptions are the mongooses of the genus *Herpestes,* which may hunt singly or in bands and may be either nocturnal or diurnal depending on the species involved.

The senses for seeing, hearing, and smelling are well developed in viverrids. Predatory birds apparently are recognized almost instantly and at considerable distances. Viverrids, like mustelids, tend to be silent, but such sounds as purring, grunting, chirping, low-pitched coughing, growling, hissing, whining, twittering, and barking have been described. More elaborate or complex "language" usually is found in gregarious forms. Specific warning or danger cries, signifying the approach of either a ground or aerial predator, are known in the social meerkats. The noisiest viverrid, however, is the nocturnal, arboreal binturong, which growls, hisses, and, at times, howls loudly. Also extremely noisy are the cusimanses (*Crossarchus* species), which travel in large groups grunting, chattering, and twittering.

Viverrid habitats

Most viverrids are terrestrial, but at least 10 genera have forms that inhabit trees to a considerable extent. Mostly arboreal are the African palm civet (*Nandinia binotata*), the small-toothed palm civet (*Arctogalidia trivirgata*), the musang (*Paradoxurus* species), the masked palm civet (*Paguma larvata*), the binturong, the fossa (*Cryptoprocta ferox*), and the slender mongoose (*Herpestes sanguineus*). The linsangs (*Prionodon* species), although spending some time on the ground, are good climbers and live in the hollows of trees. Palm civets (subfamily Paradoxurinae), a few other civets, and genets also spend considerable time in trees. Mongooses, however, seldom climb and are unable to move about easily in trees. This inability is evident when they descend from trees. An efficient well-adapted climber comes down a tree head first. Of the mongooses, only the slender mongoose, and perhaps the Indian gray mongoose, have been observed descending head first. All other mongooses back down. Although most viverrids can swim if need be, three genera have semi-aquatic members: the Congo water civet, the African civet (*Viverra civetta*), and at least three species of mongooses of the genus *Herpestes (H. brachyurus, H. urva,* and *H. vitticollis).* The otter civet (*Cynogale bennetti*) has slightly webbed feet and is aquatic. An even more aquatic viverrid, or at least more agile one, is the water mongoose, which dives and swims like an otter, although, unlike the otter, its feet are not webbed.

Some mongooses exhibit curious antics to attract their prey. For example, the white-tailed mongoose may dance outside of chicken pens and bite off the heads of chickens that curiously stick their heads through the netting. The banded mongoose is said to stand upright and fall over on its side, using such antics to attract curious guinea fowl so that they approach close enough for capture. An interesting habit found in the meerkat plays a role in heat regulation. The belly of this animal has little hair, much less than the back. In the winter or during cold weather, the meerkat basks in the sun, standing on its hind feet with the ventral surface directed toward the heat source. When the temperatures are high, the animal seeks out cool surfaces, on which it lies on its stomach. It may even dig out an area to expose the cool subsurface layers of earth before lying down.

The fossa walks on the soles of its feet and thus is plantigrade or semiplantigrade, but most other viverrids walk on their toes (digitigrade).

Information concerning reproduction in viverrids is sketchy. In most species studied there are two litters a year (usually one in spring and one in fall), although breeding can occur throughout the year. The gestation period is usually between 49 and 64 days, but in various mongooses of the genus *Herpestes* it ranges from 32 to 49 days. Most litters consist of three or four young, but the black-legged mongooses (*Bdeogale* species) may have only a single young. Most female viverrids have two pairs of abdominal nipples, thus perhaps restricting the litter size to four, but some females of the genus *Herpestes* have three pairs. The young are born in burrows, in a nest of grass on the ground, or in hollow trees.

Viverrids are found in the Old World tropics and subtropics, throughout Africa, southern Asia, Madagascar, and southeastern Europe. Only the genet (*Genetta genetta*) occurs naturally in Europe. Although restricted to southern France and Spain today, it was at one time found as far north and east as Belgium and Germany. Various mongooses (*Herpestes* species) have been introduced into the West Indies, Hawaii, and New Zealand, as well as Italy and Yugoslavia. Viverrids are most abundant in Southeast Asia (India and the Malay Peninsula). In habits and habitat they fill the niches of the mustelids and procyonids found more abundantly in the northern regions or in the New World. Viverrids are the only members of the Carnivora to reach Madagascar, where they have flourished and undergone adaptive radiation. About a sixth of the known genera of viverrids are restricted in distribution to this large island.

Importance to man. Probably of most significance economically is the musk, called civet, produced by the anal gland in members of the viverrid group known as "civet cats." Civet is a yellowish substance with the consistency of butter and is of great importance as a base for perfumes. It is produced by the anal scent glands and stored in a sac near the genital region. The Oriental and African civets (*Viverra* species) and the lesser Oriental civet, or rasse (*Viverricula indica*), are kept in captivity so that this musky secretion can be collected several times weekly. The secretion is obtained by scraping the sac with a wooden spatula. As much as four grams of civet can be obtained in one week. Long, narrow cages are used to prevent the animal from turning and biting while the secretion is being collected.

Viverrids as the source of musk

This musk has been in great demand for many years and has commanded high prices. Civet is used as a base for perfume extracts, and in India the product of the lesser Oriental civet has been used to flavour tobacco. In some areas, parts from viverrids (the binturong and the linsang, *Poiana richardsoni*) are believed to have various medicinal values.

The only civets of value to the fur trade are certain members of the subfamily Viverrinae. Of these, the large Indian, or "Chinese," civet (*Viverra zibetha*) is the only one to be utilized for fur to any great extent. Fur of the small Indian civet and the genet is used for trimming clothing. Some members of tribes in Madagascar use the tails of

civets for ornamentation. The meat of both the large and small Indian civets has been described as delicious.

The genet, white-tailed mongoose, and fossa often become pests in areas where poultry is raised. Most viverrids are efficient rat catchers, and some, such as the binturong, suricate, cusimanse, and certain mongooses, are domesticated because of their affectionate nature and their ability to eliminate mice, rats, and cockroaches from the residence of the owner. In some areas, Greece and southern Italy for example, genets have been kept as house pets, and this animal may have been the "cat" of the ancient Greeks. In other areas, the binturong is tamed to follow its master like a dog.

Form and function. Viverrids have from 36 to 42 teeth. The canines are slender and elongate and the carnassials well developed.

Viverrids have short legs with five toes on each foot, except for the meerkats; *Suricata* has four toes on each foot and *Cynictis* four on the hindfoot. The pollex ("thumb") and hallux ("big toe") are functionless and do not touch the ground. Each toe is capped with semiretractile claws that are more retractile in digitigrade species than in those that are semiplantigrade. Most viverrids have a long, slim body and a long, bushy tail. Except for some marsupials, the binturong is the only Old World mammal with a prehensile tail. The elongate head ends in a pointed, foxlike muzzle. Coloration in the civets can be quite variable in pattern, from uniform to spotted, blotched, or striped, and ranges from black and gray through various shades of brown to rufous and yellow. The stripes may run longitudinally (*Mungotictis, Galidictis*) or vertically (*Mungos, Chrotogale, Hemigalus*). Some forms of civets (especially *Bdeogale, Ichneumia*) have black legs.

Viverrids range in size from the dwarf mongoose (*Helogale parvula*) with a weight of about 500 grams (about one pound) and a total length of 300 millimetres (12 inches) to the binturong with a weight of up to 14,000 grams (about 30 pounds) and a length of 1,800 millimetres (72 inches).

Hyenas (family Hyaenidae). *Natural history.* Hyenas form a relatively homogeneous family that includes two basic types: the true hyenas (two species of *Hyaena* and one of *Crocuta*) and the aardwolf (*Proteles cristatus*). Hyenas are noted for their scavenger feeding and bone-crushing habits, although they do hunt live animals. Their food consists largely of the remains of artiodactyls, such as the antelope, left by the large cats. Vultures and jackals often feed on such kills after the cats and before the hyenas move in. The spotted hyena (*Crocuta crocuta*) eats large amounts of bone and skin, seeming to prefer them to meat. These, of course, are the remains after the predator has had its fill. Spotted hyenas hunt in aggressive packs and at times attack and kill larger animals, even as large as a young rhino. If the animal attacked puts up a fight, the hyenas stop the attack and move away. In addition to carrion, brown hyenas (*Hyaena brunnea*) feed on small mammals, young or newborn larger mammals, reptiles, eggs, insects, fruits, and berries. In coastal areas they may feed on dead crabs, whales, and other sea life washed up on shore. The brown hyena is less apt to attack living animals than is the spotted hyena. The striped hyena (*Hyaena hyaena*) has food habits similar to those of the brown hyena. The aardwolf feeds almost exclusively on termites but eats a few other insects, such as beetles. A large quantity of soil and grass is usually eaten accidentally while feeding at termite mounds.

Feeding habits of hyenas

Since carrion produces a distinctive odour, hyenas depend on the sense of smell to a great degree. Most of their food probably is located by this sense alone, sight and hearing being less important. Sight is poorly developed in hyenas.

The well-known howl of the "laughing hyena" is the call of the spotted hyena. The other species are not as noisy.

In areas undisturbed by man, hyenas hunt during the day. Members of this family inhabit open areas of plains and brushland but at times may be found in open forested areas. Hyenas form packs and are tireless trotters and runners, at times wandering great distances.

Spotted hyenas breed during the dry months, from April to July; the aardwolf breeds from October to November. The gestation period is approximately three months (95 to 110 days in the spotted hyena). Hyenas usually have three or four young. Dens may be located in abandoned burrows, caves, dense brushy areas, or among rocks. Some hyenas use the burrows of the aardvark for nesting spots, as does the African hunting dog.

The Hyaenidae have a distribution similar to, but more restricted than, that of the viverrids. They are found throughout Africa and from southern Asia east to eastern India.

Importance to man. The scavenger habits of most of the members of the Hyaenidae make them of considerable benefit to the ecosystem as a whole and often directly to man. They perform the service of eliminating the remnants left by the large predators and thus fit into a niche unused by most mammals. In some parts of Africa, hyenas are allowed to roam the villages, cleaning up garbage. In some areas people place their dead out in the open to be removed by these inexpensive "natural undertakers."

Hyenas at times will attack and kill livestock and sometimes people, usually sleeping persons or young children. Being cowardly, they seldom fight if the individual puts up any type of defense. When hungry, they may raid camps and carry off various items made of leather. In attacking large livestock or antelope, they usually approach from the rear and kill their prey by tearing open the abdomen.

Form and function. Hyenas have a total of 34 teeth, fewer than the generalized number found in other carnivores. The aardwolf has fewer upper and lower premolars and molars, resulting in from 28 to 32 teeth. The incisors are unspecialized in all members of the Hyaenidae, with the third incisor being the largest. The canines are well-developed, sharp, and elongate. Premolars of hyenas have well-developed crowns, useful in bone crushing, and tend to be conical. In the aardwolf the premolars are small, widely spaced, and, along with the small molars, tend to be vestigial. The lack of development of the cheek teeth is related to its insect-food diet, which does not require a powerful dentition. The molars of the hyenas are large and powerful, aiding in bone crushing, as do the jaws, the most powerful found in any mammal of comparable size. The head tends to be massive, with a broad muzzle, not long-nosed as in the canids or short-faced as in the cats. Anal glands are present in most forms and are well developed in the aardwolf, which can eject an obnoxious fluid when threatened.

All hyenas have a characteristic stance. The front legs are longer than the hindlegs, causing the back to slope conspicuously from a high point at the shoulders to a low point at the hindquarters. Hyenas walk on their toes and run well. There are four toes on each foot, capped by blunt, nonretractile claws. *Proteles* has five toes on the front foot. Most members of the family have a well-developed mane on the neck, which may extend down the midline of the back. This mane often is more noticeable in the young hyenas, although well developed in all age classes of the aardwolf. All have large, rounded ears. These animals are grayish or brownish in colour, marked with spots (*Crocuta*), stripes (*Proteles, Hyaena hyaena*), or uniform with bars on the legs (*Hyaena brunnea*). The tail is medium in size, bushy at the base, and pointed at the end.

Cats (family Felidae). *Natural history.* Almost all cats are strictly carnivorous and feed on small mammals and birds or on the larger herbivorous artiodactyls like deer and various types of antelope. The fishing cat (*Felis viverrina*) feeds largely on fishes and clams or snails and thus fits into a slightly different niche from that of most cats. The flatheaded cat (*Felis planiceps*) is the only felid known to feed to any extent on vegetation, with fruit and items like sweet potatoes being preferred when available. The large cats sometimes drag the kill into a tree or place it under a bush after the initial gorging. Cats live on a feast-or-famine routine, gorging themselves when a kill is made and then fasting several days between kills.

Cats have good senses of sight and hearing. Long, sensitive whiskers on the face aid the cat during the stalking of the prey by brushing against any obstacle in a path and enabling the cat to avoid making excessive noise as it stalks its prey. The easiest, most open path can be fol-

lowed through vegetation, even at night. Smell apparently is not as well developed as in some other carnivores, such as the dogs. Cats are nocturnal in habits. The cheetah is an exception to this rule, hunting predominantly by day. Their large eyes are especially adapted for seeing at night, when the amount of light is low. The cat stalks its prey to a point as close as allowed by available cover, then closes the final distance by a leap or a short dash. A chase may ensue, with the predator relying on superior speed and reflexes to overtake the dodging prey, which often has greater endurance. Although the endurance of most felids is sufficient for a chase of only a few hundred metres at most, the cheetah may pursue its prey for much greater distances, up to about 5,500 metres (3.4 miles). If overtaken, the prey is thrown down and dispatched with a deep bite, usually in the neck. Almost all hunting by these animals is done alone and quietly. The major exception to the solitary habit is the lion (*Leo leo*), in which the group, or pride, may number as many as 30 individuals. Cheetahs sometimes are found travelling in small groups but stalk and chase their prey individually.

Hunting method of cats

The gestation period of most smaller cats is approximately two months (50 to 68 days), and that of the large cats is three or three and a half months (88 to 113 days). Two or three kittens make up the usual litter. The jaguar (*Leo onca*) tends to have only one kitten. The domestic cat (*Felis catus*) sometimes has more than six kittens. Female cats may have from four to eight nipples. The breeding season usually is in the late winter or early spring. In smaller species those females that have early litters may produce a second litter in late summer or early fall. Some cats are capable of breeding at any time during the year (lion, tiger, and leopard). The size of the animal does not seem to determine the litter size, number of litters, or the time of the breeding season. In the larger cats, however, the age of initial breeding is greater; the females may be three or four years of age and males as much as five or six years old. Smaller cats may breed when less than a year old. Most litters are born in places seldom disturbed by man, such as in a rocky cavern, under a fallen tree, or in a dense thicket. The serval (*Felis serval*) uses an old porcupine or aardvark burrow. In most species the male does not aid in the care and raising of the young, and in fact, the female may have to guard against his attacks on the kittens. The male jaguar and ocelot (*Felis pardalis*) do help in raising the young.

Vocalizations of cats

Cats are noted for their purring when apparently content and for their snarling, howling, or spitting when in conflict with another member of their kind. The larger forms, especially the lion, often roar, growl, or shriek. Usually, however, cats are silent. Many cats have "clawing trees," upon which they leave the marks of their claws as they stand and drag their front feet downward with the claws extended. House cats are well known for this habit, and an owner may supply an artificial tree to save damage to furniture. Whether such behaviour is for the purpose of cleaning or sharpening the claws is debatable, but the behaviour is innate; kittens raised in isolation soon begin to claw objects. Other characteristics of cats are the constant movement of the tail when stalking, the habit of washing the face, and the habit of digging a hole in which to bury the fecal material and urine.

Almost every area on earth has some type of native cat. Australia is the only continental area that has no native felids, though the family also is absent from Madagascar and a number of oceanic islands. Cats are found in most habitats, from the forest to the desert, although they are typically animals of wooded environments. Many are in danger of extinction because of their incompatability with the activities of man or because of their combined value to zoos and the fur trade.

Importance to man. The cat, like the dog, was originally taken into the home of man as a pet. Unlike the dog, the cat has not been trained to hunt or to serve as a guardian. Cats may perform the function of "mouse catcher" in those areas where mice become a problem, but most urban house cats do not have the opportunity to perform this unless garbage and trash are allowed to accumulate. On farms, where storage of grain is common and rodents become numerous, cats help to hold their numbers in check.

The fur of cats is sometimes in great demand, especially when high fashion calls for "fun furs," those with contrasting colours involving spots or stripes of the type found on the pelts of many cats. The demand is such that numerous kinds of cats are in danger of becoming extinct. In some regions of the world, catskins have been in great demand for many years. In parts of Africa the royal furs are usually those of a cat, as those in Eurasia are from the ermine or the sea otter. Chiefs and other high dignitaries of some native tribes use the pelt of the serval as a sleeveless cloak thrown over the shoulders.

Man-eaters

The larger cats are strong, fierce, and extremely dangerous when hungry. Should one of these animals learn that the flesh of man is edible, it must be eliminated. In many, if not most, cases the "man-eater" is an old cat no longer able to kill native wild prey. Although tigers and leopards are most famous for man-eating activities, lions and jaguars also may become dangerous. The American lion or puma (*Felis concolor*) tends to avoid contact with man and most records of its attacking man, except when cornered, are questionable. Similarly, if a large cat learns that a ready meal is available in the form of fenced or domesticated livestock, the rancher will have problems until the predator is eliminated. As in the case of the man-eater, the problem is caused by one animal and not all the individuals should be condemned.

Cats are intelligent and can be trained. Many kinds of cats can be tamed as pets, but some, especially the larger species, really can never be trusted when they become older or when they are sexually active. Most wild animals, to be trustworthy, must be fed by bottle before their eyes open, in which case the cat apparently will regard the human as its actual mother. The domestic cat, when liberated in the wild or allowed to roam in the field, may become a more serious and damaging predator on wildlife than the native predators in the region.

Form and function. Cats have a reduced number of premolar and molar teeth; the typical dental formula includes only 30 teeth. The incisors are small and chisel-like and the canines long and pointed. The premolars are sharp, and occasionally an upper premolar may be lacking. The lower molar is elongate and sharp, the upper molar rudimentary. Because of the reduction in the number and size of the cheek teeth, a space remains between the canines and premolars in all cats except the cheetah. Felids form the most carnivorous group in the order, and the highly developed carnassial teeth reflect this specialized food habit. There is little if any specialization in the teeth for grinding or chewing. The strong masseter (chewing) muscles of the jaw restrict the amount of lateral movement and primarily permit vertical movements of the jaw, for holding the prey in a viselike grip and for slicing off pieces of meat with the carnassials. Thus meat is cut off and swallowed in relatively unchewed chunks that are broken down by strong enzymes and acids in the digestive tract. Cats use the rasplike tongue to remove the flesh from the bone.

Because of the reduction in the number of teeth, cats have a short face and a rounded, compact head. The ears tend to be short or, in some forms, tufted with hair. Most forms, except the lynxes (*Felis lynx, Felis rufa*), have an elongate tail that comprises about a third of the total length. The agility of a cat is evident in its anatomy. The clavicle, or collarbone, is much reduced in size. It does not connect with other bones but is buried in the muscles of the shoulder region. This type of construction allows the animal to spring on its prey without danger of a supporting or connecting structure, like the clavicle, being broken. The hindlegs are well developed, with powerful muscles that propel the animal in its spring toward or onto the prey animal. In addition to the power of the hindlegs, the animal uses strong back muscles to straighten the spinal column and provide extra force in springing and running. Most of the skin covering the body of the cat is loose, allowing for even more freedom of movement as well as providing some protection during a fight.

Anatomical basis of feline agility

The legs of the jaguarundi (*Felis yagouaroundi*) are rather

short, and those of the cheetah elongate. All cats, except the cheetah, have retractile claws. Such claws can be drawn back into a sheath when the animal is walking, rendering the footsteps noiseless and keeping the claws sharp; or they can be extended as an aid in pulling down and holding the prey when the animal is attacking or in protecting the cat when attacked. The sharp, strongly curved claws are also utilized when the cat climbs trees. There are five toes on the front foot and four on the rear. The first toe and its pad on the front foot are raised so that only four toes register in a track.

Cats are usually some tone of brown in colour. Some, like the American mountain lion and jaguarundi, are uniform in coloration, but most are marked with spots, stripes, or rosettes. The young of the mountain lion are spotted, which may indicate a spotted ancestor. Characteristic of all cats is a dark stripe extending laterally from the external corner of the eye. This stripe tends to exaggerate the expression on the animal's face toward one of meanness but probably serves to conceal the face by breaking the roundness of the eyes. Many of the colour patterns are repeated again and again in the different species of cats, from the largest to the smallest. Several species have black or nearly black colour phases mixed into normally coloured populations. The only cat with a well-developed mane or long hair on the back and chest, or both, is the male African lion, which is also the only terrestrial member of the order Carnivora that shows obvious sexual dimorphism. In many felids the male is larger than the female, but this sexual difference is not easy to distinguish, because the difference in size is slight.

Walrus and seals (families Otariidae, Odobenidae, and Phocidae). *Natural history.* Members of the families Otariidae (eared seals), Odobenidae (walrus), and Phocidae (earless seals), commonly called pinnipeds, are strictly carnivorous and mostly marine. A few venture up freshwater rivers, and two forms (*Pusa* species) are landlocked on inland lakes. In actuality, pinnipeds are amphibious, being aquatic as to food habits but terrestrial for mating, bearing young, and resting. The diet of pinnipeds consists mainly of fishes, cuttlefishes, octopuses, and crustaceans. Most of the fishes fed upon by seals are those which school. A few pinnipeds are specialized for feeding on krill (macroscopic pelagic plankton, primarily small shrimplike euphausians). Seals known to feed largely on krill are the ringed seal (*Pusa hispida*), harp seal (*Pagophilus groenlandicus*), and the crabeater seal (*Lobodon carcinophagus*); all take fishes, as well. The crabeater seal is misnamed, as it is not known to feed on crabs. Some seals, such as the leopard seal (*Hydrurga leptonyx*), are quite predatory, feeding on penguins, other birds that land on water, and other seals. The Australian sea lion (*Neophoca cinerea*) feeds largely on penguins. Such predators take advantage of the surface light at night and probably feed by visual discrimination from below. The bearded seal (*Erignathus barbatus*) and walrus (*Odobenus rosmarus*) are bottom feeders, taking largely sessile organisms such as mollusks. The walrus occasionally feeds on small whales, narwhals, and seals.

Pinniped feeding

Generally, the senses of hearing and smell are poorly developed in the pinnipeds, although smell is fair in the phocids. Both the ears and the nostrils are closed when the animal is underwater. Sight is good and is adapted for use underwater and at night. All pinnipeds have well-developed facial whiskers and a well-developed system of nerves in the facial area. Seals probably have an excellent sense of touch in the whiskers, which aids in locating their prey or warns the animal of a possible collision with a submerged object. Young walrus have such well-developed whiskers that these appear as a brushlike mustache. The walrus and bearded seal utilize these whiskers to filtre food organisms found on the bottom and as a tactile organ. Phocid seals have patches of elongate hairs above the inside corners of the eyes. There is no evidence that seals use echolocation (reflected pulses of sound) in their movements or food-seeking activities, as has sometimes been postulated.

Pinnipeds may be solitary at certain times of the year but are usually gregarious, much more so than terrestrial carnivores. During the breeding season more than a million individuals may congregate on an island. Each male sets up a territory and gathers as many females as he can. The territory is maintained by frequent loud vocalization, a threat stance with neck outstretched, rushing at an approaching rival, and possibly through the use of olfactory odours thought to be produced in the mouth. Although the term harem is usually used to describe the large number of females in the territory of a male, this concept is misleading, since the females can move from one group to another. The groups are a result of the females staying clear of the males as the latter patrol their territories. Once the territory is established the male remains to protect it, going as long as two months without eating (sea lions and fur seals). Eared seals (the otarids) and walrus tend to gather harems, while the earless seals (the phocids) are mostly monogamous (one male, one female) in their breeding activities. The gathering of harems forces the mating activity to be terrestrial rather than aquatic. In many, if not most, of the northern species of seals the breeding season is restricted in many ways by the environment. In most areas only four months exist with a climate favourable for the production of young.

Territoriality in pinnipeds

Pinnipeds have a fairly long gestation period, ranging from eight months in the leopard seal to 12 months in fur seals and sea lions. Breeding often takes place soon after the birth of the young (five days in the fur seal) or, as in the leopard seal, four months after the young are born. Copulation is usually on land, but some phocids mate in the water. Most seals, except the walrus and perhaps the crabeater seals, exhibit delayed implantation in embryonic development. The young are born on land or on

Drawing by R. Keane: Hawaiian monk seal based on San Diego Zoological Garden photograph

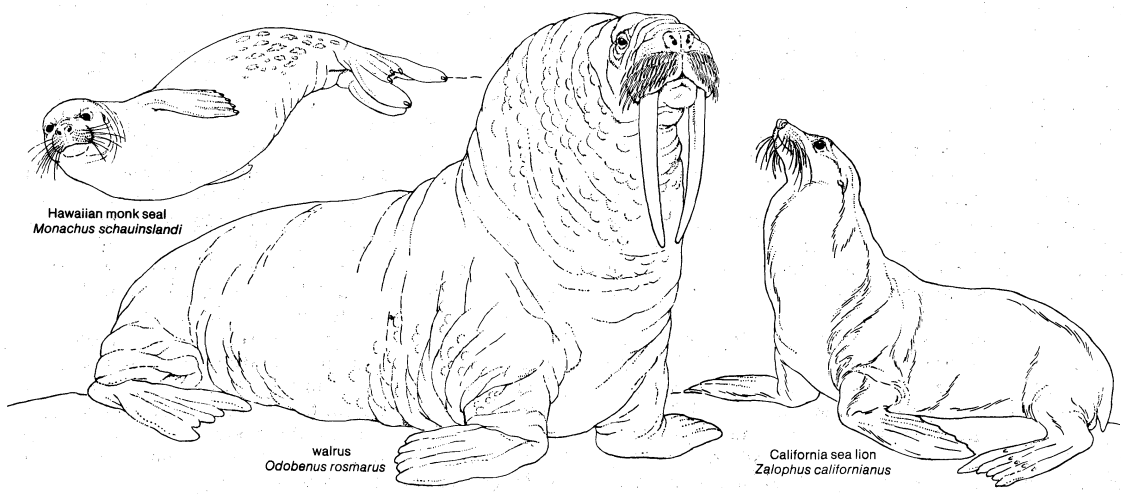


Figure 36: Representative carnivores: Pinnipeds.

ice. Usually there is but one young born, although most females have two pairs of mammary glands (some phocids have one pair). The mother goes out to sea to feed soon after the young is born. Since she must search to find her young when she returns, there has probably been a selection against the production of more than one young, at least in highly gregarious forms, for the time spent searching for the second pup would be at the expense of the first. It is evolutionarily better to produce one vigorous offspring than two weak ones. Nursing usually takes place on land. At birth the newborn pup can travel on land and swim, although young otarids are several weeks old before they develop enough blubber to keep them floating and insulated against the cold water. The young of the earless seals nurse for only three to six weeks, and the females remain near the young throughout this period. Pinniped milk is low in water content and high in fat. Such high fat content results in the rapid growth of the young, but perhaps of more importance is the low water content, of benefit to the mother in a habitat where water conservation is so important.

Female pinnipeds are sexually mature when two to eight years of age, usually at the age of three or four years. Males, at least in the polygamous species (those which gather a "harem"), may not be allowed to mate for several years after reaching sexual maturity. In these species the younger males are driven by the old males to a bachelor group located outside of the breeding group and restricted to a given area. After the breeding season, pinnipeds are largely pelagic (open-sea dwellers), travelling long distances either singly or in small groups.

A number of animals prey on seals: large sharks, killer whales, polar bears, and other seals, such as the leopard seal and walrus. Seals, especially the young, spend a good deal of time playing. They resemble the river otter in this regard.

Seal barking

Pinnipeds produce a number of different types of sounds. Barking by male seals may be related to social dominance and territorial defense, as the polygamous species are extremely vocal and the monogamous forms extremely quiet. The "barking" of the circus seal (the California sea lion, *Zalophus californianus*) is familiar to most people. The study of underwater communication in seals is difficult and in its infancy, but a variety of sounds have been recorded. These sounds may play a role in navigation, social behaviour, and foraging. Seals also roar or bellow, honk, chirp, bleat, grunt, or cough.

These aquatic, carnivorous animals are found throughout most of the coastal regions of the world. They are especially abundant in the cold polar waters. The walrus and phocids are well distributed throughout the northern coastal and ice-front areas. The otarids do not extend as far north, southern Alaska and Kamchatka being the most northern points, and are absent from the mid-Atlantic and North Atlantic Ocean areas. Pinnipeds are absent from the region of the Indian Ocean, southern Asia and northern Australia, the coastal regions of Central America and northern South America, and most of the coastline of central Africa. A few seals are landlocked on the Caspian Sea (*Pusa caspica*) and Lake Baikal (*Pusa sibirica*), where they subsist mainly on fish.

Importance to man. The seal is of significant importance to the survival of the Eskimos and other inhabitants of the north, who use almost every part of the animal: the meat is eaten; bones used for implements; tendons for sewing; hides for leather to make footwear, boats, shelters, bags, clothing, and ornaments; and the oil and fat for fuel, lubrication, and tanning. The ivory tusks of the walrus are carved into statues and knife handles. Seals also are taken commercially for their oil, for meat that can be used for food or fertilizer and for their hides that are used as leather.

A few seals are of importance to the fur trade. Of special interest are the true fur seals, *Callorhinus ursinus* in the northern regions of the Pacific Ocean and *Arctocephalus* species in the southern regions off the coasts of South America, Africa, and Australia. Most of the species utilized by the fur trade are members of the family Otariidae, which have dense underfur. The young of members of the Odobenidae and Phocidae have an insulating coat of short, dense, wool-like hair at birth. Skins of these young animals are of some interest to the fur trade but are used mostly for novelties. Adult phocids have skins with stiff hair; the underfur is reduced to the point that the hide appears naked. However, some of the extremely striking phocids were in demand for furs in the late 1960s.

California sea lions are trained to entertain at circuses and zoos. Some seals become pests to commercial fisheries and may gather during salmon or herring runs, sometimes being killed in large numbers by fishermen.

Form and function. Basically, the difference in the tooth pattern of pinnipeds, in comparison to the dentition of terrestrial carnivores, reflects an adaptation toward grasping and tearing rather than chewing. Thus there is less variation in the form of the different teeth (commonly said to be homodont, or "similar teeth"), and there tends to be a reduction in the number of teeth. There are, generally, two or three pairs of upper incisors and one or two pairs of lowers. There is a pair of conical canines that form elongate tusks in the walrus. The premolars and molars are conical and so similar that they are difficult to differentiate and are usually referred to as postcanines. There are no carnassial teeth. Usually there are ten postcanine teeth each in the upper and lower jaws, but the exact number can vary so greatly, even within one species, that the total number of teeth may range from 26 to 38. In the walrus the tooth number is reduced even further, the usual number being two upper incisors, no lower incisors, four canines, and 12 postcanines, a total of only 18 teeth. This number is variable, with some individuals having as many as 24 teeth. Weddell's seals have canines and incisors adapted for cutting holes in the ice so that areas are available for breathing when the water freezes over. The upper incisors extend forward and contact the ice before the canines.

Adaptations for aquatic life

Pinnipeds have a number of aquatic adaptations that separate them from the terrestrial carnivores. The body is streamlined, allowing the animal to pass through the water with the least amount of friction. External ears are reduced or absent, also aiding in making the pattern of the body more streamlined. Adaptive features of the skull include an overlapping of the bones, a feature even more exaggerated in the whales, and a flattening of the head, an advantage during diving. The neck is thickened by increased musculature but is more flexible (at least in the eared seals) than in the terrestrial carnivores, allowing the animal to capture its prey in the water with greater ease. In most pinnipeds the face is short, with the cranial part of the skull longer than the facial region. The eyes are large, located forward, close together, and imbedded in a cushion of fat. The iris of the eye is contracted when out of water, dilated and almost circular when diving. This type of eye can accommodate rapidly when the animal moves from dim to bright light. The ears and nostrils can be closed while diving. The pelvic girdle is more nearly parallel to the vertebral column than in terrestrial forms. Such a change in the location of the pelvis enables the hindlimbs to be held in more of a trailing position and to be more efficient in propulsion through the water. The limbs are enclosed within the body skin to or beyond the elbows and knees. This enclosure also streamlines and results in a more efficient passage through the water. The extremities of the limbs are flattened to form paddles used to steer or propel the animal through its environment. Each paddle contains five toes, extending from bones in the palm and sole (metapodials), elongated to increase the surface area of the paddles. The first digit on both forefeet and hindfeet is more elongate than the other four. The tail is short and rudimentary, a condition approached by the sea otter but quite unlike the paddlelike tail of the freshwater otter. Mammary teats and external reproductive organs are so constructed that they can be withdrawn beneath the skin to aid in preserving the smooth outline of the body. The skin, adapted to a water environment, is tough and thick and contains a thick layer of subcutaneous fat or blubber. The blubber provides a source of energy that can be used during lactation or at times of fasting. Such fasting is characteristic of pinnipeds. This layer also provides the animal

with more buoyancy, so that sinking in the water is less apt to occur. In addition, the blubber insulates the body so that the body temperature remains nearly constant. A pinniped living in cold water may be exposed to a surface temperature of about 0° C (32° F) and yet have a body temperature of 37° C (98.6° F), requiring an efficient system of heat production and insulation. In warm waters, and especially in warm air, the animal cannot get rid of excess body heat rapidly enough to prevent overheating. The flippers, so useful for swimming, can be used on land as radiators to get rid of excess body heat. One of the most obvious behavioural features of a colony of seals on land is the waving or extension of the naked flippers.

Physiological adaptations for diving

There are a number of physiological adaptations to life in the sea. Being air breathers, seals must be able to carry as much oxygen as possible with them in diving and to conserve what they do take. This adjustment is obtained by the increased volume of blood (very large for an animal their size) and the decreased rate in the use of oxygen by the body (decreased metabolism). The heart rate decreases in the elephant seal from about 85 beats per minute when the animal is on the surface to approximately 12 per minute during a dive. This drop in heart beat rate conserves oxygen in terms of the amount being carried in the blood. There cannot, however, be a decrease in the amount of oxygen being carried to some parts of the body. The heart and brain, for example, must have a constant and sufficient supply, or the cells in these organs will be damaged. This problem is alleviated by closing down the small arteries in the tail and flippers, while the larger arteries to the head and brain remain unchanged. The muscles function without free oxygen during submergence; the energy is obtained by partly breaking down the blood sugars and without expending the limited supply of oxygen. Most dives last less than 15 minutes, and upon return to the surface there is a greater exchange of oxygen and carbon dioxide than is found in terrestrial mammals. The amount of tidal air in one respiratory cycle may be as much as 90 percent, compared to 20 percent in man. Most dives are deep vertical dives rather than extensive horizontal dives because of the problems of orientation toward the breathing hole in the ice when horizontal directions are taken. Since there is no exchange of gases with the environment while submerged, seals have the additional problem of an accumulation of carbon dioxide. They seem to be less sensitive than terrestrial animals to increased amounts of this gas. The body temperature is lowered, a result of the decreased rate of metabolism. Seals are also adapted to withstand the great pressures during their dives. Weddell's seal (*Leptonychotes weddelli*), for example, dives as deep as 335 metres (1,100 feet), more than four times the distance that man can go in a diving suit. The longest known dive by a Weddell's seal lasted 43 minutes and 20 seconds (to 350 metres [1,150 feet]), with one elderly bull reaching a depth of 600 metres (2,000 feet) but remaining down a shorter time. Exactly how these animals can accomplish such feats is unknown. The anatomy of the seal reveals part of the answer. Many structures in the body are built to yield to such pressures, not to resist them. Some of these adaptations are smaller lungs; elimination of most air from the lungs before the dive; a thorax that permits the lungs to collapse; incomplete rings of cartilage in trachea, permitting complete collapse; driving air out of the middle ear by partly filling the ear cavity with blood; and redistribution of blood in general. The phocids are more efficient divers than the otarids, and, of the phocids, those living in the Antarctic dive the deepest.

The clavicle, or collarbone, is absent, allowing more flexibility in the shoulder. The kidneys, unlike those of most mammals, are composed of numerous lobes resembling a bunch of grapes, a condition also found in the terrestrial bear and otter.

Members of the families Otariidae and Odobenidae can walk on their limbs when on land. Phocids must move by undulations of the body as the limbs cannot be turned in a position to support the body. This does not mean that the phocids are helpless on land; the crabeater seal, a phocid, is probably the fastest moving seal on land, being difficult for even a running man to catch.

EVOLUTION AND PALEONTOLOGY

Opinion is divided as to whether the order Carnivora arose from the ancient creodonts or had a separate and independent origin from the order Insectivora. Most paleontologists now favour the latter view. The Cretaceous insectivore *Procerberus* seems to be morphologically close to the most primitive carnivores (miacids), the creodonts, and the primitive hoofed mammals (ungulates). *Procerberus* was an unspecialized rat-sized predator. The order seems to be a separate and distinct line that includes the modern Fissipedia (terrestrial forms) as well as Pinnipedia (marine and aquatic forms), while also including the more ancient and extinct family Miacidae.

The extinct Miacids

Although not much is known about them, the miacids were probably arboreal forest dwellers of the tropics. Animals from this type of habitat are rarely preserved as fossils, which accounts for our incomplete knowledge of the group. These primitive carnivores may have resembled our modern weasels, with elongate bodies, short limbs, and long tails. Miacids possessed several creodont features, such as an unossified tympanic bulla (a large bubble of bone or cartilage surrounding the internal ear) and lack of fusion of the carpal (wrist) bones. Miacids differed from the creodonts in having a larger brain capacity, toe sections that were not grooved, and well-developed carnassial teeth. The carnassials, as in modern carnivores, involved the fourth upper premolar and first lower molar.

In the late Eocene and early Oligocene these primitive miacids underwent an adaptive radiation; *i.e.,* they produced diverse lines that seem to have been the beginnings of the present-day families of the Carnivora. The changes needed to convert a miacid into a modern terrestrial carnivore are minor: fusion of the separate bones of the carpus and ossification of the tympanic bulla. Carnivores that began to resemble the modern weasels, dogs, and civets were in evidence in the Oligocene, but these early carnivores were still quite similar to each other and had not differentiated to the extent of representing different families.

Two major lines within the Carnivora were distinct by the late Eocene. One, the Feloidea (or Aeluroidea), today contains cats, viverrids, and hyenas; the other, the Arctoidea (or Canoidea), contains the mustelids, dogs, bears, and raccoons. Some fossils are intermediate between the two lines, but most show definite relationships to one or the other.

Feloids had the tympanic bulla made up of an external (tympanic) and an internal (endotympanic) bone, each being a separate ossification. In many forms the claws were retractile. The transition between the miacids and viverrids (civets) was so gradual that some authorities have placed the miacids in the family Viverridae. Because of the similarities to the miacids, the viverrids are considered to form the basal stock of the feloid line. Hyenas appear to be a more recent line, having evolved from the viverrids in the early Miocene. *Ictitherium,* a mid-Miocene viverrid, had characteristics intermediate between the viverrids and hyenas, but apparently was more like the "true" hyenas. Other fossil viverrids also showed similarities to the hyenas. Although the cats seem to have evolved "suddenly," the resemblance of certain catlike viverrids (such as the Oligocene *Stenoplesictis* and the modern *Cryptoprocta* of Madagascar) to the recognized felids is regarded by some authorities as indicating the close relationship of the felids to the viverrids. In the late Eocene and early Oligocene, the felids were quite distinct, with a number of highly specialized cats already present. Of particular interest are the sabre-toothed cats such as *Hoplophoneus* and the false sabre-tooth *Dinictis,* which had large upper canines fitting into a flange on the lower jaw. The Pleistocene *Smilodon,* a more recent genus, is probably the best-known sabre-toothed cat. The genus *Felis* dates back to the early Pliocene, by which time its members had features similar to those of the modern cat.

Sabre-toothed cats

In contrast to the feloids, the arctoids almost never had retractile claws; the tympanic bulla was made up of the tympanic bone alone; and there was a canal below the bulla, through which the carotid artery passed, supplying blood to the brain. Some of the earlier miacids are difficult to separate from the mustelids, which were distinct

as early as the late Eocene. They probably were one of the first arctoid lines to differentiate. By the close of the Miocene, several general types of mustelids were evident: the weasels, badgers, skunks, and otters. Another ancient arctoid line included the dogs. Oligocene canids, represented by the genera *Cynodictis* and *Hesperocyon,* resembled elongate, short-legged weasels and civets. Both genera, although considered to be primitive dogs, were generalized enough to be ancestral to many carnivores. Genera that were more typically canid appeared in the

Drawing by R. Keane based on A.S. Romer, *Vertebrate Paleontology.*
Copyright © 1966 by the University of Chicago Press; all rights reserved

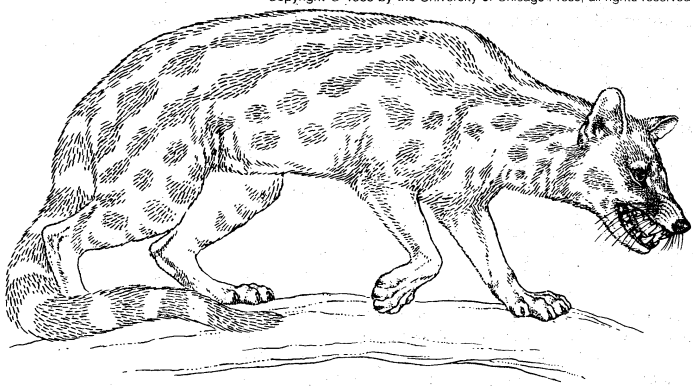


Figure 37: Early Eocene carnivore, *Hesperocyon,* an ancestor of the dogs.

early Oligocene. These were running animals, with four well-developed toes and reduced hindtoes (hallux and pollex), each ending in a blunt claw. More than 50 genera of doglike forms are known, but the main line of evolution seems to have passed through forms like *Temnocyon* (upper Oligocene) and *Cynodesmus* (lower Miocene) to the modern genera *Canis* (first seen in the lower Pliocene deposits), *Vulpes* (perhaps in the upper Miocene), and *Urocyon* (Pleistocene). Two lines of extinct canids are the hyena-like dogs and the bear dogs, neither of which has any relationship to the living hyenas or bears. Because of the close relationships of the bears, dogs, and raccoons, various taxonomists have disagreed on the placement of intermediate genera. For example, *Dinocyon, Hemicyon,* and *Cephalogale* were once considered to have been canids but are now placed with the bears and regarded as perhaps part of the ancestral lineage that diverged from the canid line in the late Oligocene or early Miocene. *Phlaocyon* and *Aletocyon* were once considered to be primitive procyonids because of their teeth, but the skull and ear

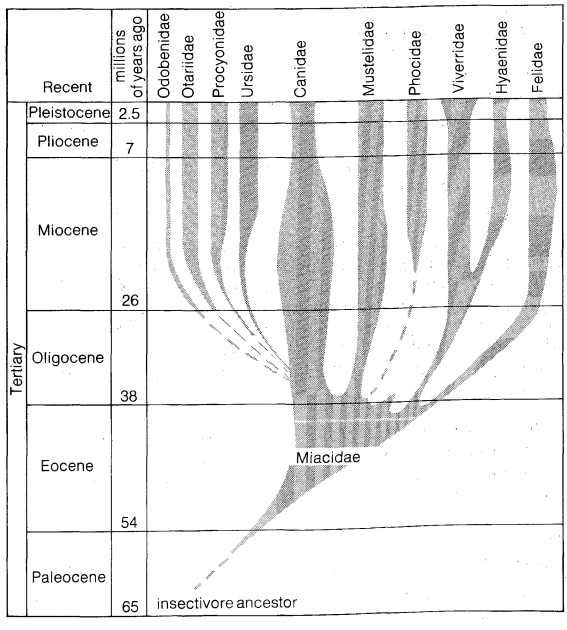


Figure 38: Phyletic dendrogram of Carnivora.

structures are more reminiscent of the dogs. A recent form, the giant panda (*Ailuropoda*), considered by some to be a bear and by others to be a procyonid, probably diverged from primitive ursid stock when the ancestor of the bear and raccoon lines still had features common to both. Procyonids (raccoons) are poorly represented in the fossil record, perhaps because of their arboreal habits. They appear to have diverged from the canid line in the late Oligocene.

The preponderance of carnivorous and carrion-eating fishes and mammals in the open sea, along with the buoyant characteristics of the bodies of the pinnipeds, results in the almost immediate destruction of any seal that dies. Dead pinnipeds seldom settle intact and become buried (the first steps toward becoming a fossil) and so are poorly represented in the fossil record. Relationships must be inferred through studies of recent forms. Although there is no fossil record of pinnipeds prior to the Miocene, it is probable that the ancestors of the pinnipeds evolved during the adaptive radiation of the miacids in the Eocene. There are two rather distinct evolutionary lines in the animals called pinnipeds. One includes the two rather closely related families Otariidae and Odobenidae. This line may have become distinct in the early Miocene or late Oligocene, evolving from a form closely related to the canid ancestor of the bears. The other line led to the modern phocids, which are quite different from the otarid–odobenid group. Their affinity to any other group of the Carnivora has not been shown, although the mustelids have more frequently been suggested as related to the phocids than any other group. *Semantor,* an aquatic carnivore from the lower Pliocene, was once considered to belong to a separate pinniped family, but is now considered a mustelid. The separation of the phocid and otarid–odobenid lines must have occurred at an extremely early date, if the two lines ever had a common pinniped ancestor. Some paleontologists postulate a separation in the early Miocene or late Oligocene. Possibly the two lines never evolved from a single aquatic group but had separate origins within the miacids, sometime in the late or even mid-Eocene. Or perhaps the otarids stemmed from an ancestral ursid type in the early Oligocene. Such an early separation, or lack of any real ancestral association, could have produced the dissimilarity of the two types of pinnipeds, as well as the difficulty in associating the phocids with any other family of the Carnivora.

Evolution of pinnipeds

CLASSIFICATION

Distinguishing taxonomic features. If one groups the seals with the terrestrial carnivores, the number of characteristics common to all members of the order Carnivora is small.

The characteristics used to separate the Carnivora from other mammalian orders and to define the subdivisions of the Carnivora are primarily structural. Of great importance are certain features of the skull (the type of jaw articulation and the shape of the paroccipital process), of the feet (the number of toes, lack of opposability of the hindtoe, type of claws, and fusion of the scaphoid and lunar bones), and of the teeth (both the overall tooth pattern and the shape of the individual tooth). The dentition is especially important in determining the relationships of fossil forms. Also useful in the taxonomy of modern carnivores are the convolutions (irregular ridges) around the lateral or sylvian fissure of the brain, the relative weights of the adrenal and thyroid glands, the type of uterus and placenta, and the position of the nipples.

Annotated classification. Although the terrestrial and marine carnivores are grouped together in the following classification, many authors separate them into two orders. In this article the split is indicated by erecting two suborders, the Fissipedia and Pinnipedia, based largely on the degree of specialization undergone by the members as they evolved and adapted to radically different environments. Basically the taxonomy presented combines the classifications by G.G. Simpson and A.S. Romer, North American paleontologists.

Numerical dental formulas are used below to indicate the number of each type of tooth in the right or left half

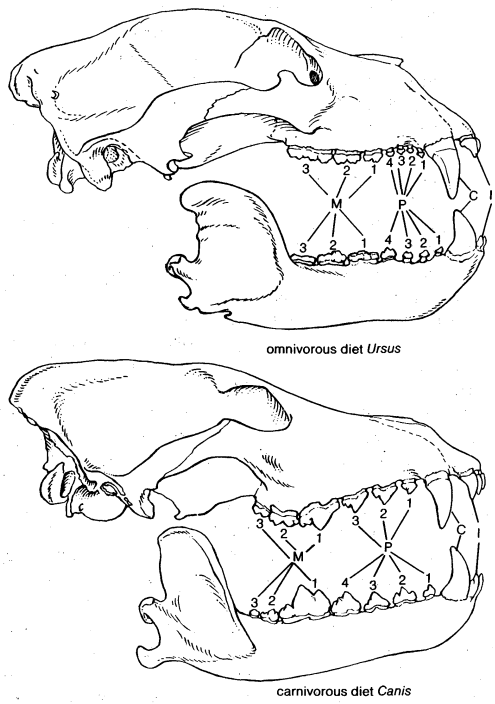


Figure 39: Tooth patterns in omnivorous and carnivorous members of Carnivora. Side view of lower mandible and cranium.

From Hall and Kelson, *The Mammals of North America*, vol. II, copyright © 1959; The Ronald Press Company, New York

of the upper and lower jaws and the total number of teeth in the mouth. Four tooth types are indicated in the following manner:

$$\text{Upper jaw: } \frac{\text{incisors . canines . premolars . molars}}{\text{Lower jaw: incisors . canines . premolars . molars}} = \text{total number of teeth}$$

ORDER CARNIVORA

Medium- to large-sized terrestrial or aquatic meat eaters distributed worldwide; composed of 10 extant families, about 115 living genera, and 274 living species.

Suborder Fissipedia

Terrestrial (except otters and polar bear); usually 3 pairs of incisors above and below, adapted for biting off flesh; premolars vary in number from the usual 4 on each half of each jaw (4/4) to 4 above and 2 below (4/2); molars vary from 2/3 to 1/2; carnassial teeth well developed and formed by the 4th upper premolar and 1st lower molar; usual dental formula: $\frac{3 \cdot 1 \cdot 4 \cdot 2}{3 \cdot 1 \cdot 4 \cdot 3} = 42$ teeth; all-inclusive dental formula: $\frac{3 \cdot 1 \cdot (2\text{–}4) \cdot (1\text{–}4)}{(2\text{–}3) \cdot 1 \cdot (1\text{–}4) \cdot (1\text{–}5)} = 28\text{–}50$ teeth; facial part of skull usually elongate; tympanic bullae large; alisphenoid canal (a perforation in the skull for a branch of the carotid artery) usually present; lacrimal duct present; pelvis at distinct angle from vertebral column; femur (thigh bone) straight, moderately slender, with small head; fibula (minor lower bone of hindlimb) much smaller than tibia (shinbone); 1st metatarsal and bones of phalanges with epiphyses only at proximal ends; metapodials never elongate; vertebrae with interlocking processes; external ears (pinnae) well developed; intestine short (6 to 8 times length of body in dog); kidneys simple, not lobulate except in otters and young bears.

Superfamily Miacoidea

All extinct generalized carnivores with carnassial teeth. Middle Paleocene to upper Eocene. There is one family, Miacidae, about 13 genera.

Superfamily Feloidea (cats, civets, and hyenas)

Tympanic bullae divided, composed of external tympanic and internal endotympanic bones; paroccipital process in close contact with bullae; most have retractile or semiretractile claws; cecum small; baculum (bone of penis) present or absent; usually polyestrous; upper Eocene to Recent; 42 living genera and 115 species.

Family Viverridae (civets and mongooses). Upper Eocene to Recent. Distributed throughout most of the southern regions of the Old World (Asia, Europe, Malay Peninsula, most of Africa, Madagascar, and associated southern oceanic islands) but most abundant in the India–Malay region; introduced on New Zealand, Hawaii, and the West Indies. Small to medium sized (0.7 to 14 kg; 1.5 to 31 lb). Well-developed carnassials; upper carnassials usually lack an anterior lobe and the lower carnassials have a well-developed talonid process; middle incisor on each side located farther inward than the outer 2; incisors broad; 4 small premolars, with 1st minute or absent; molars 2/2 or in some cases 1/2 (*Prionodon*), 2/1 (*Salanoia*), or 1/1 (*Cryptoprocta*); molars large, especially the 1st; dental formula usually $\frac{3 \cdot 1 \cdot (3\text{–}4) \cdot 2}{3 \cdot 1 \cdot (3\text{–}4) \cdot 2} = 36\text{–}40$ teeth. Alisphenoid canal usually present (absent in *Eupleres* and most members of subfamily Galidiinae); entepicondylar foramen of humerus usually present; auditory bulla externally constricted and divided by septum. Baculum well developed. Intestinal cecum small, sometimes absent. Elongate body, short legs; digitigrade or semiplantigrade gait; 5 toes usually present, with semiretractile claws (retractile in *Cryptoprocta*); pollex and hallux located on side of foot above other toes. Ears small and rounded; tail usually long (12–90 cm; 4.7–35 in.) and bushy or tapering to a point from a bushy base. Scent glands often well developed, producing secretion called "civet." Usually uniform in coloration but some spotted or striped.

Subfamily Viverrinae (true civets, genets, and linsangs). Mainly African; approximately 7 living genera and 15 species.

Subfamily Paradoxurinae (palm civets). Mainly Asian, 6 living genera and 8 species.

Subfamily Hemigalinae (banded palm civets, web-footed civets). Asia or Madagascar, 5 living genera and 7 species.

Subfamily Galidiinae (striped and ring-tailed mongooses). Restricted to Madagascar, 4 living genera and 7 species.

Subfamily Herpestinae (true mongooses and meerkats). All African, 13 living genera and approximately 37 species.

Subfamily Cryptoproctinae (fossa). Monotypic. *Cryptoprocta ferox*, found only in Madagascar.

Subfamily Stenoplesictinae. Fossil. Upper Oligocene to middle Miocene of Europe.

Family Hyaenidae (hyenas and aardwolves). Middle Miocene to Recent. Found today throughout most of Africa and southern Asia (Turkey to eastern India). Medium-sized (10 to 80 kg; 22 to 175 lb) carrion feeders; teeth specialized for crushing bones; incisors unspecialized, outer (3rd) larger than other 2; canines powerful; premolars strong and conical; molars strong and large; carnassials well developed (hyenas); long, slender canines in aardwolf (*Proteles*) with small and widely spaced premolars, small molars, and no development of carnassials. Dental formula $\frac{3 \cdot 1 \cdot 4 \cdot 1}{3 \cdot 1 \cdot 3 \cdot 1} = 34$ teeth for hyenas; premolars $3/(2\text{–}1)$, molars $1/(1\text{–}2) = 28\text{–}32$ for aardwolf. Alisphenoid canal absent; auditory bullae partially divided by a rudimentary septum in hyenas, divided in aardwolf; front legs longer than rear legs; digitigrade; 4 toes (5 on forefoot in aardwolf); claws blunt and nonretractile; entepicondylar foramen absent; baculum absent; more thoracic vertebrae (15) than any other feloid: tail of medium length (20 to 33 cm) and bushy. Preanal scent glands absent. Cecum 15 to 23 cm (6 to 9 in.) long. Short mane usually present. Ears large and rounded.

Subfamily Protelinae (aardwolf). African. 1 species. *Proteles cristatus.*

Subfamily Hyaeninae (hyenas). Africa and southwest Asia; 2 living genera, 3 species.

Subfamily Ictitheriinae. Fossil. Miocene–Pliocene. 1 Old World genus.

Family Felidae (cats). Upper Eocene to Recent. Now worldwide except for Antarctica, Australia, Madagascar, and some oceanic islands. Size medium to large (2.5 to 275 kg; 5.5 to 600 lb). Highly carnivorous; well-developed carnassials. Total number of teeth reduced from primitive carnivore condition; dental formula typically $\frac{3 \cdot 1 \cdot 3 \cdot 1}{3 \cdot 1 \cdot 2 \cdot 1} = 30$ teeth; alisphenoid canal absent; auditory bullae not constricted externally; posterior palatine foramina on maxillopalatine suture instead of on maxilla. Elongate body with relatively short legs; digitigrade; 5 toes on front, 4 on rear feet, capped with sharp, curved, retractile claws. Entepicondylar foramen present; baculum absent or vestigial; cecum small; tail short (lynx) to long (up to a third of total length). Preanal scent glands absent; tongue covered with horny papillae.

Subfamily Felinae (lynxes, typical cats, lions, tigers, leopards). Worldwide except Antarctica, Australia, Madagascar, and most oceanic islands. 2 living genera, 35 species.

Subfamily Acinonychinae (cheetah). Africa and southwest Asia. 1 species, *Acinonyx jubatus.*

Subfamilies Proailurinae, Nimravinae, Machairondontinae, Hyainailourinae. All fossil. About 39 genera.

Superfamily Canoidea (dogs, bears, raccoons, and weasels)

Tympanic bulla formed from tympanic bone only; not divided into 2 chambers; long canal beneath bulla; paroccipital process prominent and independent of bullae. Claws nonretrac-

tile. Cecum present or absent; baculum well developed; usually monestrous (1 litter of young per year). Eocene to Recent; 53 living genera and 133 species.

Family Canidae (dogs, foxes, jackals). Eocene to Recent. Now worldwide except Antarctica and most oceanic islands. Size medium (1.5 to 80 kg; 3.3 to 175 lb). Carnivorous or omnivorous; well-developed carnassials; dental formula typically $\frac{3 \cdot 1 \cdot 4 \cdot 2}{3 \cdot 1 \cdot 4 \cdot 3} = 42$; bush dog (*Speothos*) with (1–2)/2 molars (1 upper molar is sometimes minute or absent) = 38 to 40 teeth; Cape fox (*Otocyon*) with (3–4)/(4–5) molars = 46 to 50 teeth. Facial part of skull elongate; alisphenoid canal present; legs long and semirigid; digitigrade with 5 toes on front (1 dew claw) and 4 on rear foot; claws blunt and nearly straight. Entepicondylar foramen absent; baculum well developed and grooved. Cecum always present, short and simple or long and characteristically folded. Tail relatively long (up to about a third of total length) and bushy; scent glands often on dorsal part of base of tail. Ears pointed, erect, and large.

Subfamily Caninae (dogs, jackals, and foxes). Worldwide except Antarctica and most oceanic islands; dingo probably introduced on Australia in prehistoric times; wild dogs of New Guinea probably derived from domestic dogs. 11 living genera and approximately 37 species.

Subfamily Simocyoninae (dhole, hunting dog, bush dog). Asia, Africa south of the Sahara, and northern South America. 3 living monotypic genera: *Cuon alpinus, Lycaon pictus, Speothos venaticus.*

Subfamily Otocyoninae (big-eared fox). Eastern and southern Africa; monotypic, *Otocyon megalotis.*

Subfamilies Amphicynodontinae, Amphicyoninae, Borophaginae. Fossil only. Total genera about 34.

Family Ursidae (bear and giant panda). Middle Oligocene to Recent. Restricted to larger land masses of world, ice floes of Arctic, and in South America to Andes; bear of Atlas Mountains of northwest Africa now extinct. Size medium to large (27 to 780 kg; 60 to 1700 lb); usually omnivorous; carnassials poorly developed; dental formula variable, but typically $\frac{3 \cdot 1 \cdot 4 \cdot 2}{3 \cdot 1 \cdot 4 \cdot 3} = 42$ teeth. Skull elongate; alisphenoid canal present. Legs short, large, and powerful; plantigrade; 5 toes on each foot, capped by elongate, powerful, nonretractile claws. Entepicondylar foramen of humerus usually absent; lacrimal bone of skull, clavicle, and cecum absent; tail short, practically absent; ears small and rounded; lips mobile and free from gum. 6 genera and 8 living species.

Family Procyonidae (raccoons, coatis, lesser panda). Upper Eocene (probably) to Recent. Fossils through Pleistocene in Europe. Except for the lesser panda (*Ailurus*) of the southeastern Himalaya region of Asia, family is restricted to northern South America, Central America, and southern North America. Size medium (0.8 to 22 kg; 1.75 to 48.4 lb); omnivorous; carnassials poorly developed; dental formula usually $\frac{3 \cdot 1 \cdot 4 \cdot 2}{3 \cdot 1 \cdot 4 \cdot 2} = 40$, with some variation in number of premolars (kinkajou 3/3, lesser panda 3/4); alisphenoid canal absent except in *Ailurus;* legs of medium length; semiplantigrade to plantigrade; 5 flexible toes on each foot capped by nonretractile or semiretractile claws; entepicondylar foramen sometimes present; baculum well developed and, except for *Ailurus,* bilobed and hooked at distal end; cecum absent; clavicle vestigial; tail long (20 to 70 cm; 7.9 to 27.6 in.), often ringed, prehensile in *Potos;* ears small to medium in size and rounded. 7 living genera and approximately 14 species.

Family Mustelidae (weasels, badgers, skunks, otters). Eocene to Recent. Now worldwide except Madagascar, Australia, and most oceanic islands; introduced into New Zealand; more common in Northern Hemisphere. Size from smallest of carnivores (35 gm; 1.2 oz) to medium; mainly carnivorous; carnassials usually well developed; dental formula usually $\frac{3 \cdot 1 \cdot 3 \cdot 1}{3 \cdot 1 \cdot 3 \cdot 2} = 34$ teeth but may range from 30 to 38 teeth, with variation in incisors (3/2 in sea otter), molars (1/1 in ratel and African striped weasel), but especially premolars (2/3 in the African striped weasel, 4/2 in the river otter, 4/4 in American badger); alisphenoid canal absent; legs short in relation to elongate body; plantigrade to digitigrade; 5 toes on each foot, capped with nonretractile claws; entepicondylar foramen present or absent; tail usually long (up to 1/3 of total length); ears small and rounded; anal scent glands well developed.

Subfamily Mustelinae (weasels and Old World polecats). Subfamily with widest distribution and most kinds. There are 11 genera and 33 recent species.

Subfamily Mellivorinae (ratel or honey badger). African and southern Asia. There is 1 species, *Mellivora capensis.*

Subfamily Melinae (badgers). North America, Europe, Asia, with most forms found in Southeast Asia. There are 6 genera and 8 recent species.

Subfamily Mephitinae (skunks). New World only. There are 3 living genera and 11 species.

Subfamily Lutrinae (otters). Worldwide. There are 5 living genera and 19 species.

Subfamily Leptarctinae. Fossil. Miocene and Pliocene. New World. 3 genera.

Suborder Pinnipedia

Aquatic; teeth nearly homodont (uniform) and adapted for grasping and tearing, not chewing; 2 pairs of incisors below; postcanines (premolars and molars) similar, never more than 2-rooted, and usually 5 on each jaw (varies from 3 to 7); carnassials absent; usual dental formula $\frac{3 \cdot 1 \cdot 5 \text{ (postcanines)}}{2 \cdot 1 \cdot 5 \text{ (postcanines)}} = 34$ teeth; all inclusive dental formula $\frac{1\text{–}3 \cdot 1 \cdot (3\text{–}7)}{0\text{–}2 \cdot 1 \cdot (3\text{–}6)} = 18\text{–}38$ teeth; cranial part of skull longer in proportion to facial part than in most fissipeds; face profusely innervated by extremely large trigeminal nerve; brain more spherical and with greater development of convolutions than fissipeds. Tympanic bullae small in some; alisphenoid canal present or absent; lacrimal duct absent; basioccipital and sphenoid relatively large as compared with fissipeds; pelvis small and nearly parallel to vertebral column; femur broad, flattened, with globular head on short neck; fibula almost as large as tibia; metatarsals and all bones of phalanges with epiphyses at both ends of shaft; metapodials elongate; 5 well-developed digits on each limb; 1 digit on manus and 1st and 5th digits of pes usually longer than other digits; feet fully webbed; nails small to absent; baculum massive. Fusiform-shaped body; external ears (pinnae) reduced or absent; tail short or vestigial (5 to 20 cm [2 to 8 in.] long) with 8 to 15 caudal vertebrae. Intestine long; cecum short; kidneys lobulate. Thick layer of subcutaneous fat (blubber). One large, precocial young.

Family Otariidae (eared seals). Lower Miocene to Recent. Coastlines of Pacific, South Atlantic, and Indian oceans. Hindlimbs useful for support on land; nails well developed on middle 3 digits, rudimentary on outer ones. Small external pinnae of ears. 2 pairs of lower incisors; upper incisors of 1st pair notched transversely; usually 20 to 22 postcanine teeth; dental formula $\frac{3 \cdot 1 \cdot 4 \cdot (1\text{–}3)}{2 \cdot 1 \cdot 4 \cdot 1} = 34$ to 38 teeth. Males are 2 to 4 times as large as females (males 120 to 1,000 kg [265 to 2,200 lb]; females 60 to 270 kg [130 to 600 lb]); testes suspended in distinct external scrotum; small but distinct tail; obvious neck; skin whitish or light gray.

Subfamily Otariinae (sea lions). Coastal regions of northern and eastern Pacific Ocean, southern Australia and southwestern New Zealand; coastal regions off eastern coast of South America. There are 4 living genera and 5 species.

Subfamily Arctocephalinae (Pacific and southern fur seals). Southern coastal areas of all southern continents; north Pacific region. There are 2 living genera, 7 species.

Family Odobenidae (walrus). Upper Miocene to Recent. Circumpolar holarctic distribution, seldom south of about 58° north latitude. Hindlimbs useful for support on land. External pinnae absent; postorbital process absent; alisphenoid canal present. Lower incisors absent in adult; upper canines of both sexes form tusks; usually 12 postcanine teeth; dental formula $\frac{(1\text{–}2) \cdot 1 \cdot (3\text{–}4) \cdot 0}{0 \cdot 1 \cdot (3\text{–}4) \cdot 0} = 18\text{–}24$ teeth; males larger (up to 1,270 kg; 2,800 lb) than females (up to 860 kg; 1900 lb). Testes abdominal (internal). No free tail. Skin whitish or light gray. Monotypic, *Odobenus rosmarus.*

Family Phocidae (earless seals). Middle Miocene of Europe and North America; Pliocene of Asia; lower Pleistocene of Africa; Recent along northern and southern coastal areas, sporadic in tropical and southern regions. Hindlimbs useless on land (cannot be placed forward under body); nails equally well developed on all digits; external pinnae absent; postorbital process rudimentary or absent; alisphenoid canal absent; 2 to 4 lower incisors; usually 16 to 20 postcanine teeth; dental formula $\frac{(3\text{–}2) \cdot 1 \cdot 4 \cdot (0\text{–}2)}{(1\text{–}2) \cdot 1 \cdot 4 \cdot (0\text{–}2)} = 26$ to 36 teeth; males may be much larger than females (elephant seal, hooded seal), slightly larger (some members of subfamily Phocinae), equal in size (some Phocinae and some Monachinae), or females may be larger than males (some Monachinae); olfactory apparatus reduced; testes abdominal; stubby tail; no visible neck; skin brown or black.

Subfamily Phocinae (northern seals). Northern temperate to Arctic waters; 7 living genera and 9 species.

Subfamily Monachinae (southern seals). Tropical and Southern Hemisphere waters; 6 living genera and 9 species.

Critical appraisal. The taxonomy of the major categories of major groups placed in the Carnivora has been in a state of flux for at least 100 years, and these categories do not seem to be stabilizing, even today. Most

mammalogists at present regard the seals and terrestrial carnivores as belonging to different orders, the Pinnipedia and Carnivora. There are, in reality, only a few features common to the seals and their terrestrial relatives because of the extensive and numerous adaptations the aquatic forms have undergone to make them efficient carnivores of the sea. Mammalogists who have studied seals intensively now realize that there is no anatomical structure unmodified by the extensive aquatic adaptations; every organ and tissue examined has been found to be different in some way from its counterpart in terrestrial forms. In many, if not most, ways there is more difference between seals and terrestrial carnivores than between the bats and the insectivores, and few mammalogists would place the bats and insectivores in the same order. Other mammalogists, tending toward conservative taxonomy, feel the relationship of the terrestrial and aquatic carnivores can be best expressed by retaining them in two suborders, the Fissipedia and Pinnipedia, of the single order Carnivora. The conservative attitude is retained in this article.

Divergence of seals

Of the 10 living families recognized in the Carnivora, two have separated from their lines most recently and are most easily associated with other existing families: the Odobenidae with the Otariidae and the Hyaenidae with the Viverridae. The family Procyonidae has perhaps the greatest number of taxonomic problems. It appears to be a collection of primitive forms whose relationships are not yet clear. Most of the living genera have at one time been placed in distinct families or subfamilies (Nasuidae, Bassariscidae, Ailuridae, Bassaricyonidae, Procyoninae, Potosinae). In most other families there are also genera sufficiently different from other members of their family that they have been given separate families (the canids *Lycaon* and *Otocyon;* the giant panda *Ailuropoda;* the otters; the viverrid fossa *Cryproprocta;* and the aardwolf *Proteles*).

The arrangement of the ten families into two distinct superfamilies, Canoidea and Feloidea (or Aeluroidea), appears to be a natural arrangement dating back to the works of W.H. Flower and H. Winge in the late 1800s. In the Canoidea, as revealed by studies in comparative anatomy and the fossil record, the families Canidae, Ursidae, and Procyonidae seem to be most closely related. Also placed in the Canoidea is the family Mustelidae, although some of the more primitive members show resemblances to the primitive viverrids as well as the canids. In the Feloidea, the families Viverridae and Hyaenidae seem most closely related, the Felidae being the most aberrant.

Those families that contain rather diverse lines have been subdivided into subfamilies; the number of subfamilies in each family indicating the amount of evolutionary divergence that has occurred. The groups that have probably been distinct the greatest length of time have the most subfamilies; the Viverridae with six and the Mustelidae with five. Within the viverrids some of the subfamilies are distinct, a few less distinct. Most distinct are the subfamilies Herpestinae and Cryptoproctinae. The Cryptoproctinae seems to be an ancient group with the genus *Cryptoprocta* retaining characteristics of the felids as well as the viverrids, so much so that some taxonomists, have placed *Cryptoprocta* in the felids, although most mammalogists regard them to be more similar to the viverrids. The remaining four subfamilies have been combined and split in a number of ways. Usually retained, along with the two above subfamilies, are the subfamilies Viverrinae and Galidiinae. Each of the five mustelid subfamilies seems to be distinct from the others. Some authorities feel that the skunks and badgers may be close enough to place them in the same subfamily. Of the five subfamilies, the most diverse is the Mustelinae with 11 distinct genera.

What much of the discussion above means, in terms of critical taxonomic appraisal, is that a system of classification is in some ways an artificial system set up for the convenience of man. Ideally, the system reflects real evolutionary relationships, but these must be inferred from a scanty fossil record and from comparisons of modern species. Since there are differences of opinion among specialists as to which taxonomic characters should be given the most weight, there are certain to be alternate classifications, the acceptability of which depends on new information continually being discovered. Just as the animals have evolved, so does the taxonomic system. (H.J.S.)

Cetacea (whales, dolphins, porpoises)

The order Cetacea consists of a group of primarily marine mammals occurring throughout the seas of the world and in certain tropical rivers and lakes. The name whale is often applied by scientists as a general name for the larger species; most of the smaller members of the order are called dolphins or porpoises. (The traditional grouping of these mammals in the single order Cetacea is not universally accepted. This disagreement is reflected in varying systems of classification [see below, *Annotated classification* and *Critical appraisal*]).

General features. The order Cetacea includes three distinct suborders: Archaeoceti, an extinct group of toothed whales; Odontoceti, the modern toothed whales; and Mysticeti, the baleen, or whalebone, whales. In addition to skull characteristics, the three groups are distinguishable by dentition: the archaeocetes were heterodont (*i.e.,* the anterior teeth were quite different in shape from the posterior ones), and milk teeth in the young were replaced by permanent teeth in the adult, as in most mammals; the odontocetes are mostly homodont (teeth of uniform shape), and most lack milk teeth; adult baleen whales lack teeth, and the animal feeds by straining minute organisms from the water by the baleen plates (see below). The odontocetes include about 70 species—the porpoises, dolphins, pilot whales, or "potheads," false killer whales, beaked whales, the sperm whale, and a few lesser known forms. They range in length from about 1.3 metres (4.3 feet) in the smaller porpoises to about 18 metres (60 feet) in the male sperm whale. The length of the ten species of baleen whales ranges from six metres (about 20 feet), in the case of the pygmy right whale (*Caperea marginata*), to about 30 metres (just under 100 feet) in the largest recorded specimen of the blue whale (*Balaenoptera musculus*). A great range of adult weight is encompassed by the Cetacea, from about 45 kilograms (100 pounds) in some small porpoises to about 136,000 kilograms (150 tons) in the blue

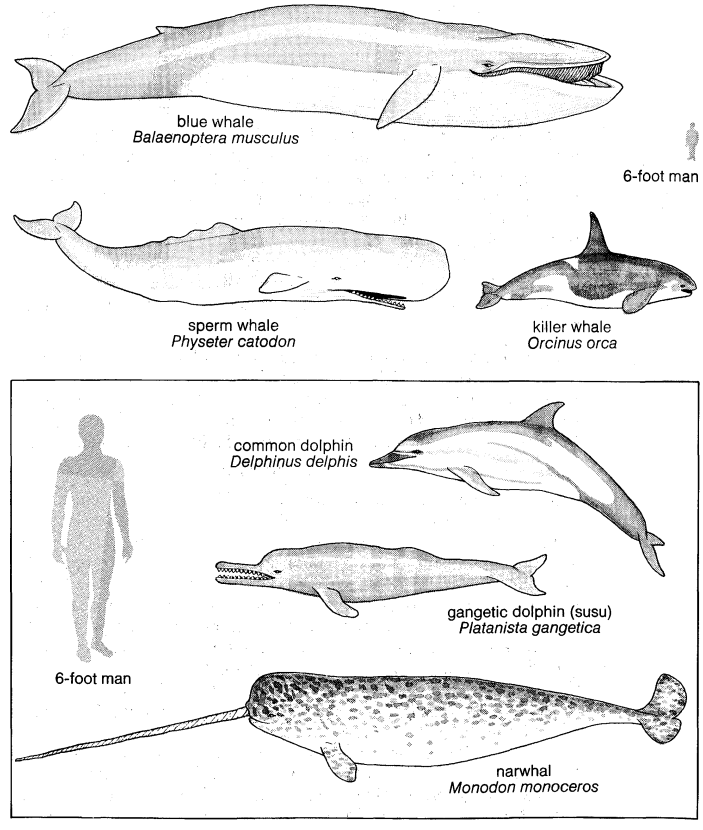


Figure 40: Body plans of representative cetaceans.

whale. The body of modern whales is fusiform (tapered); the tail, which ends in a horizontal blade (the flukes), is used for propulsion. The flukes are unsupported by bone. There is usually a dorsal fin on the back, also unsupported by bone. Breathing is accomplished rapidly at the water surface through paired (in mysticetes) or single (in odontocetes) blowholes, which are generally located on top of the head, some distance behind the tip of the snout. The exhalation produces the familiar spout, through the sudden condensation of water contained in the breath. Following inhalation, the animal holds its breath for variable periods, as it swims beneath the surface. Dives of smaller cetaceans last a few minutes, but those of some larger species may last an hour and perhaps longer.

Breathing through blowholes

The cetaceans are an ancient group, having split off from other mammals early in the history of the Mammalia. In spite of their many and profound adaptations for aquatic life, cetaceans are typical mammals, developing their young internally and giving birth at sea, nursing the young with milk, and possessing internal heat control mechanisms, or warm-bloodedness. Although hair is largely absent in most adult whales, the evidence from fetal or newborn animals indicates that the lack of hair is a secondary loss (*i.e.,* that hair was present in the ancestors of whales).

Importance to man. Many of the more abundant whales and porpoises are commercially important, their meat being used as food for animals and (especially in Japan) for humans and their oil for industrial lubrication and for conversion into soaps and fatty acids, which are used in cosmetics and detergents.

About two dozen countries have engaged in commercial whaling at one time or other, but by the early 1970s only Japan and the Soviet Union maintained whaling fleets in open ocean, the fleets of other countries having ceased operations as a result of declining markets for whale products and drastically reduced populations of the commercially valuable whales. Several other countries have continued to take whales from shore stations. Limitations on catches of the larger cetaceans, imposed by the International Whaling Commission (IWC), were too late to prevent the near extinction through overexploitation of the blue whale by the late 1950s. Populations of the fin whale (*Balaenoptera physalus*) have declined sharply under heavy fishing pressure. The decline in these species has led to increased pressure on the sei whale (*B. borealis*) and the sperm whale (*Physeter catodon*), by 1970. Effective regulation of commercial whaling in international waters to a level that would allow a sustained yield without depleting whale populations has been handicapped by the inability of the IWC to impose obligatory controls on its members. In 1971 the United States declared the commercially exploited whales to be endangered species and prohibited by law the importation of all whale products.

NATURAL HISTORY

Behaviour. *Social behaviour.* Most toothed whales spend their entire lives in tightly organized schools, which may range in numbers from a few animals to 1,000 or more individuals. Baleen whales are more often found singly, although small schools, or pods, occur on the breeding grounds, and whales may congregate in feeding areas in considerable numbers. Odontocete schools, at least, may be quite complex in structure, involving family groups, groups segregated by age and sex, and even schools composed of two or more species. Bottle-nosed dolphins (*Tursiops truncatus*), for example, are often found associated with schools of pilot whales (*Globicephala* species) in some areas and move with them, apparently because the food-finding capability of the large whale school is greater than that of the smaller dolphin group. Schools of toothed whales show a variety of swimming patterns, including movements as a rank or front, single file, or a more formless pattern. Within schools, subgroups usually swim, feed, and dive independently, except when the school is frightened or travelling rapidly; then members may pack together and leap and dive more uniformly. Groups of mothers with young are usually found together near the centre of a school.

Swimming patterns of toothed whales

Studies of captive odontocete schools, especially of bottle-nosed dolphins, have shown that mother–young relations may persist for several years, with the young returning to its mother in times of stress even when full grown. Other nonrelated females may also be part of such groups and may assist during birth and in defense against males. Adult males tend to segregate within the school, except during the breeding season; juvenile males may gather in small groups. Dominance hierarchies exist in which both males and females exert leadership, though dominance is strongest in large adult males. Aggression by the leader is used to establish such relationships and may take the form of biting, hitting the offender with the tail, clapping the jaws intimidatingly, or ramming with the snout. Juvenile males of many species are especially apt to be heavily scarred with tooth marks. Biting or raking with the teeth is also a marked part of sexual activity, and females are therefore apt to be scarred, especially in the genital area.

Play. Play, although common in odontocetes, is almost unknown in mysticetes. In porpoises, much play is sexual in nature, consisting of rubbing or prodding the genitals with the snout or fins and flippers, usually accompanied by sound production of a complicated character. Other forms of play consist of balancing floating objects, such as kelp, sticks, or feathers, on the fins and flukes, or holding objects in the mouth. Younger killer whales and sperm whales have been seen pushing small logs or planks; food items are also used. Fish may be thrown into the air or proffered to other fish lurking in crevices, apparently to lure them into the open. Underwater observations at sea have revealed porpoise groups swimming and diving together in complicated ballet-like movements, riding ocean swells or breaking surf, in precise formation swimming. In pilot whales, playlike swimming patterns have been observed, the animals resting vertically, head down, with a metre or so of their tails in the air, or swimming on their backs.

Epimeletic behaviour. Behaviour in which one animal assists another in trouble (epimeletic behaviour) is marked in whales, having been noted in both mysticetes and odontocetes. It consists of "standing by," in which school members refuse to leave a wounded animal or actually support a sick animal, sometimes one of a different species. An extreme form sometimes occurs in porpoises: a mother may support the body of her stillborn calf until it literally rots away. Most whales and porpoises appear to be solicitous of their young, and there are several reports of babies being shielded by the mother or taken on a pectoral flipper and rolled part way out of water.

Fright. Fright is typically shown by a bunching together of school members and rapid swimming, a reaction that has been used by fishermen to capture certain species. The school is frightened by noise production underwater—banging on pipes or metal bars or clacking two rocks together. The animals are run until their shortage of breath prevents deep dives and underwater evasion, and eventually they are herded ashore or into nets. Porpoises sometimes signal moderate levels of fright by slapping the water with the tail flukes.

Causes of fright

Feeding. Two general modes of feeding exist in cetaceans. Mysticetes are strainers who envelop large amounts of water, forcing it through horny fibrous mats of baleen (slatlike horny plates) that hang from the roof of the mouth and swallowing the food that remains on the inner surface of the baleen. The odontocetes actively pursue and capture swimming prey, especially fish and squid. Two types of feeding have been described in baleen whales: the skimming mode, in which the whale swims forward with the mouth slightly open, filtering a continuous column of water as it moves; and the gulping mode, in which the animal engulfs a mouthful of water and forces it out past the baleen before taking another mouthful.

Life history. *Reproduction.* Although all cetaceans are seasonal breeders, this period may be quite long. Adult male porpoises tend to accompany the females who are receptive. Impregnation occurs during the long period from spring to fall and, in both the large baleen whales and the porpoises, birth takes place about 11 to 12 months afterward. The sperm whale is an exception, gestation taking about 16 months. Growth is fairly rapid in all cetaceans,

and the birth size is large; a 1.6-metre (5.2-foot) harbour, or common, porpoise (*Phocoena phocoena*) is known to have given birth to a calf 0.6 metre (two feet) long, and a 30-metre (100-foot) blue whale was found to contain a full-term fetus 7.54 metres (24.7 feet) long.

Birth of whale calves takes place underwater; usually, the tail emerges first. The newborn porpoise is often pushed to the surface by its mother or by an attendant female for its first breath. After the umbilical cord snaps, close to the umbilicus, the baby begins to swim. The fins and flukes are folded and flaccid at birth but harden into proper position within a few hours. The young of both porpoises and larger whales accompany the mother very closely for the first few weeks, often positioning themselves at about the midbody of the parent and taking advantage of water flow, so that they are carried along with a minimum of effort.

Baleen whales apparently breed once every two years and nurse young from seven to more than ten months. The nursing span is more protrated in toothed whales: the young bottlenose dolphin may continue to nurse up to the age of 18 months and the young sperm whale for 12 to 13 months.

Mammary glands and milk

The paired mammary glands in all whales are hidden beneath the body blubber, the nipples being buried in a longitudinal fold located lateral to the genital aperture. Milk, which is secreted into the enlarged lacteal duct, is thought to be forced into the mouth of the baby by contractions of body muscles underlying the mammary gland. In baleen whales and most porpoises, the baby thrusts its snout into the mammary slit, taking the nipple in its tongue and receiving the milk; the tongue, when applied to the palate (roof of the mouth) forms a tube for receiving the milk. The baby sperm whale, which has a bulbous snout, takes the nipple sideways in its jaw. Whale milk is very high in fat content and almost lacking in milk sugar.

Migration. Most baleen whales undergo seasonal migrations, some as long as 5,000 kilometres (3,000 miles) each way, from feeding to calving grounds. With the exception of the sperm whale, odontocete migrations seem to be much more local in character and in some species apparently do not occur at all. There is a seasonal migration of sperm whales from calving grounds near the Equator toward the higher latitudes, but only adult males go beyond the temperate zones.

Summer concentrations of planktonic (free-floating) food, such as the euphausiid crustacea, or krill, are greatest in polar latitudes; hence, baleen whales congregate in these waters to feed. Movements toward the Equator occur in the austral (southern) or boreal (northern) fall and apparently terminate within about 30° of the Equator, usually far at sea, where birth of young takes place in the warm waters. The humpback whale (*Megaptera novaeangliae*) typically frequents island shores at this time, and mothers and young may be seen especially near subtropical islands. Long migrations are apparently made wholly or nearly completely without feeding, except on the summer grounds; the animals live off body fat reserves in the blubber coat. Some baleen whales feed closer to the Equator in such areas as the Persian Gulf and the Gulf of California; migrations in these populations, therefore, are probably of lesser extent. The migratory cycles of the gray whale (*Eschrichtius robustus*) and the humpback conform closely to an annual day length cycle, suggesting that a photoperiodic effect may be involved in the regulation of reproduction.

In the smaller toothed whales, some food-related seasonal movements occur in which porpoises and whales may congregate in certain areas in response to seasonal abundances of squid or fish. Other porpoise species are sedentary, and some continuously occupy discrete portions of a shoreline. Odontocetes apparently do not undergo prolonged periods of starvation and thus must have continuously available food supplies.

Locomotion. *Swimming and diving.* Swimming is accomplished by vertical undulatory movements of the tail, the thrust coming from both the horizontal flukes and the posterior part of the body. Typically, a swimming cetacean describes a roughly sinusoidal (undulatory) path as it swims, coming to the surface for a breath at the top of each curve and descending below the surface in between. Because breathing is accomplished through the paired or single blowhole on top of the head, the animal does not have to arch its neck to breathe and thus break the smooth path of travel. Only in the sperm whale is the blowhole anterior on the left side of the snout tip. The flukes are frequently thrown free of the water before steep dives, a habit uncommon in rorquals (*Balaenoptera*).

Small cetaceans normally do not dive for periods exceeding five to seven minutes, but the beaked whale (*Ziphius cavirostris*), the bottle-nosed whale (*Hyperoodon ampullatus*), and the sperm whale may stay below for very long periods; the sperm whale may routinely dive for an hour below the surface, and specimens that have become tangled in submarine cables indicate that it may reach depths of at least 1,100 metres (about 3,600 feet).

Length of dives

Vincent Serventy

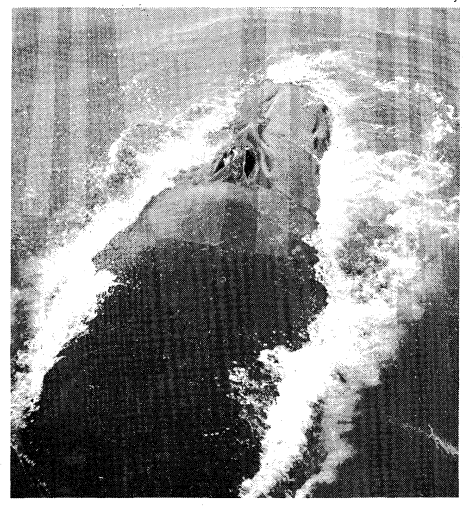


Figure 41: Humpback whale surfacing, showing blowholes.

Speed. It has been reported that porpoises travel at speeds of at least 38 kilometres per hour (21 knots, or 24 miles per hour) and whales at 56 kilometres per hour (35 miles per hour). Tests on trained animals suggest that cetaceans possess unusually low drag because the boundary layers (water near the skin) remain partially laminar (smooth) at high speeds and do not become predominantly turbulent, as is the case around moving rigid bodies. Body shape is believed to play an important part in the low drag experienced by swimming whales.

Leaping and wave riding. Many whales and porpoises leap from the water from time to time, and some perform aerobatic manoeuvres. Sperm whales reportedly can jump clear of the water; humpback whales may mate as they leap in tandem from the water; and many porpoise species leap during fast swimming. Some species trained in captivity have vaulted as high as six metres (20 feet) above the water surface. One species group of the genus *Stenella* spins rapidly along its longitudinal axis as it leaps. Some of these leaps are thought to aid in shaking off suckerfish (*Remora*), but use in social contexts is also evident. The larger rorquals seem never to leap; the gray whale spy hops, or pitchpoles, often rising to about the level of its flippers and hanging for a moment, apparently viewing the aerial surroundings; or it may leap three-quarters of its length and fall back on its side.

Many species of smaller whales and porpoises come to the bows of vessels and may ride for many minutes without beating their tails, obtaining a thrust from the pressure wave in front of the ship's bow wave. Porpoises have also been observed riding the "bow wave" produced by a large mysticete when the latter was at the surface.

Sound production and communication. Acoustic behaviour, which is highly developed in all cetaceans, involves passive listening, social signalling, and echolocation. Although the acoustic sense and passive listening are developed in all cetaceans, only odontocetes have been shown to echolocate—*i.e.*, to use the echoes of their own

signals for discrimination and navigation. Two major sorts of sounds are produced by cetaceans: low-pitched signals predominantly in the human audible range, such as barks, whistles, screams, and moans; and brief clicks of high-intensity sound, some as high as 200 kilohertz, or about 13 times the upper frequency of humans. The former sounds seem to serve mostly in social communication; the latter serve largely in discrimination and navigation. Click production in porpoises allows highly refined discrimination; in experiments, porpoises can discriminate between objects of slightly differing sizes, detect the presence of fine wires, determine the locations of tiny objects, and discriminate among different kinds of metals. Clicks are projected from the porpoise forehead and upper jaw in a highly structured beam; the highest frequencies are strongest directly ahead of the animal. Low-frequency sounds are more omnidirectional, and their source in dolphins is thought to be in structures within the left nasal passage.

Types of communicative signals

Communicative signals are used in a variety of contexts. Humpback whales produce a variety of moans and calls in their breeding areas, porpoises emit distress cries, and specific signals are produced during birth. The finback whale produces a monotonously repeated, very low-pitched, narrow band sound (about 20 Hertz) that is thought to allow long distance communication between whales. Captive porpoises have mimicked human voice signals and artificial sounds, and a variety of sounds have been shaped by training techniques. No evidence exists, however, for a humanoid language in any cetacean.

Nonvocal signalling of emotional states, which occurs in porpoises and small whales, has not been studied in the larger whales. Such signals include showing the whites of eyes, which is a sign of rage, and various jaw claps and body postures, which carry social meaning.

FORM AND FUNCTION

General features. The general body form is fusiform, or spindle-shaped, with the head end variously modified into a more or less attenuated beak, rounded, bluff, or flattened. The tail always ends in a horizontal blade (the flukes), through which vertical movements of the tail produce forward thrust. The forelimbs are paddle- or sickle-shaped flippers that function in balance and steering. A dorsal fin is usually present. The flippers have all the recognizable elements of the mammalian forelimb skeleton. No external trace of the hindlimbs remains, but a greatly reduced pelvis is represented by a pair of slender, irregularly curved bones, remote from the backbone and embedded in the flesh in the vicinity of the reproductive opening. In some of the larger whales, bony or cartilaginous remnants of the hindlimb skeleton still persist as attachments to the pelvis.

The head varies in size from about one-tenth of the total length in some toothed whales, such as dolphins, to nearly one half in the sperm whale. No ear pinnae are to be distinguished adjacent to the inconspicuous opening of the external ear tube. Eyes and eyelids are present and usually functional, but vision is said to be reduced in some river dolphins.

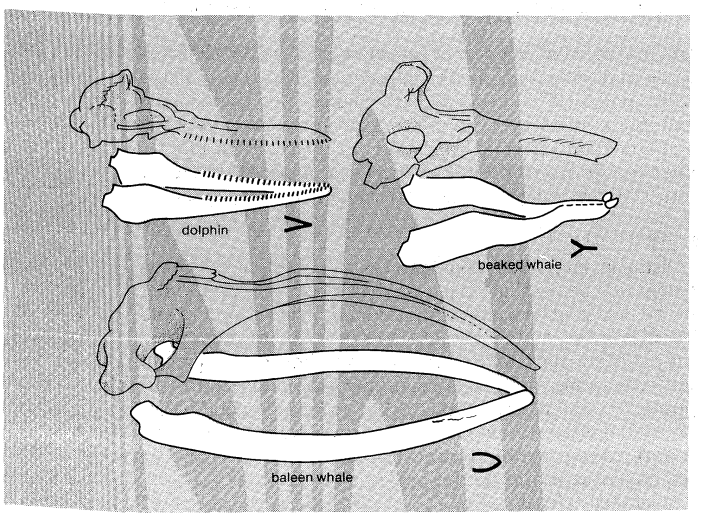


Figure 42: Three cetacean skulls. The small figures represent the general shape of the jaw as viewed from above.

The facial part of the skull is elongated into rostrum, or beak (Figure 42). The maxillary bones are spread laterally at their posterior ends and either override the frontal bones (in the toothed whales) or extend under the orbital processes of the frontal bones (in baleen whales). The telescoping of the skull and the laminated (layered) bony structure, together with the high, short, broad braincase, result in a skull form that is far removed from the usual mammalian pattern.

Feeding adaptations. The digestive tract and the dentition show several special features associated with feeding habits. Food enters the large gape and passes unchewed through the esophagus and into a specialized multichambered stomach. Specialization of the stomach is related to the simplification of dentition, which requires that the food, swallowed whole, undergo prolonged digestion. In the modern carnivorous cetaceans (the odontocetes), the teeth are simple, uniform in shape, peglike, and single rooted, frequently exceeding the primitive mammalian number (44). When the mouth is closed, the top teeth interdigitate with those of the lower jaw, rather than occluding, as do those of most land mammals. The teeth are used for seizing prey, which is not chewed before being swallowed. Among the dolphins, many species have reduced dentition. Risso's dolphin (*Grampus griseus*), which feeds predominantly on cuttlefish, has the teeth reduced to from two to six on each side of the lower jaw and none in the upper jaw. This reduction of teeth in association with a cuttlefish diet is even more evident in the bottle-nosed and beaked whales (Hyperoodontidae), whose slender jaws bear one pair (occasionally two pairs) of functional teeth in the mandible and none in the upper jaw, except in the Tasmanian beaked whale (*Tasmacetus shepherdi*), which has teeth in both jaws, totalling up to 90. The sperm whale, in which functional teeth are restricted to the lower jaw, eats large squid extensively, but fish are also taken.

Dentition

The baleen whales are plankton feeders, including in their diet small crustaceans such as krill (euphausians) and copepods (*Calanus*), and fishes such as herring, sardines, and capelin, which they strain from large amounts of water with the baleen plates. These are anchored to the roof of the mouth, somewhat spaced apart at right angles to the long axis of the head. Each plate is triangular, with its smallest side inserted in the jaw. The outer (labial) edge of the plate is smooth; the inner (lingual) side is frayed out into a fringe of bristles that collectively form a matted sieve or strainer. The strainer selectively removes the plankton from the water expelled from the whale's mouth. The plates are longest (to nearly four metres, or 13 feet) in right whales and shorter, broader, and stiffer in rorquals. Each half of the total baleen complement may comprise upward of 300 plates. Each plate is formed of parallel tubes of horn (keratin), with compacting horn between, which are enveloped, except at their fringed ends, in a covering layer. Between adjacent plates is a pad of softer horn. Baleen does not differ essentially from hair in chemical composition and, like the latter, has its origin in the skin.

The senses. *Smell.* The sense of smell is reduced or lacking in cetaceans. A reduced olfactory organ is, however, present in baleen whales; on the other hand, it is absent in odontocetes.

Sight. Underwater vision is excellent in porpoises (in species in which it has been tested), except for some river porpoises. The eyes of the Amazon River porpoise (*Inia geoffreyensis*) are reduced but apparently functional; those of the susu of the Ganges, Brahmaputra, and Indus rivers (*Platanista gangetica*) are rudimentary and vision is poor. Some evidence indicates that porpoise vision out of water is poor. The lens of the cetacean eye is modified for underwater vision: the sclerotic capsule of the eye is greatly thickened, and a variety of adaptations relating to pressure change and vision in water are present. There is no definite evidence for colour vision in cetaceans.

Hearing. Hearing is a major sense in all cetaceans.

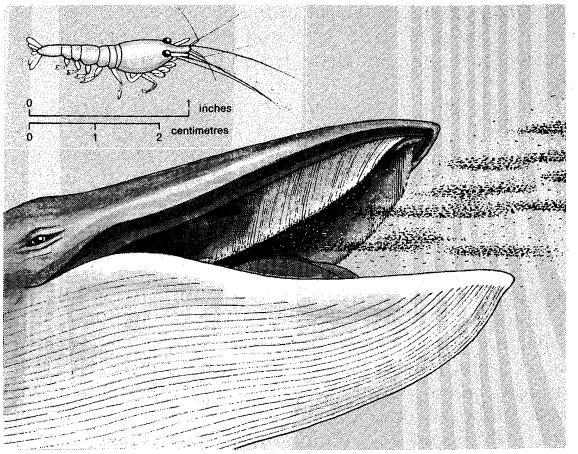


Figure 43: Fin whale swallowing krill, one of which is shown in the inset above.

In the bottle-nosed dolphin hearing capability extends through the human range (20 Hertz to 20 kilohertz) to about ten times the human upper limit (up to about 150 kilohertz). In odontocetes, the hearing of clicks is highly directional and is apparently mediated through the lower jaws and throat, rather than through the much reduced external ear canal. Three species of Japanese porpoises are approximately 10 times as capable as a human of discriminating pairs of clicks given in rapid succession.

Little is known about the sound reception of baleen whales other than evidence from whalers who know them to be extremely sensitive to boat noise.

Structure of the ear

The ear cartilages, of similar origin to those that support the pinna (external ear) in land mammals, have been sunk in the blubber over the external ear. In baleen whales, the external ear canal becomes a solid cord; it opens again near the eardrum and is filled with a solid waxy plug the laminations of which have been used for age determination, much in the manner of tree rings. In odontocetes, the tiny canal is generally open during its entire convoluted length. In one species, the Dall porpoise (*Phocoenoides dalli*), it is at least occasionally occluded at the surface. The middle ear ossicle bones differ widely from those found in terrestrial mammals, apparently in response to the conditions of hearing underwater. The middle ear bones and the inner ear (cochlea) are housed within a heavy, almost ivory-like, petrotympanic complex, which hangs nearly free from the skull and is surrounded by air sacs connecting with the respiratory tracts by way of the eustacian tube. These sacs are important in pressure equilibration during dives and in acoustic isolation of one ear from the other and of both ears from the animal's own sound signals. Masses of small blood vessels occupy much of the middle ear and are also instrumental in pressure equalization.

Taste. Marked food preferences have been noted in captive odontocete cetaceans. The presence of well-developed taste buds on the tongue point to a well-developed sense of taste.

Proprioception. The awareness of bodily position (proprioception) is as acute in cetaceans as in any other mammal. Their exquisitely refined capabilities in manoeuvering in the water provide other evidence of coordinative sense.

(K.S.N.)

PALEONTOLOGY AND EVOLUTION

It is generally agreed that whales originated in the early Paleocene or Upper Cretaceous (about 70,000,000 years ago) from some group of terrestrial meat-eating mammals. Some authorities believe that a likely parent group was the Mesonychidae, medium-sized, large-headed mammals of the order Creodonta, believed to be near the base stock of both the carnivores (Carnivora) and the ungulates (several orders of hoofed mammals). It has also been suggested that the whales evolved from a Paleocene member of the order Insectivora resembling the Eocene *Pantolestes,* a semiaquatic carnivorous animal about the size of the modern otter (*Lutra lutra*). An American paleontologist, L. Van Valen, has pointed out similarities between the mesonychids (as well as the closely related arctocyonids) and the earliest archaeocete whales, the protocetids. All that can be stated with certainty is that the ancestors of the protocetids must have evolved through an amphibious stage.

The earliest known whale fossil is a fragmentary shoulderblade from the lower Eocene of England (about 50,000,000 years ago). Remains of archaeocete whales of the family Protocetidae are found in middle Eocene formations. They were relatively small animals about the size of the modern porpoises (two to three metres long). The hindlimbs had already been lost, and the nostrils were midway between the terminal position found in most terrestrial vertebrates and the top of the head, where they are found in most advanced whales. The snout was elongated, but the skull bones were not telescoped as in modern forms; the teeth retained the differentiated (heterodont) condition found in terrestrial carnivores, the front teeth being tapered pegs and the cheek teeth serrated. At the time of their greatest development, in the late Eocene, some archaeocetes were large, the most striking being *Basilosaurus,* a slender animal that attained a length of some 20 metres (nearly 70 feet). One line of archaeocetes, the dorudontids, persisted into the early Miocene (about 20,000,000 years ago).

Paleontologists disagree as to whether the modern toothed whales (odontocetes) and baleen whales (mysticetes) arose from an early archaeocete relative or separately from different terrestrial ancestors. Recognizably odontocete remains are known from the upper Eocene, and even earlier odontocetes must have been contemporaries of the early archaeocetes. One early form, *Agorophius,* already showed some telescoping of the skull and the nostrils were above the eyes. The agorophiids appear to have had a relatively short history but were probably ancestral to the porpoise-like squalodonts, which appeared in the late Oligocene (about 28,000,000 years ago) and by the middle Miocene were the dominant toothed whales of the world's oceans. By the early Miocene most of the modern odontocete families were represented by some members. Fossils of sperm whales found in lower Miocene deposits have been assigned to the modern genus *Physeter.* These are among the oldest known representatives of the physeterid line, which produced a variety of other Miocene and later forms and must have existed as a well-differentiated group during the later part of the Oligocene.

Intermediate fossil forms

Two groups of fossil whales appear to be intermediate between the toothed and the baleen whales. In the Patriocetidae the nostrils were in an intermediate position between the snout and the forehead; the telescoping of the skull (in the symmetrical manner seen in later mysticetes) was only slightly evident. The other group, the Aetiocetidae, was placed in the Odontoceti on the basis of possession of teeth, but Van Valen has pointed out that *Aetiocetus* is mysticete in other respects and should be considered closer to the mysticete evolutionary stem.

The earliest fossils definitely assignable to the Mysticeti are the middle Oligocene cetotheres, a group that became quite abundant in the subsequent Miocene. Of the three modern families of baleen whales, two (Balaenidae and Balaenopteridae) are known from scanty Miocene and abundant Pliocene fossils; the third (Eschrichtiidae) is known only from a few Pleistocene and post-Pleistocene fossils representing the sole modern genus, *Eschrichtius.*

CLASSIFICATION

Distinguishing taxonomic features. The three suborders recognized below share the same basic body plan but differ in the degree of specialization attained. The suborders and families are separated primarily on the basis of the following characteristics: tooth structure, number, and degree of differentiation; modifications of the skull, especially the position of the nostrils, degree of telescoping of the whole skull, and extent of joining of the two halves of the jaw (rami); and degree of modification of the pelvic girdle (in the archaeocetes only).

Annotated classification. The classification presented below is based on research by cetologists F.C. Fraser, R. Kellogg, and a number of other modern authorities.

Groups marked with a dagger (†) are extinct and known only from fossils.

ORDER CETACEA

Aquatic mammals with forelimbs modified into flippers and hindlimbs lacking; pelvic girdle vestigial and not attached to vertebral column; tail laterally flattened and extended into horizontal flukes, supported by cartilage. External nares at top of head (except in sperm whales and archaeocetes); cranium telescoped variably, with elongated rostrum. Three suborders, with about 10 fossil and 9 Recent families.

†**Suborder Archaeoceti** (archaeocetes or zeuglodonts)

Fossil only; lower Eocene to middle Miocene. Anterior and posterior teeth differentiated; total teeth not exceeding 44, the basic number in terrestrial placental mammals; nostrils positioned toward the top of the head, to varying degrees. More than 12 genera described.

†*Family Protocetidae*

Lower to upper Eocene; Europe, Africa, and possibly North America.

†*Family Dorudontidae*

Upper Eocene to lower Miocene; Europe, Africa, North America.

†*Family Basilosauridae*

Upper Eocene to lower Oligocene; Europe, Africa, North America.

Suborder Odontoceti (toothed whales)

Upper Eocene to Recent; worldwide marine, a few in fresh water. Teeth relatively uniform (homodont) and numerous (up to 300), occasionally reduced. Skull asymmetrical; nostrils on top of head (except in Physeteridae) with single external opening (blowhole). Several pairs of ribs joining sternum, which is fused into a single bone only in the adult.

†*Family Agorophiidae*

Fossil only; upper Eocene of North America. Two genera.

†*Family Squalodontidae*

Fossil only; upper Oligocene to lower Pliocene; Europe, North and South America, Australia and New Zealand. More than 12 genera described.

Family Platanistidae (river dolphins)

Lower Miocene to Recent. Snout long and slender; neck vertebrae all free; sight reduced. Flippers short and broad, sternum well developed. About 10 fossil and 4 monotypic Recent genera, the latter respectively in the Ganges, Indus, and Bramaputra rivers of India; La Plata of Argentina; Amazon and Orinoco of South America; and Tung-T'ing Lake and adjacent Yangtze River, China. Adult length 1.5–2.5 metres (5–8 feet).

Family Hyperoodontidae (beaked whales)

Previously called Ziphiidae. Lower Miocene to Recent. Functional teeth greatly reduced in number, to 1 or 2 pairs in lower jaw (except for total of 90 teeth in *Tasmacetus*); vestigial (nonfunctional) teeth not uncommon. Pair of longitudinal grooves in throat region. Dorsal fin slightly more than one-third of way back along body. Tail usually not notched in middle. Flippers small. Worldwide, marine; length to about 12 metres (40 feet). About 14 fossil and 5 Recent genera; 15 Recent species.

Family Physeteridae (sperm whales)

Lower Miocene to Recent. Head proportionately large, with bulbous, squared snout; mouth narrow and ventral; lower teeth totalling 40–52, upper teeth vestigial, smaller, varying in number. Blowhole a single, asymmetrical opening on the anterior left tip of snout. Twenty fossil and 2 Recent genera; length to 19 metres (62 feet; *Physeter*) and to 3 metres (10 feet; *Kogia*) in Recent genera. Tropical and temperate oceans, and (adult males only) polar waters.

Family Monodontidae (narwhal and beluga)

Pleistocene to Recent. Dorsal fin lacking. Neck vertebrae free. Flippers broad, rounded at tips. Teeth reduced to 8 or 10 in *Delphinapterus* (beluga); all vestigial in *Monodon* (narwhal) except for 1 left tooth of male, which grows into a long, straight tusk, extending in front of animal. Two genera; length to about 5 metres (16 feet); Arctic Ocean.

Family Phocoenidae (porpoises)

Upper Miocene to Recent. Head with rostrum (forehead extending to snout); teeth with expanded spade-shaped crowns. Dorsal fin triangular (when present); tail notched posteriorly. Three fossil and 3 Recent genera. Adult length (in Recent species) to about 2 metres (6.6 feet); virtually worldwide; marine, 1 species ranging up Yangtze River.

Family Delphinidae (dolphins and killer whales)

Lower Miocene to Recent. Characteristics variable. Dorsal fin and well-defined beak present or absent. None with expanded tooth crowns. About 35 fossil and 12 extant genera. Length 1.7 to about 9 metres (5.5 to 30 feet), worldwide, marine; 1 species in salt and fresh water (Irrawaddy River).

Family Stenidae (long-snouted dolphins)

Lower Miocene to Recent. Like Delphinidae but differing in structure of air sinus system and in shape of forehead. Three genera. Widespread in tropical oceans and rivers.

†*Families Eurhinodelphidae, Hemisyntrachelidae, and Acrodelphidae*

Three families of Miocene toothed whales; about 12 included genera, from all continents.

Suborder Mysticeti (baleen, or whalebone, whales)

Teeth lacking in adult; numerous but vestigial in embryo; baleen present. Lower jaw large, mandibles bowed outward and loosely united in front. Skull symmetrical; nasal bones relatively well developed; blowholes paired slits, slightly divergent posteriorly.

†*Family Aetiocetidae*

Upper Oligocene. Toothed but with symmetrical skull and other typical mysticete features. One genus (possibly 2); North America.

†*Family Patriocetidae*

Upper Eocene to upper Oligocene; Europe and North America; 4 genera.

†*Family Cetotheriidae* (cetotheres)

Middle Oligocene to lower Pliocene; North and South America, and Europe; about 30 genera.

Family Eschrichtiidae (gray whale)

Formerly Rhachianectidae. Head less than one-quarter of total length; neck vertebrae not fused. Dorsal fin lacking. Flippers with 4 digits. Pelvis large. One recent species; length to about 13 metres (45 feet); found along both coasts of the North Pacific Ocean. Fossils from Pleistocene and Recent, from Atlantic Ocean.

Family Balaenidae (right whales)

Lower Pliocene to present. Skull much arched; neck vertebrae fused; baleen long, narrow, flexible strips. Ventral grooves on throat lacking. Four fossil and 2 Recent genera (4 species); length 6–18 metres (20–60 feet); polar and subpolar oceans.

Family Balaenopteridae (rorquals and humpback)

Upper Miocene to present. Skull broader and less arched than in Balaenidae; baleen plates shorter, broader, less flexible, neck vertebrae not fused. Dorsal fin present; flippers narrow. Conspicuous longitudinal grooves present on throat. Eight fossil and 2 Recent genera (8 species); length 10–30 metres (33–100 feet); blue whale (*Balaenoptera musculus*) is the largest animal that has ever lived. Family occurs in all oceans and, at various times, from Equator to poles; migratory.

Critical appraisal. Authorities disagree on whether the whales should be treated as one order, Cetacea (as above), two, or three (as in the classification used in the section *The class Mammalia*). The determining factor is the degree of shared ancestry, the evaluation of which remains controversial. The possibility that certain protocetids could have given rise to both the modern groups of whales leads to the use of a single order, but the absence of intermediate fossils linking the baleen whales with the toothed forms supports the use of separate orders. Resolution of this problem must await the discovery of relevant fossil material.

Below the ordinal level, there is further disagreement at many points. Although there is no doubt that any recent cetacean is odontocete or mysticete, the relationships of many genera are in doubt. The long-snouted dolphins (Stenidae) are of uncertain affinities but are included in the Delphinidae by many authorities, as are the porpoises (Phocoenidae). At the species level, there is uncertainty about the specific or subspecific status of many populations.

(Ed.)

Proboscidea (elephants)

The order Proboscidea comprises three suborders and about 300 species of terrestrial mammals. All but two species, the Asiatic, or Asian, elephant (*Elephas maximus*) and the African elephant (*Loxodonta africana*), are extinct. The elephants are the largest surviving land ani-

mals and, among the mammals, are exceeded only by the whales in size.

The Proboscidea are characterized by columnar limbs, bulky bodies, and elongated snouts. In recent forms, testes are internal. The snout is a long boneless proboscis, or trunk; it is a combination of the upper lips, palate, and nostrils. Some of the incisor teeth develop into tusks. One extinct suborder (Deinotherioidea) lost the upper tusks; certain others have lost the lower ones and evolved upper tusks of dentine from which the enamel has partially or completely disappeared. The canine teeth were generally repressed in all groups, and the cheek teeth developed rows of blunt cones or ridges. In later forms, the temporary teeth were replaced by permanent ones, which are pushed by an escalator-like movement along a horizontal plane, so that the front teeth were replaced by teeth moving forward from the rear. The skull, which originally was elongated, became shorter, higher, and bulkier in later forms. The back of the eye orbit remained open instead of forming a complete bony ring, and the nasal opening in all Proboscidea is at a higher horizontal plane than the eye sockets. The neck shortened as the animals evolved larger, higher bodies and an elongated trunk that also functions as a hand. The skull has enlarged out of proportion to the brain in order to serve as an anchor for the trunk and to support the heavy dentition. This order occurs in all the continents except Australia. Fossils of proboscideans provide valuable information about early humans who were their contemporaries.

GENERAL FEATURES

Size range and distribution. In Europe, the landmass that broke up to form the islands of the Mediterranean Sea harboured proboscideans. Three fossil species have been found in Malta; one had had a height of 2.1 metres, or 6.9 feet (*Palaeoloxodon mindriensis*), another a height of 1.5 metres, or 4.9 feet (*P. melitensis*), and the third was less than one metre, or about three feet (*P. falconeri*). *P. creticus* of Crete was 1.5 metres, and *P. cypriotis* of Cyprus was 0.9 metre (three feet) in height. In North America, a small *Mammuthus* isolated on Santa Rosa Island, off the coast of southern California, was probably derived from *Mammuthus meridionalis,* a species that stood 4.2 metres (13.8 feet) at the shoulders.

Early human records of elephants

Elephas maximus asurus lived in Iran and Syria. Early drawings of the animal and fragmentary skeletal remains indicate that it was the largest subspecies of the Asian elephant. The war elephants employed by Pyrrhus in 255 BC and engraved upon Roman seals show animals of unusual size. "Sarus," which signified "the Syrian," was the outstanding animal in the elephant battle squadron of the Carthaginian general Hannibal. In 1500 BC elephants (*Elephas maximus rubridens*) existed in China as far north as Anyang, in northern Honan Province. Writings from the 14th century state that elephants were still to be found in Kwangsi Province.

Man as well as other environmental factors exterminated the woolly mammoth (*Mammuthus primigenius*) and the imperial mammoth (*M. imperator*) about 10,000 years ago. Several races of the living species of Asian and African elephants also died out by about 1500 BC. The small North African race became extinct by the 2nd century AD, and some of the American mastodons, such as *Cuvieronius postremus* of South America, died out as recently as the 4th century AD. The large African bush elephants (*Loxodonta africana*) were exterminated from the Transvaal in South Africa early in the 20th century, but they still occur over much of the continent south of the Sahara Desert. A smaller elephant inhabits the forests of western equatorial Africa, particularly in the Congo region. It is considered by some to be a subspecies (*Loxodonta africana cyclotis*) of the African elephant; others believe it to be represented by several subspecies; still others consider it to be a separate species (*L. cyclotis*).

In Asia, elephants have been exterminated from Iran, Iraq, Afghanistan, the northwestern part of India, and from much of the Malay Peninsula, Java, and the greater part of Borneo and Sri Lanka (formerly Ceylon). Isolated colonies remain in forest areas of Mysore in the peninsular part of India, in Assam, Nepal, Burma, Malaya, southern China, Sri Lanka, Sumatra, Borneo, and other islands of the East Indies.

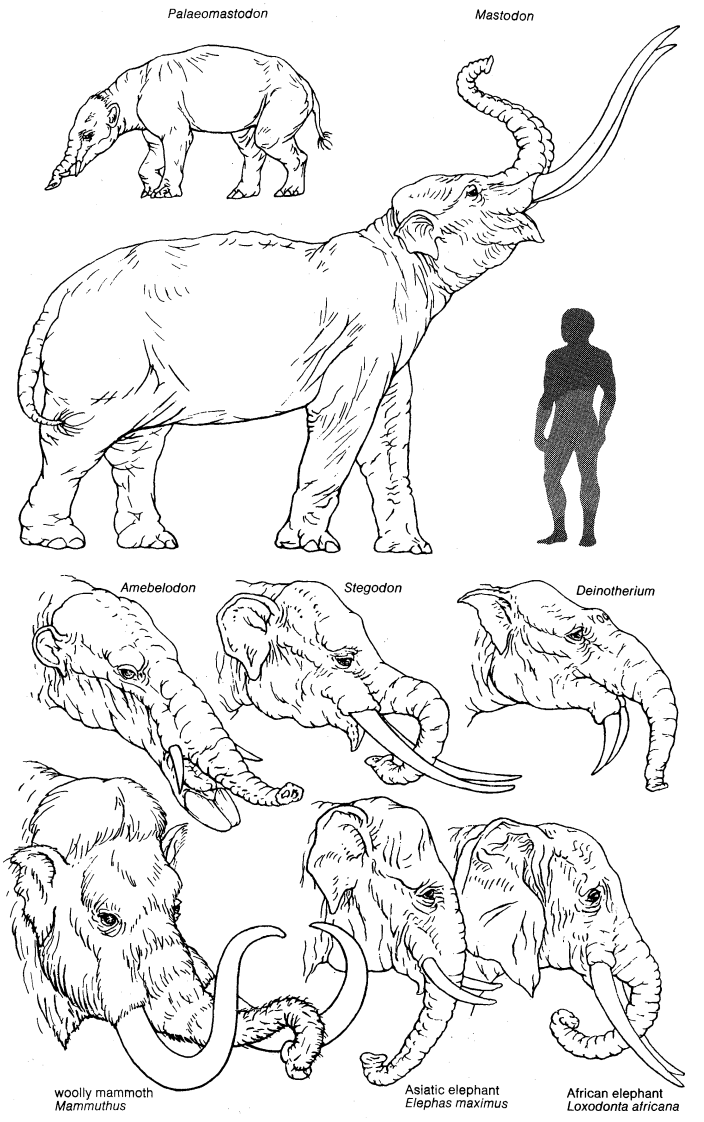


Figure 44: Representative fossil and living proboscideans.
Drawing by Christian D. Olsen

Importance to man. Elephants constitute the chief source of commercial ivory. Because of the continuing demand for this commodity, the animals are in danger of extermination. Elephant "pearls" consist of concentric layers of ivory deposited over a foreign object that has been intruded into the soft ivory at the base of the growing tusk. From early times elephants have been used as beasts of burden in India and Burma. Since they do not breed freely in captivity, new stock for domestication is often captured from wild herds. One method is to drive them through a funnel-shaped stockade into a small enclosure; trained tame elephants help to subdue, noose, and train young captives for service. The training and handling of an elephant are usually entrusted to one man, called the mahout in India and the *oozie* in Burma; an elephant and its keeper frequently became inseparable companions.

Capture of elephants

Trained elephants carry humans in a howdah, or miniature hunting lodge, on their backs. They are used to move timber or other heavy materials.

In modern times, African elephants have been trained for labour only since late in the 19th century; however, they were used by the Carthaginians in wars with the Romans. Both species were tamed at that time. Elephants were used as executioners, in amphitheatres, and for military pageants. They are still used in exhibitions at circuses, carnivals, and zoological parks. In warfare elephants have

been used to drag heavy equipment, especially through mud and up steep slopes. As late as World War II, they were of value in military movements over the mud-clogged roads of Southeast Asia.

The association between man and elephants goes far back into mythology, and a rich folklore has developed. Bracelets of hairs from the tail of a freshly-killed or living elephant are prized as good-luck charms.

Attempts to locate the legendary "graveyards" to which old elephants allegedly resort when near death have been, for the most part, unsuccessful. The groups of buried elephant remains that have been found probably represent sites where elephants drowned in bogs or quicksands.

NATURAL HISTORY

Reproduction and life cycle. Tuskers, or tusked bulls, occasionally fight brief, savage duels that may end in death for the defeated animal. A duel between tuskless elephants may last for days, with occasional periods of rest. The female selected from the herd by the winner often makes an apparent attempt to escape from him. After a brief preliminary courting, the male mounts the female from behind, leaning over her back and either gripping her body or resting his forefeet upon her pelvis, and assumes a standing posture. Copulation lasts for about 20 seconds, with very little movement or noise. Mating continues promiscuously for about two days, after which the most powerful bull drives off the others and remains with the cow for about three weeks. The period of gestation varies from 20 months for a female calf to 22 months for a male. When parturition is about to occur, the herd surrounds the cow, who assumes a squatting position while giving birth.

In regions where large carnivores, such as tigers, prey upon newborn elephants, the cow seeks a female associate. The mother and the other elephants in the herd blow dust upon the moist, newborn calf to dry it. Two hours after birth, the baby is able to stand and is suckled. The mother and calf then join the herd.

Tame cows begin breeding at the age of eight or nine years; tame bulls begin when about 11 or 12 years of age. The interval between the birth of successive calves is about four years. In captivity cows are known to continue bearing calves until 60 or 70 years of age.

Appearance of elephant calf

The newborn elephant is about one metre (three feet) high and weighs about 90 kilograms (200 pounds). It is covered with yellow and brown hair. After a few months the hair on some parts of the body is as long as that in the extinct mammoth. In *E. maximus* there is also a pinkish patch around each eye, and when the calf is about five months old, a faint whitish patch develops on each cheekbone. As this patch spreads, similar patches appear upon the trunk and ears. In the more easterly races of Malaya and Sumatra there are only a few gray spots. The hoof nails, which are dark at first, gradually become lighter. When the elephant is about eight years old, a thick oily secretion known as musth, or must, begins to flow from a gland in the temporal, or temple, region. It occurs in both sexes, but is more active in males. The secretion increases each year until it drips into the elephant's mouth. Some authorities believe that the function of the secretion is to inhibit feeding; others believe it has some effect on sexual activity.

An elephant is not fully grown until it is about 25 years old. In the wild, the average life-span is about 60 years, but under optimum conditions an elephant may live for 80 years.

Behaviour. The organization of an elephant herd is often according to age and sex. In *Elephas,* although herds of 10 or 15 females and their young may appear to be under the leadership of a large female, and their organization matriarchal, young adult males are always in the vicinity, as is the real leader of the entire group, a large male. The leader may be accompanied by one young adult male who acts as a scout, warning the leader of danger. The herd also has a system of scouts, and, before emerging into an open area, one of the scouts usually explores it. If no danger is apparent, he signals by trumpeting to the herd to advance. Individuals often serve as guards while the rest of the herd feeds or bathes.

The herd is held together both by blood relationship and by a strong sense of companionship. If an individual is injured, three or four others surround it, shielding it from danger, supporting it, and helping it to move away. A calf that has lost its mother is adopted by the other cows in the herd even if they have their own calves to raise.

Ecology. Elephants clear paths through forests that are too dense for other animals. Many modern roads in elephant-inhabited countries originated in this manner. Elephants browse to a height of about five metres (16 feet), thereby increasing the amount of sunlight available for shrubs. Their uprooting of grass and roots aerates the soil and stimulates the growth of plants that replace the ones devoured. Mud wallows frequented by elephants are fertilized by their excreta.

Feeding habits

An elephant may destroy or discard as much vegetation as it consumes. An adult may eat between 250 to 350 kilograms (550 to 770 pounds) of solid food each day. When grazing, the animal uses its trunk or forefoot to gather grass, which is slapped against a forelimb to rid it of sand. In rainy weather, when soil is more difficult to shake off, the animal browses. Asian elephants break off branches; African elephants are more likely to push over trees. The wood apple (*Feronia elephantorum*) is a favourite food of the Asian elephant. The animal also eats wild rice that grows in forest lakes and various other aquatic plants. The African bush elephant eats the fruit of various palm trees. The spongy wood of the baobab tree provides some water during periods of drought.

FORM AND FUNCTION

Extant forms. The adult elephant has a tuft of hair at the tip of the tail and sometimes a patch of hair on the head. The limbs are adapted for bearing great weight: the legs are straight and pillar-like, and the bones of the joints are flat at the articular surfaces. Each toe has a heavy hoof nail; the weight is borne on thick pads behind the toes. The nose and upper lip are extended into a long trunk, which contains the nasal passages and has nostrils at the tip. Water for drinking is sucked into the trunk and then discharged directly into the mouth. The trunk is used for placing food into the mouth, for spraying and dusting the body, for lifting or moving heavy objects, and even for throwing objects at man.

The upper second incisors are typically developed into ivory tusks, the longest and heaviest teeth of any living animal. Canine teeth are absent. The complex molars are of the high-crowned type, with transverse rows of enamel ridges on the grinding surface, which is often traversed by a longitudinal median groove. There is normally only one complete functional tooth and half of a second one at a time on each side of each jaw. These are replaced horizontally from the rear as they wear away.

The Asian elephant (*Elephas maximus*) and its races are distinguished from the African elephants by being somewhat smaller and by having relatively small ears with the upper edge curled forward. The head is more domed, is structurally more complex, and has a greater development of diploe, or bony cell cavities. *E. maximus* also has high-crowned teeth and a relatively smooth trunk with a single fingerlike projection at the tip. Adult bulls weigh up to 5,450 kilograms (six tons) and commonly stand three metres (10 feet) at the withers. Only the males develop tusks, which average 1.5 metres (4.9 feet) in length, the pair weighing about 32 kilograms (70 pounds). These develop flat, longitudinal planes of wear near the tip. Tusks 2.7 metres (8.9 feet) long and 68 kilograms (150 pounds) in weight have been recorded. About 90 percent of the males of the Ceylonese race lack tusks; Sumatran elephants are of slighter build and have longer trunks.

In contrast to the Asian species, the African bush elephant is generally larger. This and the forest form have extremely large ears (one metre in breadth), with the upper edge curled backward, and a roughened, heavily ringed trunk with two projections at the tip. The molars are of coarser construction, have fewer ridges, are less crenulated (scalloped), and consist of a thicker surface of enamel over thick plates of dentine. The top of the head is not domed, and the forehead is more convex.

The tusk tips are usually conical, the legs are longer, and the eyes are relatively larger than those of the Asian elephant.

The average height of adult bull bush elephants is 3.3 metres (10.8 feet) at the shoulder, and the average weight is six tons. Cows are about 0.6 metre (two feet) shorter. Both sexes possess tusks, which average about 1.8 metres (5.9 feet) long, the pair weighing 36 to 55 kilograms (79 to 121 pounds). A pair of tusks in the British Museum weighs about 133 kilograms (293 pounds); the larger one of the two is 3.5 metres (11.5 feet) long, with a basal circumference of 46 centimetres (18 inches). The largest elephant on record, a bull bush elephant killed in the Cuando river district of southeastern Angola in 1955, is on display at the Smithsonian Institution in Washington, D.C.; it probably weighed 8,200 kilograms (nine tons) when alive and stood four metres (12 feet) at the shoulder. Adult forest elephants are about 2.1 metres (6.9 feet) tall and weigh 1,225 kilograms (2,700 pounds), with slender tusks that are often more than three metres long. The ears are relatively small and smoothly rounded at the margins.

Albinos

Albino, or "white," elephants occasionally occur, especially in Thailand and Burma, where they are regarded by some as semisacred.

The mammoth. The genus *Mammuthus* contains some of the largest members of the family Elephantidae. Certain ones reached a height of over 4.2 metres (13.8 feet) at the shoulders. One was *M. meridionalis* of Asia and Europe, and the other was *M. imperator,* which entered North America during the upper Pleistocene.

The genus also contains one of the most specialized members of the family, the woolly mammoth, *M. primigenius,* which probably became extinct about 10,000 years ago. It inhabited the sub-Arctic area of Asia and Europe and eventually entered North America over the Bering Strait; it travelled southward across western North America almost to Wyoming, then spread eastward toward Lake Michigan. Its height of 3.3 metres (10.8 feet) at the shoulders equalled that of a large *Elephas maximus,* but the body was relatively shorter, and the hindquarters sloped downward. Its skull was compressed from front to back. As an adaptation to its cold environment, the woolly mammoth evolved small ears, a short goatlike tail, and a coat of dense, furry, short hair overlain by longer, bristly hair. It also had a humplike reserve of fat upon the top of

Drawing by Christian D. Olsen; adapted from A.S. Romer, *Vertebrate Paleontology,* © 1933, 1945, 1966 by the University of Chicago; all rights reserved

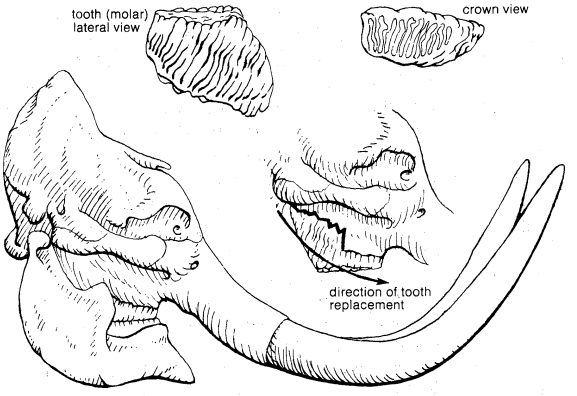


Figure 45: Skull and tooth movement in the woolly mammoth (*Mammuthus primigenius*).

the head and on the shoulders. A subcutaneous layer of fat about eight centimetres (3.1 inches) thick covered the body. The molars had as many as 27 lamellae, or plates. The tusks of the male were about 4.8 metres (15.7 feet) long. Their almost circular curvature and great size suggest that they may have functioned as shovels for exposing vegetation buried under snow. The mammoth is one of the few extinct proboscideans in which the carcass has often been completely preserved. Among the best known are two from Siberia—one discovered near the mouth of the Lena River in 1804; the other in the bank of the Beresovka River in 1900.

EVOLUTION AND CLASSIFICATION

Historical development and paleontology. The order Proboscidea has evolved from unknown ancestors that were not much larger than pigs. They flourished during the Paleocene Epoch (65,000,000 to 54,000,000 years ago). During the course of evolution, the lower jaw elongated beyond the upper, and the tusks projected well beyond the upper ones. At this stage the nose, palate, and upper lips developed into an elongated fleshy cover to the projecting lower jaw. It is probable that the nostrils opened well above the extremity of this flap and were near the eyes. A longer lower jaw proved less useful than a shorter one, so the upper flap was converted into a multipurpose tubular proboscis. Because the nostrils shifted to the tip of the proboscis, the animal was able to breathe while submerged in water. When so submerged, the animal had to rely on its sense of smell more than sight to detect approaching predators.

Origin of the proboscis

The suborder Deinotherioidea, consisting of one genus, is an early branch of the main proboscidean stock of the Eocene Epoch (54,000,000 to 38,000,000 years ago). They lost their upper tusks and developed a downward-hooked, tusk-tipped mandible. Numerous species of the suborder occurred in Asia, Europe, and Africa and persisted into Pleistocene times (2,500,000 to 10,000 years ago). The largest was the European *Deinotherium gigantissimum,* which reached a height of 3.8 metres (12.5 feet) at the shoulders and existed during the Pliocene Epoch (7,000,000 to 2,500,000 years ago).

In the suborder Mastodontoidea, the family Gomphotheriidae comprises 15 genera, including the earliest members of the order, *Phiomia* and *Palaeomastodon.* The former were the size of donkeys, but the latter were as large as a modern Asian cow elephant. In this family the skull and neck are elongated, and the teeth low crowned. The second incisors are enlarged; the upper ones are compressed and vertical, and they retain a band of enamel. In the later evolved genera, the lower pair are bent forward, depressed, and expanded into shovellike structures that do not meet the upper tusks. The canines are absent. Among this family are *Cordillerion* of North America and *Cuvieronius* of South America. The latter became extinct as recently as AD 200 to 400.

Phiomia, which occurred in Egypt and India, was an archaic shovel-tusked form with an elongated neck. It flourished during the Oligocene (38,000,000 to 26,000,000 years ago). The mandible and its tusks became more shovellike in *Amebelodon* and *Platybelodon* of the Miocene (26,000,000 to 7,000,000 years ago).

In *Palaeomastodon* the skull shortened, and the tusks assumed a cylindrical shape. The genus occurs in middle Oligocene deposits of Egypt. In some later genera, such as *Anancus* and *Stegomastodon,* the jaws shortened and the lower tusks disappeared. In some shovel-tuskers, the mandible remained enlarged and continued to function as a shovel for digging plant bulbs. It was probably protected by a horny pad.

Mastodontidae contains the single genus *Mastodon.* It lacked lower tusks. The tusks were about two metres (seven feet) long. Some species attained a height of about three metres (10 feet), and were covered with hair. Parts of carcasses of the recently extinct American species *Mastodon americanus* have been discovered in peat deposits and swamps.

The earliest proboscideans in Asia are *Phiomia, Gomphotherium, Platybelodon,* and *Serridentinus,* of the family Gomphotheriidae, and *Stegolophodon,* of the family Elephantidae. All occurred in the Miocene of China and some in Burma and India.

In the suborder Elephantoidea, *Stegolophodon* is intermediate between the mastodons and elephants proper. Although this genus first appeared during the Miocene of Europe, Asia, and Africa, it persisted into the upper Pleistocene of these continents.

Stegodon, which occurred during the Pliocene in Asia and during the Pleistocene in Africa and Asia, evolved from *Stegolophodon.* The skull enlarged, the jaws shortened, and the lower tusks disappeared. *Stegodon zhaolongensis* of China was a large species with the most primitive teeth

in the genus. The tusks of some species such as *Stegodon magnidens* of India were about 3.3 metres (10.8 feet) in length; in others, however, they were feeble. The molars were usually low crowned, and the depression between each pair of dental folds or plates was Y-shaped, rather than V-shaped or U-shaped as in other elephants.

The genus *Mammuthus* includes all species formerly placed under *Archidiskodon, Metarchidiskodon, Parelephas,* and *Stegoloxodon.* Its species occurred in the Pleistocene of Asia, Europe, America, and Africa. Several species had great bulk and heavy tusk development.

Development of pygmy forms

On various occasions during the Pleistocene Epoch, normal-sized elephants that inhabited a continent were isolated when a part of the landmass was separated into islands by the submergence of low-lying land. As these isolated colonies of elephants multiplied, their fodder supply decreased, and the animals gradually became smaller—an unsuccessful measure against extinction. This process is evident in the East Indies and Philippines, in the islands of the Mediterranean Sea, and in certain islands off the coast of southern California.

Classification. *Distinguishing taxonomic features.* The proboscideans are classified largely according to body size; shape of the skull; dentition; and the shape, size, and degree of reduction of enamel in the tusks. Groups marked with a dagger (†) are extinct and known only from fossils.

Annotated classification.

ORDER PROBOSCIDEA

Oligocene to present; North America, Eurasia, and Africa. Heavy-bodied (graviportal) animals with snout prolonged into a fleshy proboscis (trunk). Tusks developed from upper or lower incisors or both; canines absent; cheek teeth with transverse rows of blunt cones or ridges. Skull short, high; nasal openings at a higher horizontal plane than eyes. Body size medium to large; shoulder height from about 1 m (about 3 ft) to more than 4 m (13 ft). About 300 species, all extinct but two.

†Suborder Deinotherioidea

†*Family Deinotheriidae*

Lower Miocene to upper Pleistocene; Europe, Asia, Africa. Upper tusks lacking; lower tusks curving downward from tip of lower jaw. One genus (*Deinotherium*); many species; height to about 3 m (10 ft).

†Suborder Mastodontoidea

†*Family Gomphotheriidae*

Lower Oligocene to Recent (AD 200–400); Europe, Asia, North and South America. Skull and neck elongated. Teeth low-crowned; succession vertical. Later genera with protruding, shovellike lower incisors, others with substantial upper tusks and no lowers; premolars with 2 transverse crests, molars with 3. About 15 genera, several dozen species; shoulder height about 1 to 3 m (3 to 10 ft).

†*Family Mastodontidae* (mastodons)

Lower Miocene to upper Pleistocene (possibly to early historic times); Europe, Asia, Africa, North America. Molars with rounded prominences, but no ridges; lower tusks absent, but upper incisors substantial, reaching over 2 m (6.6 ft) in length in males of some species. One genus (*Mastodon*), many species; shoulder height to at least 3 m (10 ft).

Suborder Elephantoidea

Family Elephantidae (elephants and mammoths)

Lower Miocene to present; fossils from Europe, Asia, East Indies, Africa, and North America; Recent species from Africa (*Loxodonta*) and southern Asia (*Elephas*) which have probably evolved from *Stegolophodon.* Six genera, with several dozen fossil and 2 Recent species; shoulder height from 1 to about 3.5 m (3 to 11.5 ft). The epiphyses (ends of long bones in limbs) do not fuse until the last molars appear. The molars are replaced at least three times. Marrow disappears from the limb bones early in adulthood.

(P.E.P.D.)

Sirenia (dugong, manatees)

Sirenians are a group of large aquatic mammals that have become rare or extinct as a result of exploitation by man for meat and oil. The largest, Steller's sea cow (*Hydrodamalis gigas*), which reached lengths of about eight metres (about 26 feet), was eaten out of existence by hungry seal hunters within a few decades of its discovery in 1741 in the Bering Sea. The remaining forms, to which the common name sea cow is also sometimes applied, are the dugong (*Dugong dugon*) and manatees (three species of *Trichechus*), which, if their stocks were allowed to rebuild themselves, could again become of economic importance. Being the only large aquatic herbivores, other than some turtles, they could provide meat from the vast expanses of marine and freshwater vegetation, at present quite unused by man, and so bring another marginal area into production.

Extinction by hunting

Dugongs and manatees are both usually seen up to lengths of three to four metres (10 to 13 feet), but larger specimens, up to six metres (20 feet), have been mentioned in tales of early travellers.

The dugong seems to have had a wider distribution in the past, but in recent times it has been restricted to the warm coastal waters of the Indo-Pacific region. Throughout most of its range it is now much depleted in numbers but exists from the Solomon Islands in the east to the head of the Red Sea in the west and from the Philippine Islands and the Persian Gulf in the north to Brisbane, Perth, and Mozambique in the south. It is entirely marine and rarely even enters estuaries. Manatees, however, seem more adaptable and inhabit the coastal, estuarine, and riverine waters of the tropical and subtropical Atlantic region: *T. senegalensis* from West Africa; *T. manatus,* with two subspecies, from the Caribbean; and *T. inunguis,* landlocked in the Amazon Basin.

Natural history. Sirenians tend to live in groups, and dugongs particularly may be highly gregarious at times. An enormous herd, three and a half by one and a half miles in extent, was recorded off Brisbane, Australia, early in the last century. Both genera are totally aquatic, living almost continuously submerged, and this, coupled with their wary and surreptitious habits, makes them exceptionally difficult to study. They are unable to come out of the water, but manatees placed on land nearby may just manage to return to their natural environment by vigorous wriggling of the body. Breathing is by frequent brief visits to the surface, where the tip of the snout is protruded and the nostrils are silently opened by muscular valves. A manatee has been known to last as much as 16½ minutes between taking breaths. Manatees may also be seen lying at the surface, usually with only their backs visible above the water. Even when basking thus, they are easily frightened by sound or disturbance and will silently submerge. On occasion both dugongs and manatees will thrust their heads entirely out of the water, and they sometimes may even disport themselves like cumbersome seals.

Aquatic behaviour

Extremely little is known of the breeding biology of sirenians and particularly of the time scale involved. Manatees are known to be long-lived, and dugongs are believed to have similar lifespans. But there is no information about age of maturity, rate of growth, or number of young produced in a lifetime. Mating in manatees has been observed with the animals in very shallow water lying on their sides. Gestation is known to occupy more than 152 days, and the normal birth is a single calf, which receives much maternal care. Very occasionally twin fetuses are found in dugongs. Suckling from the single pair of pectoral mammary glands normally occurs in a horizontal position, but the occasional suckling of a young one held vertically by a flipper may have given rise to the mermaid myth. There is no marked difference in size between the sexes.

Sirenians are totally herbivorous. Whereas Steller's sea cow fed on marine algae, the dugong and manatees feed entirely on green higher plants, a circumstance that limits their distribution to relatively shallow waters. The habit of the dugong is to feed where there is good growth of "dugong grass" (*Zostera,* etc.), the leaves and underwater stems of which are pulled off by the animal's powerful lips. Manatees seem prepared to feed on marine or freshwater plants whether growing on the bottom, floating at the surface, or even growing on banks of rivers up to a foot above water level. It is this catholicity that has led to the experimental use of manatees in Guyana to keep clear irrigation and transport channels in cane fields that otherwise must be cleared by hand. Virtually the sole enemy of both the dugong and manatees is man. There does not seem to be any commercial exploitation of these

animals at present, but both are hunted sporadically and locally for food with the use of nets or harpoons. The dugong still has a considerable residual stock off northern Australia and with protective legislation enforced, could have a valuable future. Manatees, because they come into more confined waters, are still in danger of extermination.

Form and function. Sirenians have torpedo-shaped bodies and, like whales, have lost all external trace of hindlimbs. Their tails likewise are flattened horizontally and provide the main propulsive force in swimming. The forelimbs are small and may help in turning and manoeuvring. Though usually somewhat sluggish, sirenians are capable of considerable speeds for short distances.

Distinguishing features

The main distinguishing characteristics between the dugong and the manatees are that the former have a downturned snout and a forked tail, while the latter have a straight snout and rounded tail. Both have immensely thick, tough skin, which is nearly hairless. In the dugong there are individual hairs about six millimetres (0.24 inch) long, spaced five to seven centimetres (two or three inches) apart. There is no substantial layer of blubber beneath the skin, but the body contains much fat. The eyes are small and circular, without lids, and the minute external ear openings are to be found only by careful examination. All of the bones are of exceptional density. Sirenians, particularly the dugong, have a much enlarged and intensely muscular upper lip, the corners of which serve to pluck the vegetation on which the animal feeds. In the dugong there are remains of two incisors, the second of which becomes large in the adult, remaining unerupted in the female and erupting to form a tusk in the male. There are basically six cusped molars in each jaw, which fall out progressively from the front so that only two remain in the adult. These teeth, together with horny pads at the front of each jaw, crush the food. In manatees each jaw has a series of 20 to 30 crushing molars, which move progressively forward during the life of the animal. Manatees are also exceptional in having only six vertebrae in the neck, instead of the seven standard in mammals. Almost nothing is known about the physiological processes of these animals.

Drawing by J. Helmer based on (manatee) photograph courtesy of Field Museum of Natural History, Chicago; (dugong) from G.M. Allen, *Extinct and Vanishing Mammals of the Western Hemisphere*; Cooper Square Publishers

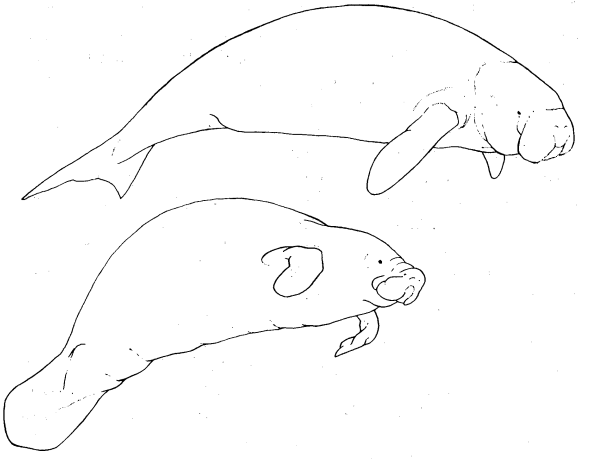


Figure 46: *Body plans of two major sirenian groups.* (Top) dugong (*Dugong dugon*); (bottom) manatee (*Trichechus manatus*).

Classification and paleontology. Sirenians comprise the order Sirenia of the superorder Subungulata, which also includes elephants and hyraxes. Two points of resemblance with elephants are the mode of tooth succession and the pectoral position of the mammary glands. The extensive fossil record indicates that the Sirenia date from the Eocene, about 50,000,000 years ago. More than 20 genera of fossil sirenians have been described. They seem to have been widespread near the coasts of the warmer seas of the world, and some think that the manatee is of more recent origin and has replaced the dugong in the Atlantic region.

(G.C.L.B./C.K.B.)

Perissodactyla (horses, asses, zebras, tapirs, rhinoceroses)

The order Perissodactyla is a group of herbivorous mammals characterized by the possession of either one or three hoofed toes on each hindfoot. The name (Greek *perissos,* "odd," and *daktylos,* "finger") was introduced to separate the odd-toed ungulates from the even-toed ones (Artiodactyla), all of which had previously been classified as members of a single group.

GENERAL FEATURES

The Perissodactyla comprise three families of living mammals: six species of horses (Equidae), four species of tapirs (Tapiridae), and five species of rhinoceroses (Rhinocerotidae). These families are remnants of a group that flourished during the Tertiary Period (from 65,000,000 to 2,500,000 years ago), a time when it was much richer in species and in variety of form than at present and played a dominant role in the fauna of the world.

The horses, asses, and zebras are long-legged, running forms with one functional digit in each foot and with high-crowned, molariform (*i.e.,* modified for grinding) cheek teeth. The tapir is a rather rounded, piglike, semiamphibious forest and woodland animal with a small proboscis (trunklike snout) and a coat of short, bristly hairs. Tapirs have primitive features, such as four hoofed toes in the forefoot and three in the hind, and they have rather simple molar teeth. Rhinoceroses are massive, graviportal (ponderous) creatures with a thick and nearly hairless hide and three digits on each foot. They bear hornlike structures on the head.

The Perissodactyla are of particular scientific interest because their fossil history is so well-known. The evolution of horses from the tiny "dawn horse" (*Hyracotherium,* formerly *Eohippus*) to the present form is a classic sequence, knowledge of which has played an important role in evolutionary thought. The order also provides a notable example of parallel evolution. Following completely different evolutionary paths, both perissodactyls and artiodactyls (*e.g.,* cattle, antelope, swine) independently evolved features such as high-crowned grinding teeth and elongated limbs with a reduced number of digits, in adaptation to a similar running (cursorial), herbivorous mode of life.

Indian and square-lipped rhinoceroses

Living perissodactyls are of medium or large size. Asses and tapirs, the smallest representatives of the order, attain a length of approximately two to 2.5 metres (6.6 to 8.2 feet), stand one metre or more at the shoulder, and weigh up to 250 or 300 kilograms (550 to 660 pounds). The largest forms are the Indian and square-lipped rhinoceroses (*Rhinoceros unicornis* and *Ceratotherium simum,* respectively), which are four to five metres (13 to 16.4 feet) long and measure up to two metres at the shoulder. The maximum weight has not been well established, but a figure of 2,070 kilograms (4,550 pounds) has been recorded. *Baluchitherium,* relative of the rhinoceroses known from the Oligocene (about 30,000,000 years ago), was the largest known land mammal, standing about 5.5 metres (18 feet) at the shoulder.

Ecologically, the Perissodactyla were the dominant large herbivore group during the Tertiary, a position now held by the Artiodactyla. All feed either by grazing (*i.e.,* cropping grasses) or by browsing (taking shoots and leaves from trees and bushes). The Equidae in particular were abundant and important members of the Old World fauna until their numbers were reduced by modern man. Zebras are still numerous and ecologically important in a few parts of Africa. The importance of the domestic horse and the ass in the history of mankind is very great indeed. Both have served extensively as pack, draft, and riding animals. The horse is sometimes eaten by man, and its flesh is widely used as pet food; through centuries of domestication, it has been developed into a number of different breeds (for more information on domesticated horses, see HORSES AND HORSEMANSHIP).

The living wild Equidae are confined to the Old World. Zebras and the true wild ass (*Equus asinus*) are African, with the zebras confined to the southern and eastern parts,

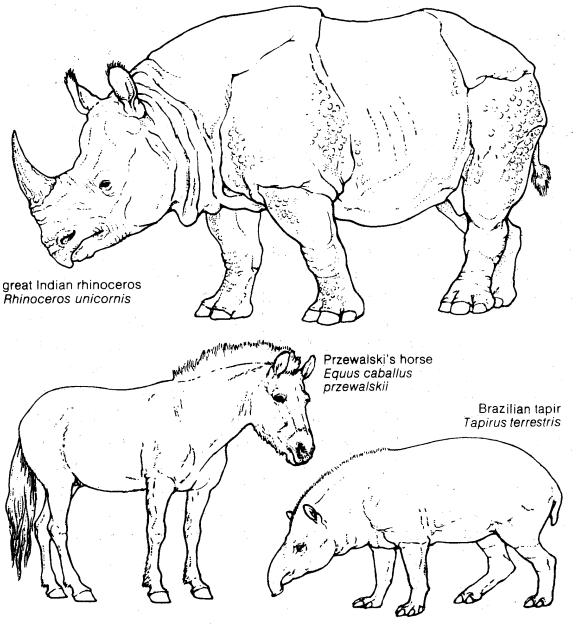


Figure 47: Major body plans among perissodactyls.
Drawing by R. Keane

while the ass originally ranged over northern and northeastern Africa.

The wild horse (*Equus caballus*), ancestor of the domestic horse, occupied the low country north of the great mountain ranges from Europe across central Asia; it may now be extinct as a wild animal. The half-asses, races of *E. hemionus,* were found in the arid zone of Asia from Persia to the Gobi Desert, as well as in Arabia, Syria, and northwestern India.

The living rhinoceroses are also Old World forms, with two species in Africa and three in Asia. There are three species of tapirs in the New World tropics, one in Middle America and two in South America. The fourth species of tapir is Asiatic.

NATURAL HISTORY

Distribution, ecology, and conservation. The Equidae are highly specialized for a cursorial, herbivorous mode of life. They are absent from forests and other densely vegetated regions, but apart from this limitation, the range of the group is relatively unrestricted by the type of vegetation, climate, and topography. The species replace one another geographically for the most part, and each occupies a somewhat different habitat.

Dietary needs as range factors

Grass plays a major role in the diet; the zebras, for example, are known to feed on tall, coarse grasses avoided by most antelopes. Some species also take shrubs, herbs, and even bulbs. Water requirements vary in different species. In South Africa, Burchell's zebra has been found to drink about once every 36 hours. By contrast the mountain zebra (*Equus zebra*), Przewalski's horse (*Equus caballus przewalskii*) and the half-ass, all living in semidesert areas, are reported to survive if they can drink once in three or four days. The ass, too, can manage with less water than the horse. The mountain zebra and Przewalski's horse dig for water in dry riverbeds and depressions.

Zebras. The mountain zebra occupies parts of the arid rocky escarpment separating the interior plateau of the southern African subcontinent from the coast lowlands. The race *E. z. hartmannae,* still relatively numerous over much of its original range in South West Africa and southern Angola, enjoys legal protection and is represented in game reserves. Conflict between these zebras and farm livestock for the meagre pasture of the semidesert regions, however, has led to a reduction in zebra numbers. The race *E. z. zebra* was originally common in the mountain ranges of the Cape Province, but now survives only as a remnant of perhaps 100 animals. About one-half of these have sanctuary in a national park.

Burchell's zebra (*E. quagga*) formerly inhabited a great area of grassland and savanna from the Cape to the southern Sudan. The southernmost race (*E. q. quagga*), which was only partly striped, became extinct in the 19th century. The populations of the other races have been much reduced in many places and the range of the species has shrunk considerably. There are large populations in reserves, however, and the species is not in any immediate danger of extermination. Grevy's zebra (*E. grevyi*), which shares a narrow zone in northern Kenya with Burchell's zebra, is confined to sparsely wooded, semidesert plains and low hills in northern Kenya, southern and eastern Ethiopia, and western Somaliland. Its status appears to be generally satisfactory.

Asses. The true ass (*Equus asinus*), ancestor of the domestic donkey, is the equid of arid North Africa whose range extends south to approximately 6° N latitude. Its natural distribution probably included all habitable parts of North Africa. At present, asses are known from semidesert country extending from the east bank of the Nile (in the Sudan) to the Red Sea, and in parts of Eritrea, Ethiopia, and Somalia. There are also isolated pockets in the Tibesti Mountains in the Sahara, and in the countries of central and western Africa. There is a great deal of uncertainty about the identity of all asses now described as "wild." Some may be merely feral (escaped or released) donkeys, and interbreeding with feral donkeys is likely to have occurred in many, if not all, existing populations.

The two surviving races of wild horses

The wild horse. The wild horse was widely distributed in Eurasia north of the mountain chains. The Romans encountered it in Spain. Two races have survived to modern times. A gray race, known as the tarpan, was the horse of southern Russia. It became extinct in the Ukraine during the mid-19th century. The endangered Przewalski's horse (*E. c. przewalskii*), a small, reddish-brown race (considered a species by some authors), was last seen in the wild in 1968 in the remote, semidesert steppe country on the boundary between Mongolia and China. Wild horses enjoy legal protection in Mongolia and China, but nomadic pastoralists have been encroaching on previously uninhabited country and competing with the horses for pasture and the scarce water supplies.

The half-asses, races of *Equus hemionus,* occupied the dry belt from Mongolia through Central Asia to Syria, with a northern limit at about 50° N latitude. The chigetia or kulan (*E. h. hemionus*), which was formerly widespread over an immense region of the Gobi, now occurs only in semidesert steppe country in central Mongolia. Hunting and competition for water by pastoral tribesmen are responsible for its decline. The kulan is slightly smaller than the kiang (*E. h. kiang*), which is found on the cold, arid steppes of Nepal, Sikkim, and western Tibet at altitudes of 4,270 metres (14,000 feet) and more. The kiang is now said to be rare but not endangered. The Persian onager (*E. h. onager*) lives in a lower semidesert or desert environment, with a range that formerly included northeastern Iran, northwestern Afghanistan, and Russian Turkestan. It is now extremely rare and unlikely to survive outside northeastern Iran and the Badkhyz Reserve in Turkmeniya. A small nucleus has sanctuary in the semidesert salt plains of the recently established Kavir Protected Region in Iran. The Indian wild ass is a closely related, probably identical, form sometimes distinguished as the race *E. h. khur.* A fairly small population occupies salt flats in the Rann of Kutch, a remnant of the thousands found there at the end of World War II. The Syrian onager (*E. hemionus hemippus*) is the smallest member of the group and stands about one metre (three feet) at the shoulder. It was once found in the desert region of Palestine, Syria, and Iraq, and was domesticated by the ancient Sumerians before the introduction of the domestic horse into Mesopotamia. This race may survive in the Djezireh Desert, Syria, or north of the Syrian–Turkish border; if so, the number must be extremely small.

"White" and "black" rhinos of Africa

Rhinoceroses. The two African species of rhinoceros are the black or prehensile-lipped rhinoceros (*Diceros bicornis*) and the white or square-lipped rhinoceros (*Ceratotherium simum*). The terms black and white are misleading, since both species are grayish to brownish, but the names are well established in common usage.

The black rhinoceros was originally widespread from the Cape to southwestern Angola and throughout eastern Africa as far as Somalia, parts of Ethiopia and The Sudan. Its range also extended westward through the northern savanna zone to Lake Chad, the northern Cameroons, Volta, and the northern Ivory Coast. The animal was extremely numerous in some parts. It now occupies a much smaller area, within which it is found in scattered pockets, many of them in parks and reserves. South Africa has an important population of some 300 of them in its Zululand reserves. The species still occurs in northern South West Africa and southwestern Angola, Rhodesia, Mozambique, Malawi, Zambia, and Zaire. Northern Tanzania and Kenya have more black rhino than any other country but the future of the animals outside the parks and reserves is far from secure. The species is also found in parts of Somalia, Ethiopia, northern Uganda, and The Sudan, and remnants remain in the Central African Republic and northern Cameroon. The decline in numbers is largely the result of expanding human settlement and of poaching to obtain the horns, which fetch high prices.

The black rhinoceros occupies a variety of habitats, frequenting open plains, sparse thorn scrub, savannas, thickets, and dry forests, as well as mountain forests and moorlands at high altitudes. It is a selective browser, and grass plays a minor role in its diet. Succulent plants, such as euphorbias, assume great importance in dry habitats, and where these plants are abundant the animals appear to be able to survive without free water. Where water is available, drinking is regular and frequent; the animals also may dig for water in dry riverbeds.

The much larger white rhinoceros is a grazing species with a broad, square muzzle. It prefers short grasses seven to 10 centimetres (about three to four inches) high. The animal makes much use of shade trees for resting and is dependent on surface water. The range of the white rhinoceros is markedly discontinuous. South of the Zambesi River it was once extremely common over a fairly large area of bushveld. It has since become confined to the game reserves in Zululand, where the population has risen to some 1,700; some of the animals have been redistributed to several other parks and reserves in southern Africa.

A northern race formerly inhabited the southern Sudan and adjacent areas of Uganda and Zaire, extending westward into the Central African Republic. It has also been much reduced, but considerable populations still survive in the Parc National de la Garamba (Zaire) and in the Bahr el Ghazal region of The Sudan, and a small remnant has found sanctuary in the Ruwenzori National Park, Uganda.

The smallest living rhinoceros

The smallest of the three Asian rhinoceroses (also the smallest living member of the family) is the Sumatran, or Asiatic, two-horned rhinoceros, *Didermocerus* (or *Dicerorhinus*) *sumatrensis,* standing one to one and a half metres (three to five feet) at the shoulder. It was originally found from eastern Pakistan and Assam throughout Burma, much of Thailand, Indochina (Cambodia, Laos, and Viet Nam), Malaya, Sumatra, and Borneo. Small isolated populations still occur in a few widely separated localities in Burma, Thailand, Malaya, Sumatra, and Sabah, and possibly in other nearby territories. The total population is thought to number between 100 and 170. Some of the survivors in Sumatra are protected in reserves.

The Javan, or lesser one-horned, rhinoceros (*Rhinoceros sondaicus*) occupied the islands of Java, Borneo, and Sumatra, the Malay Peninsula, and a region extending northwards through Burma into Assam and eastern Bengal. It is now restricted to the Udjung-Kulon Reserve in western Java where there are at least 25 and perhaps as many as 50 to 60 animals.

Both the Sumatran and Javan rhinoceroses inhabit forests as well as marshy areas and regions of thick bush and bamboo, climbing actively in mountainous country. They are mainly browsers.

The great Indian, or one-horned, rhinoceros (*Rhinoceros unicornis*) is more or less equivalent in size to the square-lipped rhinoceros and is distinguishable from the smaller Javan rhinoceros by the presence of a large horn, tubercles on its skin, and a different arrangement of skinfolds. It previously occupied an extensive range across northern India and Nepal from Assam in the east to the Indus Valley in the west. It is found in a range of habitats—open grassland, savanna, forests, and hilly country—and appears to be mainly a grazer, raiding grain fields in some areas. Hunting and the pressure of expanding human populations have greatly reduced both the range and numbers of this animal. It is now found almost entirely in eight reserves or sanctuaries in India, notably the Kaziranga Sanctuary in Assam (estimated population 300) and in the Rapti Valley region of the Nepal Terai. The total population is estimated at about 600 animals, and the prospects for survival appear to be reasonably good.

Tapirs. The Malayan tapir (*Tapirus indicus*), largest member of the family Tapiridae, is found in Sumatra and the Malay Peninsula, as far north as the Burmo-Siamese border in latitude 18° N. It is found from sea level to high altitudes and occupies forests and thickets but may feed in more open areas. It is still abundant and widespread.

Distribution of New World tapirs

The three New World species occupy distinct, nonoverlapping but contiguous ranges. The mountain tapir (*Tapirus pinchaque*), the smallest and most primitive, inhabits the temperate-zone forests and bordering grasslands of the Andes in Colombia and Ecuador and in northern Peru, up to altitudes of nearly 4,600 metres (about 15,000 feet). Agricultural and pastoral expansion have resulted in some decline in the status of this species, but it is still fairly common. The Central American, or Baird's, tapir (*T. bairdii*) is the largest of the American species. It is essentially middle American, with a range extending from Mexico into coastal Ecuador, and it occupies undisturbed climax rain forest. It is shy and adjusts poorly to the disturbance caused by settlement. This disturbance, together with the destruction of habitat accompanying human occupation, has greatly reduced its range and numbers. The species is said to be much in need of active conservation. The most widespread species is the Brazilian tapir (*T. terrestris*), which is found throughout the Brazilian subregion east of the Andes and in a small area west of the Andes in northwestern Venezuela and northern Columbia. Like the other species, it is largely a forest form requiring the proximity of water. The three New World tapirs are mainly browsers and are remarkably similar in habits.

Behaviour. *Expression and communication.* The Equidae communicate by means of calls and changes in facial expression. Six different sounds are made by Burchell's zebra. A whinny, consisting of a series of two- or three-syllabic "ha" sounds, serves to maintain contact between members of a group. The repertoire includes an alarm call ("i-ha"), an alarm snort, a drawn-out snort of satisfaction, and a squeal of pain and fear. Other species utter similar sounds, the whinny of the horse and the bray of the ass being well-known examples. Characteristic facial expressions have been described for greeting ceremonies (mouth open, ears up), threat (mouth open, ears back), and submission (mouth open, nibbling movements, ears down). In all species studied except the horse, females assume a particular expression ("mating face") when permitting the male to mount.

In the rhinoceroses and tapirs snorting, squealing, bellowing, and, in some forms, whistling sounds play a major role in communication. Visual signals are not well developed in these nonsocial animals, but a few facial expressions are used.

Feeding. The Perissodactyla are mainly grazers or browsers. The quality and quantity of grasses available to grazing species may vary considerably with the season and the area. The animals may accordingly move great distances to reach attractive sources of food. Migrations of Burchell's zebras to succulent pastures during the rainy season are a feature of the Serengeti Plains and the Etosha National Park in Africa. The distribution of asses, half-asses, and horses inhabiting arid areas largely follows that of rainfall and pasture.

The food of the browsers is fairly readily available throughout the year; thus, species in this category are relatively sedentary. The browsing rhinoceroses may break down trees and shrubs, and use their forelimbs to help get at otherwise inaccessible leaves and twigs. Food is plucked

with the lips. In the tapirs, the upper lip is fused with the short proboscis. The rhinoceroses (excluding the white) have a pointed upper lip with a fingerlike process that is used to pluck leaves and twigs. The white rhinoceros, with its broad square muzzle, is the most specialized grazing rhinoceros, feeding on grass.

Herd behaviour among equids

Social organization and territory. In both the mountain and Burchell's zebras the family group is the basic social unit. It generally consists of a single adult male and two or three adult females with their foals. The groups are stable, apparently because of strong mutual ties among the females rather than because of herding by the male. The stallion is dominant, and there is a hierarchy among the mares, the highest ranking (alpha) animal usually leading the group. Other males are either solitary or live in bachelor groups of two or three, sometimes up to 10. Juveniles leave the family group when they attain sexual maturity at one and one-half to two years. When large aggregations occur on favoured grazing grounds, the groups retain their identity. Among Grevy's zebras and wild asses, territorial males and groups of mares and foals and of stallions are found. There is no evidence for territorial behaviour among any of the zebras except Grevy's. Individual groups occupy home ranges that overlap to some degree with those of other groups.

The social organization of other equids is not as well documented. Observers studying the wild horse and half-asses have noted that females and juveniles form a group dominated by a single stallion, which keeps them together by active herding; the unattached males are solitary or live in small herds.

The pattern of social organization among the rhinoceroses is quite different from that of the Equidae. Dominant adult males of the white rhinoceros occupy territories that, in the Natal reserves, average about 200 hectares (500 acres). Within its area a male may tolerate subadult or aged bulls, which have subordinate status. Adult females accompanied by their calves inhabit home ranges encompassing the territories of six or seven dominant bulls. Juveniles consort with other juveniles or with calfless females, but groups of more than two usually do not stay together long.

The black rhinoceros is basically solitary. Adults of both sexes usually occupy home ranges of 1,000 hectares (2,500 acres) or more, the size depending on the characteristics of the environment; occasionally the ranges may be as small as 200 hectares, however. There is a good deal of overlap in the utilization of these home ranges.

The Asiatic rhinoceroses also are essentially solitary, but detailed information on the nature of the areas they inhabit is not available. Individual great Indian rhinoceroses are said to occupy tracts as small as eight to 20 hectares (20 to 50 acres).

Little is known about the social organization or territorial behaviour of the tapirs. All species are reported to be found alone or in pairs.

Scent marking and territories

Male zebras and horses follow mares in estrus. The stallion, after smelling the spot where a mare has urinated or defecated, exhibits "flehmen" (a characteristic display in which the head is lifted and the upper lip raised) and then urinates or defecates on the same spot. In similar fashion, members of stallion groups often urinate or defecate consecutively; communal dung heaps formed by five to eight animals often arise in this way. The significance of such behaviour is not clear.

Among the Rhinocerotidae excretory products play an important role in marking territories and home ranges. Dominant male white rhinoceroses defecate almost entirely on heaps within their territories. They then scatter the material by kicking vigorously, presumably leaving an individual scent mark in this manner. In addition, they urinate in a ritualized fashion, spraying the urine in powerful jets in a manner peculiar to them and shown by no other sex or age group. Other members of the population also use dung heaps (either in territories or in communal areas, such as along paths) but not exclusively, and they do not scatter dung.

In the black rhinoceros, dung-scattering behaviour does not appear to be exclusive to dominant males. The function of the communal heaps may be mainly to establish the presence of the inhabitant in his home range, and to maintain contact between known animals.

Dung heaps and urine spraying are also observed among other species of rhinoceroses and among tapirs; their significance is presumably of a similar nature.

Fighting. The pattern of fighting is related to the amount of lethal equipment the various groups possess. The Equidae, unarmoured, do not employ stylized fighting techniques to reduce the danger of serious injury—as among certain other species. Fighting is largely confined to adult males competing for estrous mares. Various techniques occur in the zebras, which may serve as an example of the family. Circling, neck fighting, biting (either in a standing or sitting position), rearing combined with biting and kicking, and kicking on the run all are used, either alone or in combination. No set pattern is followed.

Fights among rhinoceroses consist of charges and striking with the horns, usually accompanied by vocal threats. Goring is not common, the stylized pattern having probably been evolved to minimize the danger of serious injury from the formidable horns.

Mutual grooming and greetings

Amicable behaviour. Mutual grooming is well known among horses. Two animals stand facing in opposite directions and groom each other by nibbling at the root of the tail and the base of the neck. Burchell's zebra behaves similarly and so, presumably, do other members of the family.

Zebras greet each other simply by nose-to-nose contact, except that adult stallions go through a ceremony involving nose-to-genital contact. Nose-to-nose greeting is also characteristic of tapirs and rhinoceroses. The latter also rub their bodies together.

Courtship and mating. Courtship is relatively simple among the social equids. The true ass is apparently exceptional. The partners are strangers when the first approaches are made and the female requires violent subjugation by the male, which bites, kicks, and chases her before she will stand for him. This may be the result of separation of the sexes outside the mating season. The wild horse and Burchell's zebra are not at all violent. The stallion often grooms the mare before attempting to mount. The estrous mare (especially, or exclusively, young mares in the case of Burchell's zebra) adopts a typical posture with legs slightly apart, tail lifted, and, except in the horse, a characteristic facial expression (the "mating face" already mentioned).

The more or less solitary rhinoceroses and tapirs go through a more elaborate courtship, presumably because the partners are strangers. After a chase, the male and female may engage in low-intensity fighting, ending with the male laying his head on the female's rump and then mounting and copulating for an extended period. Several males may mate with an estrous female.

Relations between parent and offspring. The perissodactyls bear well-developed (precocial) young, usually a single offspring. After the mother has assisted in removing the placenta and has licked her offspring in the usual mammalian fashion, the young animal soon attempts to stand. A Burchell's zebra foal has been observed to stand quite firmly 14 minutes after birth, and a black rhinoceros calf 25 minutes after birth.

Newly born equids follow any nearby object during the first few days of life. At this time, zebra mares drive away all other zebras from their foals. The behaviour ensures that the foal will form a bond with its mother during the initial period of imprinting. Foals follow their mothers closely and are groomed frequently.

Although precocial, black rhinoceros calves appear to have a lying-out period; that is, an initial period when they rest quietly in thick cover except when being suckled. Thereafter they follow their mothers closely. A young white rhinoceros tends to walk ahead of the mother and may be guided by her horn.

Most young perissodactyls remain with their mothers until the next offspring is born. A young rhinoceros may, therefore, accompany its mother until it is two and one-half years old, or older. Although grazing starts early, suckling proceeds for a considerable time, perhaps for its psychological rather than its physiological value.

Running games and mock fighting

Play. As in other mammals, play is a prominent form of behaviour among young perissodactyls. Zebras up to the age of one year frequently engage in running games. Foals gallop wildly about on their own, jumping and kicking up their heels, sometimes chasing other animals, such as gazelles, mongooses, or birds. In groups they play catching games, running after one another in close succession. Mock fighting sometimes takes place. Groups of adults have also been seen to chase foal groups in play, and indeed stallion groups carry out playful gallops. Stallions also engage in play greeting and in mock fights.

Playful romping and mock fighting with the horns are common among rhinoceros calves. Young tapirs play running games.

Rolling and wallowing. Behaviour for the care of the body is widespread among the perissodactyls. Equids frequently roll in dry, loose soil forming rolling hollows—a common feature of zebra country.

Wallowing, which may help regulate body temperature, probably is mainly a form of self-grooming; it is practiced by all species of rhinoceroses. They often spend hours lying in pools during the middle of the day in hot weather. Mud of suitable consistency induces wallowing, which may be followed by sand bathing. Prolonged rubbing on tree trunks or suitable stumps follows a wallow; old rubbing stumps and stones may take on a shine from repeated use.

Tapirs may have the most pronounced tendency to bathe and wallow, but few details of their behaviour are known. They are also said to enter water when disturbed.

Reproduction. Female equids of all the species for which information is available attain puberty at about one year, but are not normally successfully mated before the age of two to two and one-half years, and possibly as late as three to four years in the case of Grevy's zebra. Zebras probably breed until about 20 years of age. The domestic species are seasonally polyestrous (repeatedly fertile), coming into breeding condition in spring and, unless mated, undergoing repeated estrous cycles at intervals of approximately three weeks until the end of the summer. The wild species studied also tend to mate seasonally; most young are born in spring and summer.

Reproductive potential of equids

The gestation period of equids is between 11 and 12 months. In most species a postpartum estrus occurs, usually within two weeks of the birth of the young; thus, the maximal potential reproductive rate is one young per year. This potential is not always attained. Only about 50 percent of domestic mares that are mated produce foals, and nearly half of a study group of Burchell's zebra mares bore only one foal in three years.

The gestation period of three species of rhinoceroses is about 15 to 17 months. For the Sumatran rhinoceros the period is said to be only seven months. No information is available for the Javan rhinoceros. Female white and black rhinoceroses attain sexual maturity at the age of four to five years and are capable of calving at intervals of approximately 2½ years. Rhinoceroses probably breed until between 30 and 40 years old. The white tends to have a mating peak in spring, corresponding with the flush of green grass, and a calving peak in autumn.

The Malayan and Brazilian tapirs have gestation periods of 13 months' duration. The Brazilian tapir is reported to mate before the onset of the rainy season.

FORM AND FUNCTION

Integument. The skin of rhinoceroses is extraordinarily thick. The great Indian and Javan rhinoceroses are covered with large, practically immovable plates, separated by joints of thinner skin to permit movement. The two species differ in the arrangement of the folds. The hair of all rhinoceroses is sparse or absent except that of young Sumatran rhinoceroses, which have a dense coat of crisp, black hair. The skin of the tapirs is also thick with a sparse covering of short hairs arranged in irregular groups. The equids have a normal hide with a well-developed hairy coat.

Structure of rhinoceros horns

The "horns" of the rhinoceroses are noteworthy structures of epidermal origin. The horn is composed of a mass of fused epidermal cells that are impregnated with a tough, fibrous protein (keratin) and that rest on a roughened bony cushion on the fused nasal bones. The male Javan rhinoceros has a short horn about 25 centimetres (10 inches) long; the horn of the female is rudimentary. The great Indian rhinoceros has a single horn up to 60 centimetres (25 inches) long. The other species of rhinoceroses have a second horn that stands on a protuberance of the bones (frontals) between the eyes. Hornlike structures were also present in the titanotheres and extinct rhinoceroses.

In all living perissodactyls the terminal digital bones are flattened and triangular, with evenly rounded free edges, and are encased by keratinous hooves derived from the integument. The single hoof of the equids—the only mammals to walk on the tips of single digits—is the most highly developed structure of this kind among mammals. The keratinous wall is analogous to the nail of mammals that have claws or nails.

Skeleton. *Backbone.* The vertebral column acts as a firm girder, with high dorsal (neural) spines on the thoracic vertebrae, above the forelimbs and ribs. Spines and ribs serve as compression struts above and below. The column balances largely on the forelegs and is pushed from behind by the hindlegs, which are the main propellants. This skeletal structure permits running and also enables great weights to be borne in such animals as the rhinoceroses. There are never fewer than 22 thoracolumbar (trunk) vertebrae.

The neck, or cervical, vertebrae are opisthocoelous; *i.e.,* with the bodies (centra) of the vertebrae hollowed behind to take the convex heads of the succeeding centra. This feature facilitates rotatory movement of the neck and is most highly developed in the horses.

Limb girdles. The shoulder blade is long and narrow with a small coracoid process (a ridge to which muscles are attached) and a low spine. There is no clavicle (collarbone). The pelvic girdle has a broad, vertically raised ilium to which are attached the large gluteal (thigh) muscles, important for locomotion, and the abdominal muscles, which carry the weight of the belly.

Limbs. There is a clear evolutionary tendency in the Equidae for the limbs to become long and slender, with a reduction in the number of digits in the swift-running forms. These changes are accompanied by an increase in rigidity and specialization for movement fore and aft. The upper (proximal) segments, the humerus of the forelimb and the femur of the hindlimb, have remained short. In contrast the lower (distal) parts, consisting of an anterior radius and posterior ulna in the forelimb and an anterior tibia and posterior fibula in the hindlimb, have become longer and slimmer. The humerus is short and broad. Its articulation with the radius and ulna only permits fore-and-aft movement. The proximal end of the stout femur has a third projection, or trochanter, in addition to the usual mammalian two, which serves as an additional point of attachment for the large locomotory muscles of the hindlimbs.

Evolution of swift-running forms

The anatomical feature of the order now considered most significant is that the axis of symmetry of the limbs passes through the third or middle toe, the most strongly developed and the one on which most of the weight is borne. This is called the mesaxonic condition and is contrasted with the paraxonic condition of the Artiodactyla, in which the axis passes between the third and fourth toes.

Originally the five toes of the limb were held in the semidigitigrade position—*i.e.,* with the weight of the body being borne on the soles of the toes and on the lower ends of the elongate metacarpal and metatarsal bones of the forefeet and hindfeet. The upper (proximal) ends of these bones were raised above the ground, a condition still to be seen in the tapirs.

As the third digit became increasingly dominant, it became longer and thicker. The upper ends of the third metacarpus and metatarsus broadened and forced the other digits to the side. The first (inner) digit was the first to disappear. The earliest known forms already bore only three digits on the hindfoot. The loss of the first toe on the front foot led to the four-toed condition common in the Eocene; the fifth digit persisted, although it was somewhat

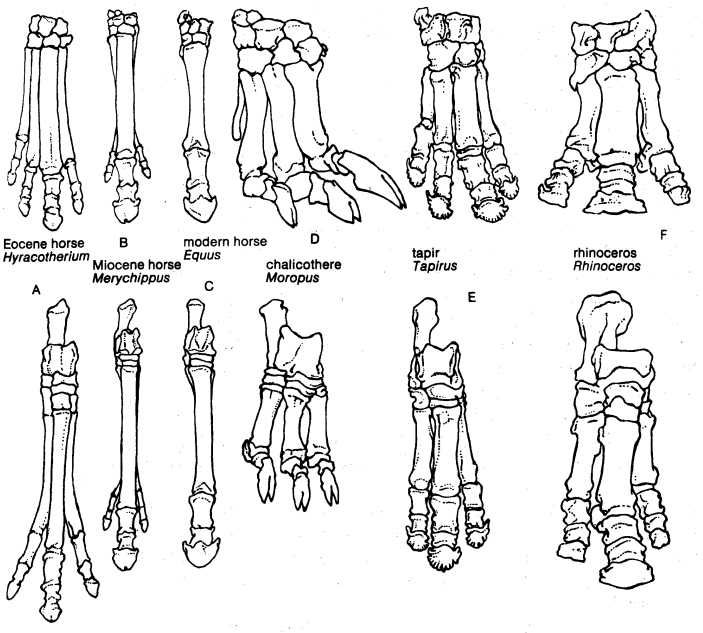


Figure 48: Forelimbs (above) and hindlimbs (below) of some recent and fossil perissodactyls.

Drawing by R. Keane based on (A, B, C, D) A.S. Romer, *Vertebrate Paleontology*, Copyright 1933, 1945, 1966 by the University of Chicago; all rights reserved

weak. The tapirs, with four toes in front and three behind, have retained this early condition.

The fifth digit of the hindfoot was the next to disappear in the evolutionary sequence, and all known forms beyond the lower Oligocene had only three toes in each foot. The living rhinoceroses illustrate this condition. A massive limb with a broad foot is essential in such heavy animals. The feet have cushions of elastic connective tissue well suited to bear the weight of the body.

In the most highly specialized forms, the second and fourth digits also underwent reduction. These digits are retained in the living Equidae only as functionless, vestigial slivers of bone on either side of the third metacarpal and metatarsal.

Growth of the third digit's terminal joint

With reduction in the number of digits, the third has assumed a progressively more vertical position. Its terminal joint, or phalanx, has become larger and the hoof surrounding it bigger and thicker. At the same time the ulna, the smaller bone of the forelimb, decreased in size until, in the modern Equidae, its upper end fused with the radius and its lower end remained merely as part of the articulating surface of the radius with the wrist bones (carpals). In the hindlimb, the fibula became reduced in similar fashion. It articulates with the ankle bones (tarsals) only in a few extinct forms, such as the titanotheres. In the tapirs and rhinoceroses it is slender. In the equids the proximal end remains as a small splint of bone, while the distal end has fused with the tibia.

The arrangement of the small bones of the carpus ("wrist") and tarsus (ankle) is another characteristic feature of perissodactyl limbs. In most other mammalian orders, the carpals and tarsals provide the limbs with regions of flexibility. Their development in the perissodactyls has been toward increasing rigidity, following the trend in the limbs as a whole. The basic mammalian arrangement is one of three rows of bones: a proximal row of three, adjacent to the lower arm and leg bones; an intermediate row of four (the centralia); and a distal row of five (carpalia of the forefoot, tarsalia of the hind), one to each digit. The centralia, carpalia, and tarsalia are numbered one to four or five, beginning on the side of the thumb and big toe. In all orders the number of wrist and ankle bones has been reduced. Usually only one of the centralia, known as the navicular, remains, while there are the same number of carpalia and tarsalia as there are digits.

An interlocking plan is characteristic of the Perissodactyla. In the forefoot, the third distal carpal (the magnum or capitate) is enlarged and interlocks with the proximal carpals. The elongated third metacarpal thrusts up against these interlocked bones. In the equids, distal carpal I (the trapezium) is absent, and the arrangement in the hindfoot is similar. In the most advanced, modern forms metatarsus III thrusts against the enlarged and flattened tarsal III (ectocuneiform), and this in turn is in contact with the large, flat navicular (centrale II and III). The navicular abuts on the flattened astragalus (or talus), the intermediate bone of the proximal row. The articulation between the upper surface of the astragalus and the tibia is pulley-like and permits only fore-and-aft movement of the limb.

Teeth. The full complement of mammalian teeth consists of three incisors, one canine, four premolars, and three molars in each half of each jaw. The arrangement may be expressed by the formula $\frac{3 \cdot 1 \cdot 4 \cdot 3}{3 \cdot 1 \cdot 4 \cdot 3} = 44$ teeth. The figures represent the number of incisors, canines, premolars, and molars in each half of the upper (above the line) and lower (below) jaws, respectively.

Fully-toothed early hoofed mammals

The Condylarthra, a group of mammals that first appeared in the Paleocene (about 65,000,000 years ago) and were ancestral to most of the later and recent hoofed mammals, had a full complement of teeth. In many of the early perissodactyls, only the first lower premolar had been lost. The subsequent evolutionary sequence led to losses and specializations of the incisors and canines. Lengthening of the facial part of the skull resulted in the formation of a gap, the diastema, between the incisors and the premolars. The first upper premolar was reduced or lost in consequence. The canines, when present, were situated in this diastema, as they are in male horses.

Among the living perissodactyls, the tapirs have the least specialized battery of teeth. In this, as in many other features, they have remained primitive. The dental formula of the family Tapiridae is: $\frac{3 \cdot 1 \cdot 4 \cdot 3}{(2-3) \cdot 1 \cdot 3 \cdot 3} = 40\text{–}42$ teeth. The first upper premolar is noteworthy in being the only premolar with a milk, or deciduous, predecessor.

The cutting teeth are reduced in the Rhinocerotidae. Incisors and canines are absent in the two African forms. The Asiatic species have one or two upper, but generally no lower, incisors in each half of the jaw. They have no upper canines; the Javan and Sumatran rhinoceroses have one short, sharp lower canine on each side. There are three upper and lower molar teeth in all five species. The white and Sumatran rhinoceroses have three premolars, and the others have three or four premolars. The dental formula for the family is set up, therefore, as follows: $\frac{(0-2) \cdot 0 \cdot (3-4) \cdot 3}{0 \cdot (0-1) \cdot (3-4) \cdot 3} = 24\text{–}30$ teeth. The dental formula of the Equidae is $\frac{3 \cdot 1 \cdot (3-4) \cdot 3}{3 \cdot 1 \cdot 3 \cdot 3} = 40\text{–}42$ teeth.

The form of the premolars and molars is of great interest and their evolutionary history has been studied in some detail. Primitive, browsing members of the order had brachydont cheek teeth (*i.e.*, with low crowns and long, narrow root canals), with separate low, rounded cusps—the bunodont condition. Increasing specialization for grazing resulted in fusion of the cusps into ridges (lophs), thus teeth of this kind are called lophodont. Lower molars typically have two transverse lophs, the protoloph and the metaloph. In the upper molars these ridges are fused with a longitudinal ridge (ectoloph), which runs along the outer edge of the tooth. Further development leads to a convoluted arrangement of the lophs, such teeth being termed selenolophodont.

Associated with these changes in the tooth surfaces is a tendency for the crown to become higher. High-crowned teeth are termed hypsodont. The hollows between the lophs of hypsodont teeth are filled with a deposit of secondary cement, which strengthens the teeth and makes them more resistant to wear. A further evolutionary trend is for premolars to become as large as molars. Where the process of molarization is complete, as in horses, all grinding teeth are identical.

The Tapiridae have primitive brachydont premolars and

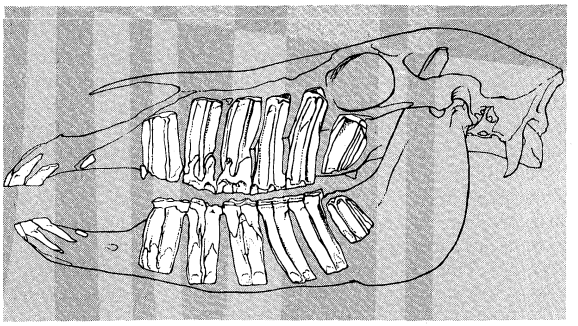


Figure 49: Skull and dentition of horse, *Equus caballus*, cut away to show roots of teeth.

From Horses: *The Story of the Horse Family in the Modern World and Through Sixty Million Years of History*, by G.G. Simpson. Copyright 1951 by Oxford University Press, Inc. Reprinted by permission

molars. They possess two simple transverse lophs, and are thus termed bilophodont. They lack secondary cement on the crowns.

The Sumatran rhinoceros, the most primitive of the living rhinoceroses, and the Javan rhinoceros have similar brachydont, lophodont cheek teeth. The great Indian rhinoceros, which is less of a specialized browser, has hypselodont (hypsodont and selenodont) premolars, with a layer of cement on the crowns. The black rhinoceros has brachydont and lophodont teeth, with a thin layer of cement. The white rhinoceros is more specialized, for the cheek teeth are hypselodont and have a thick cement layer.

The grinding teeth of the Equidae are highly specialized, high crowned, with a complicated selenodont surface and thick cement deposits.

Comparative size of the stomach

Digestive system. The stomach of perissodactyls is small, simple, and undivided. In the horse its capacity is only 8.5 percent of the whole digestive system. The comparable figure for the ox is 71 percent. The intestine is very long and the cecum (blind gut) and colon are huge and sacculated (*i.e.*, with many blind pockets). Here food is macerated and fermented and the fibrous portions are dissolved. The liver has no gall bladder.

EVOLUTION AND PALEONTOLOGY

The Perissodactyla appeared early in the Eocene, about 55,000,000 to 40,000,000 years ago. Together with most other ungulate mammals, they were probably derived from the Condylarthra. The condylarths were abundant in Europe and North America, mainly during the Paleocene (65,000,000 to 55,000,000 years ago). Condylarths were unspecialized mammals, rather carnivore-like in appearance. The larger species attained the size of tapirs. The limbs were fairly short and primitive but the third toe was somewhat enlarged and the phalanges ended in hooves. There was a full complement of teeth with some specialization of bunodont molars for herbivorous diet.

Horses. The earliest horses appeared during the lower Eocene in Europe and North America. They are generally known as *Eohippus* ("dawn horse"), but *Hyracotherium* is the correct taxonomic designation. Some species of these little forest-dwelling, browsing animals were no larger than a terrier. They had moderately long, slender limbs with only four toes in the forefoot and three in the hindfoot, all equipped with hooves. The molars were essentially bunodont (with low, rounded cusps) and the premolars simple.

Hyracotherium-like animals persisted in Europe until the end of the Eocene. Another group, the paleotheres or "native" European horses, evolved as a specialized side branch, which died out in the Oligocene. North America was the centre of horse evolution. During the Eocene, *Hyracotherium* was succeeded by forms such as *Orohippus* and *Epihippus,* which are known only from that epoch.

The Oligocene (40,000,000 to 26,000,000 years ago) saw a major change with the appearance of three-toed horses, *Mesohippus, Miohippus,* and others. All of the premolars were similar to the molars, low-crowned but lophodont (ridged). *Anchitherium* was an early Miocene form as large as a modern pony, which migrated from North America to Europe. These primitive three-toed horses or anchitheres survived until Pliocene times, some of their descendants attaining the size of a rhinoceros.

The main course of horse evolution entered a third stage in North America in the Miocene Period (26,000,000 to 7,000,000 years ago). A line of grazing horses developed, almost certainly to exploit the new grasslands that were spreading over the surface of the earth. The degree of lophodonty of the molariform teeth increased, changing the pattern of the crests on the surface and increasing their grinding efficiency. Of greater importance, these teeth became hypsodont (high-crowned) and thus maintained a good grinding surface as grinding of the harsh, siliceous grass caused them to wear down. Another substance, cement, came to supplement the dentine and enamel forming the teeth of earlier types, and provided additional material to resist abrasion. The evolution of these specialized teeth was a tremendous advance.

Trends in horse evolution

The limbs of the grazing horses became increasingly rigid and specialized for fore-and-aft movement, better fitting the animals for running in open country. In *Merychippus* the ulna was fused with the radius and the fibula was much reduced. In some advanced forms the central toe was much larger than the two lateral toes and carried most of the weight of the body on a hoof much like that of modern horses.

A number of evolutionary lines developed during the Pliocene, which lasted from 7,000,000 to 2,500,000 years ago. *Pliohippus* of North America is probably the line from which modern horses have come. The genus *Equus* is characteristic of the Pleistocene when it developed in North America and spread to all continents except Australia. By the end of the Pleistocene, horses had become extinct in the New World.

Titanotheres. Another group entirely, the titanotheres (Brontotheriidae), evolved independently from *Hyracotherium*-like ancestors and became abundant in North America during the Eocene and Oligocene, but disappeared by the Miocene. They were also found in Asia and Eastern Europe. The end forms, such as *Brontops* and *Brontotherium,* were huge, the largest standing 2½ metres (eight feet) at the shoulder. They had long, low skulls and a small brain. Many species bore a pair of large, hornlike processes on the front of the head.

Chalicotheres. The chalicotheres (Chalicotheriidae) were moderately large animals that appeared in Eurasia and North America during the Eocene. Thereafter they evolved mainly in the Old World, disappearing from America in the mid-Miocene, but persisting in Asia and Africa until they died out in the Pleistocene. Early mem-

By courtesy of the Field Museum of Natural History, Chicago

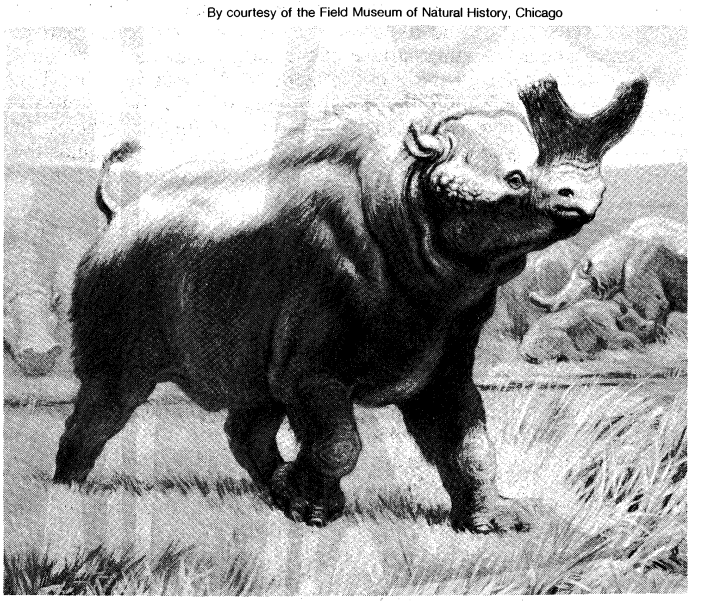


Figure 50: Reconstruction of fossil Perissodactyla. *Brontops,* an Oligocene titanothere, some individuals of which were about 2.5 metres (eight feet) at the shoulder. Detail of a painting by Charles R. Knight.

bers of the group such as *Paleomoropus,* from the lower Eocene, resembled contemporary equids. The Miocene *Moropus* typifies the peculiar characteristics of later forms. It was horselike, but the front legs were longer than the rear. Each foot bore three toes which ended in large, fissured phalanges bearing claws that may have been used for digging up roots and bulbs.

Tapirs. The fossil and living tapirs, together with similar extinct forms such as the lophiodonts, constitute a small, fairly uniform group.

Homogalax, Lophiodon and other tapir-like animals appeared in the Eocene. Some were close to the ancestry of both rhinoceroses and tapirs, and other evolutionary lines left no descendants. *Protapirus* from the Oligocene of Europe and North America was the forerunner of the modern tapirs. These tapirids persisted until the Pleistocene, when climatic changes led to their extinction.

Rhinoceroses. The fossil history of the rhinoceroses is far more complex. They evolved from the early tapiroids but diverged from that group. Unlike the equids, they have tended to grow large, with short, stout limbs with little reduction in the digits. Although the premolars have tended to molarize, the molar cusp pattern is simple; the teeth seldom become hypsodont, and cement occurs rarely.

The Hyracodontidae, or running rhinoceroses, from the Eocene and Oligocene of North America and Asia were the most primitive. *Hyracodon* had long, slender legs with three toes on each foot and typically rhinocerotid cheek teeth.

The amynodonts are known from the late Eocene and Oligocene of Eurasia and America and lived in Asia until the Miocene. They were a side branch, perhaps derived from primitive hyracodonts. *Metamynodon* and some other forms were about as large as hippopotamuses and may have lived in rivers. The premolars were simple and the incisors reduced, but canines and molars were enlarged.

True rhinoceroses are probably also descended from early hyracodonts. They became numerous in the Oligocene as large animals with molariform premolars. Many side branches evolved. *Trigonias* still had four front toes but *Caenopus,* another Oligocene representative, had the three toes common to all later rhinoceroses. Like its relatives of that period, *Caenopus* was fairly small, about the size of a tapir, and hornless.

One spectacular group contained giant hornless creatures, such as *Baluchitherium* and *Indricotherium* from the Oligocene and Miocene of Asia; the largest of known land mammals, they stood 5.5 metres (about 18 feet) high at the shoulder and had a long neck and long forelegs.

Rhinoceroses died out in North America during the Pliocene, but in Eurasia many different species survived to the Pleistocene. One of the most familiar of these later rhinoceroses is the woolly rhinoceros (*Coelodonta*), which was depicted by Stone Age artists and is known from nearly intact specimens. The surviving rhinoceroses represent remnants of once varied and abundant stock; the genera have little direct relationship to one another.

CLASSIFICATION

Distinguishing taxonomic features. Skeletal features are of greatest importance in classifying the Perissodactyla. They are, of course, the only criteria applicable to fossil forms. Distinguishing characteristics of the skull include the relative length of the facial region, length and form of the nasal bones, presence of a postorbital bar and of hornlike structures. The form and number of teeth, presence of a diastema, number of molariform premolars, the height of the cheek teeth, their cusp structure and the presence or absence of cement on the grinding surfaces all are particularly valuable taxonomic features. The limb structure is also a useful aid to classification, especially the length of upper and lower portions, the degree to which the ulna and the fibula are reduced and fused, respectively, with the radius and the tibia, the number and form of carpal and tarsal bones, and the number and relative size of the digits.

Importance of skeletal features

Among living perissodactyls, body size, form of the upper lip, number and length of horns, structure of the skin and colour pattern are the most important features for classification.

Annotated classification. The classification presented here follows that of U.S. paleontologist George Gaylord Simpson, which is generally accepted but modified in some fairly minor ways by other authors. The term Mesaxonia, introduced by the 19th century paleontologist O.C. Marsh, is essentially synonymous with Perissodactyla. Simpson used it as a convenient designation for the superorder containing the single order Perissodactyla, as he used the parallel name Paraxonia for the superorder with the one order Artiodactyla.

Groups indicated by a dagger (†) are known only as fossils.

ORDER PERISSODACTYLA

Herbivorous ungulates with either 3 digits or 1, at least in the hindfeet. Early forms digitigrade, this feature either retained or completely replaced by the unguligrade condition in later representatives. The body weight borne mainly or entirely on the third digit through which the long axis of the limb passes (mesaxonic condition). The talus (heel bone) with only 1, proximal, keeled surface, articulating with the tibia, no duplication of keel on the distal surface, as in artiodactyls. Nasals broad at the posterior end, alisphenoid canal present. Posterior premolars molariform; cheek teeth bunodont in early forms, typically lophodont or selenolophodont, rectangular; wide diastema separates them from incisors and canines, which may be reduced or absent. Testes inguinal, occasionally scrotal; mammary glands inguinal, with 2 teats; uterus bicornuate. Fifteen living species, but more than 200 fossil forms.

Suborder Hippomorpha

Superfamily Equoidea

Dentition complete, upper molars with 6 tubercles, the 2 external ones united to form an ectoloph, median and internal tubercles generally fused into a single loph. Tendency to molarization of premolars and reduction of lateral digits.

†*Family Palaeotheriidae* (paleotheres or "native horses"). Lower Eocene to lower Oligocene; Europe and Asia. Size variable, from that of fox (shoulder height about 40 centimetres or 16 inches) to that of a rhinoceros; body form tapirlike. Three digits in each foot. Dentition complete or first premolars absent; premolars molariform, separated from canines by diastema. Cheek teeth brachydont, lophodont, with or without cement. Orbits without postorbital bar. Eurasian "native horses," 7 genera.

Family Equidae (horses and relatives). A progressive group with tendency to molarization of premolars and development of cement.

†*Subfamily Hyracotheriinae* (hyracotheres). Eocene: Europe and North America. Size about that of a fox; 4 toes in forefoot, 3 in hindfoot. Dentition complete, premolars simple or last 2 slightly molariform, molars brachydont, bunodont, diastema short. Skull short, without postorbital bar. Three genera. Basic stock from which other members of family evolved.

†*Subfamily Anchitheriinae* (anchitheres). Lower Oligocene to lower Pliocene. Fossils from North America, Europe, Asia. Browsing horses. Early forms collie-sized (shoulder height about 60 centimetres [23.6 inches]), end forms as large as a rhinoceros. Three toes in each foot, central toe somewhat enlarged; metapodials elongated. First premolars reduced, second to fourth molariform. Cheek teeth brachydont, lophodont; ectoloph (longitudinal outer crest) above typically W-shaped cross-lophs retained distinct cusps; hypostyle at back edge of upper molars prominent. Lower cheek teeth with 2 crescent-shaped lophs and double cusp at junction. Diastema long, no postorbital bar. Six genera.

Subfamily Equinae (horses, asses, half-asses, and zebras). Lower Miocene to Pleistocene of North America; Pleistocene of South America; lower Pliocene to Recent of Europe, Asia, and Africa. Most forms as large as ponies or modern horses. Feet with 3 toes in early forms, 1 in later forms (but metapodia 2 and 4 persist as remnants), limbs with progressively elongated lower segments and reduced or fused ulna and fibula. Calcaneum without articulating surface for fibula. First premolars reduced or absent, others molariform; cheek teeth hypsodont, selenolophodont, with cement; canines in long diastema. Skull long, with postorbital bar in upper Miocene and later forms; nasals long and narrow. Some Recent forms with striped coat.

†*Superfamily Brontotherioidea*

†*Family Brontotheriidae* (titanotheres). Lower Eocene to lower Oligocene. Fossils from North America and Asia. Early forms as large as tapirs, end forms huge, standing 2.5 metres (8.2 feet) at the shoulder. Graviportal browsers; 4 toes in forefoot, 3 in hindfoot. Dentition complete, premolars small

and simple in early forms; later forms with incisors and first premolars reduced or absent, remaining premolars only partly molarized. Molars large, bunolophodont, above with W-shaped ectoloph as in early equids, lower with double V resembling equid pattern but lacking reduplication of cusp at union of V's. Skull elongated, low, without postorbital bar, with large rough, hornlike process in later forms. About 39 genera.

†Suborder Ancylopoda

†*Family Chalicotheriidae* (chalicotheres). Upper Eocene to lower Pliocene. Fossils from North America, Europe, Asia. Early forms equivalent in size (and similar) to contemporary horses, *Hyracotherium;* later representatives larger, up to size of a modern horse. Most with forelegs longer than hindlegs and 3 toes in each foot; rudimentary metacarpal V retained in forefoot, most with clawlike digits. Incisors of canines reduced or absent; first premolar absent, remainder small and simple. Molars bunolophodont, cusp pattern similar to Brontotheriidae. About 13 genera. Aberrant perissodactyls which may have used claws to dig for bulbs and roots.

Suborder Ceratomorpha

Superfamily Tapiroidea

Brachydont forms, molars with simple ectoloph and strongly developed transverse protoloph and metaloph. Forefoot generally with 4 toes, hindfoot with 3.

†*Family Isectolophidae.* Lower to upper Eocene. Fossils from North America and Asia. Limb structure similar to that of *Hyracotherium,* teeth already tapiroid. Four genera.

†*Family Helaletidae.* Lower Eocene to middle Oligocene. Fossils from North America and Asia. Skull advanced in structure; premolars much simpler than molars, molars of simple lophodont form. Ten to 12 genera.

†*Family Lophiodontidae* (lophiodonts). Lower to upper Eocene; Europe and probably Asia. Skull not tapir-like, lophs of molariform teeth oblique. About 6 genera.

Family Tapiridae (tapirs). Lower Eocene to Recent. Fossils from North and South America, Europe, Asia, Africa. Recent species in Central and South America and Southeast Asia. Limited to tropical equatorial forests of South America and Asia. Browsers. Skull greatly modified; premolars completely molarized, first sometimes absent; transverse lophs strongly developed in molariform teeth, ectoloph simple. Limb structure primitive; 4 toes in forefoot; 3 in hindfoot. Six genera but only 1 extant.

Superfamily Rhinocerotoidea

Molars brachydont, upper with 2 transverse lophs fused with a well-developed ectoloph; lower with 2 asymmetrical crescents. Premolars more or less molariform. Rough processes on nasals (for horns) frequently present; forefoot with 3 or 4 toes.

†*Family Hyrachyidae.* Lower to upper Eocene; North America and probably Asia. Small animals; forefoot 4-toed; molariform teeth with separate tubercles. Four genera.

†*Family Hyracodontidae* (hyracodonts or running rhinoceroses). Middle Eocene to upper Oligocene. Fossils from North America and probably Asia. Relatively small; feet 3-toed, legs long and slender. Molariform teeth typically rhinocerotid. About 7 genera.

†*Family Amynodontidae* (amynodonts). Upper Eocene to middle Oligocene. Fossils from North America, Europe, Asia. Hippopotamus-like; forefoot 4-toed, hindfoot with 3 toes. Skull short, heavy, premolars and incisors reduced, premolars simple, canines and molars enlarged. Six genera.

Family Rhinocerotidae (rhinoceroses). Middle Eocene to present. Fossils from North America, Europe, Asia, and Africa. Now restricted to tropical Africa and Asia. Large; most forms 3-toed. Skull elongated, raised at posterior end. Dentition incomplete, upper canines always absent, some incisors usually enlarged as cutting teeth; premolars molariform, last upper molar simple. About 34 genera, 4 extant.

Status of the chalicotheres

Critical appraisal. Systems of classification frequently differ in the status given to the chalicotheres. Simpson takes the view that their ancestry was probably equid and almost certainly hippomorph. He holds that the significance of their claws has been over-emphasized and has tended to distract attention from their true affinities. Accordingly he places them in the superfamily Chalicotheroidea of the suborder Hippomorpha. In the classification presented above they are given their own suborder, Ancylopoda, following the views of Alfred S. Romer, another authority on the group.

The forms united in the family Lophiodontidae by Simpson, followed here, are thought by some recent workers to warrant separation into three families. Romer distinguishes the families Lophialetidae, Deperetellidae, and Lophiodontidae. The affinities of certain primitive genera such as *Hyrachus* and *Colonoceras* remain controversial; Simpson places them in the family Hyrachyidae, superfamily Rhinocerotoidea. Romer considers them to be early tapiroids and assigns them to the Helatidae (superfamily Tapiroidea). The difference of opinion is slight, for it is generally agreed that the hyrachids are close in the common stem of the tapiroids and rhinocerotoids.

(R.C.Bi.)

Artiodactyla (pigs, goats, sheep, cattle, giraffes, deer, camels)

The mammalian order Artiodactyla, or even-toed ungulates, includes the pigs, peccaries, hippopotamuses, camels, chevrotains, deer, giraffes, pronghorn, antelopes, sheep, goats, and cattle. It is one of the larger mammal orders, containing about 150 species, a total that may be somewhat reduced with continuing revision of their classification. Many artiodactyls are well-known to man, and the order as a whole is of more economic and cultural importance than any other group of mammals. The much larger order of rodents (Rodentia) affects man primarily in a negative way, by competing with him or impeding his economic and cultural progress.

GENERAL FEATURES

Abundance and distribution. Artiodactyls were once the dominant herbivores (plant-eating mammals) of almost every continent. They are an important link in the chain by which the sun's energy, having been used by green plants, is made available to other forms of life. They tend to be medium- or large-sized animals. If they were any smaller they would compete with rabbits and the larger rodents, and if they were larger they would compete with elephants and rhinoceroses, the largest of terrestrial herbivores. The success of artiodactyls has depended on skeletal adaptations for running and on the development of digestive mechanisms capable of dealing with plant foods; none is adapted to flying, burrowing, or swimming. The individual species tend to be fairly narrowly adapted, in comparison with other mammals, but many of them nonetheless have broad distributions.

Native artiodactyls are absent only from the polar regions and from Australasia, but many have been introduced into Australia and New Zealand. In Australia, the position of medium and large herbivores is occupied by kangaroos. Through most of its evolutionary history, the order was absent from South America; only within the last few million years have some groups entered that continent. The occurrence of the majority of living artiodactyls in the Old World is a recent phenomenon; a considerable variety once inhabited North America.

Families of living artiodactyls

The order Artiodactyla contains nine families of living mammals, of which the Bovidae (antelopes, cattle, sheep, and goats) is by far the largest, containing nearly 100 species. There are five Eurasian and four African species of pigs (family Suidae) and two Central and South American species of piglike peccaries (Tayassuidae). The two hippopotamus species (Hippopotamidae) are African. The more familiar large species were until recently widespread throughout Africa south of the Sahara and in the Nile Valley; the pygmy hippopotamus has a restricted distribution in West Africa. The camel group (Camelidae) was formerly abundant in North America, the now extinct North American stocks having produced the camelids of South America (wild guanaco and vicuña, domestic llama and alpaca) and the Old World dromedary and Bactrian camel.

The remaining artiodactyls (*i.e.,* the suborder Ruminantia) are all ruminants (cud chewers), the most primitive of which are the chevrotains (Tragulidae), with three species in Asia and one, the water chevrotain, in West Africa; the chevrotains are clearly remnants of a group that was once more numerous and widespread. Deer (Cervidae) are basically Eurasian and have not spread into sub-Saharan Africa, although they have reached the Americas. There are about 30 species, the greatest number being concen-

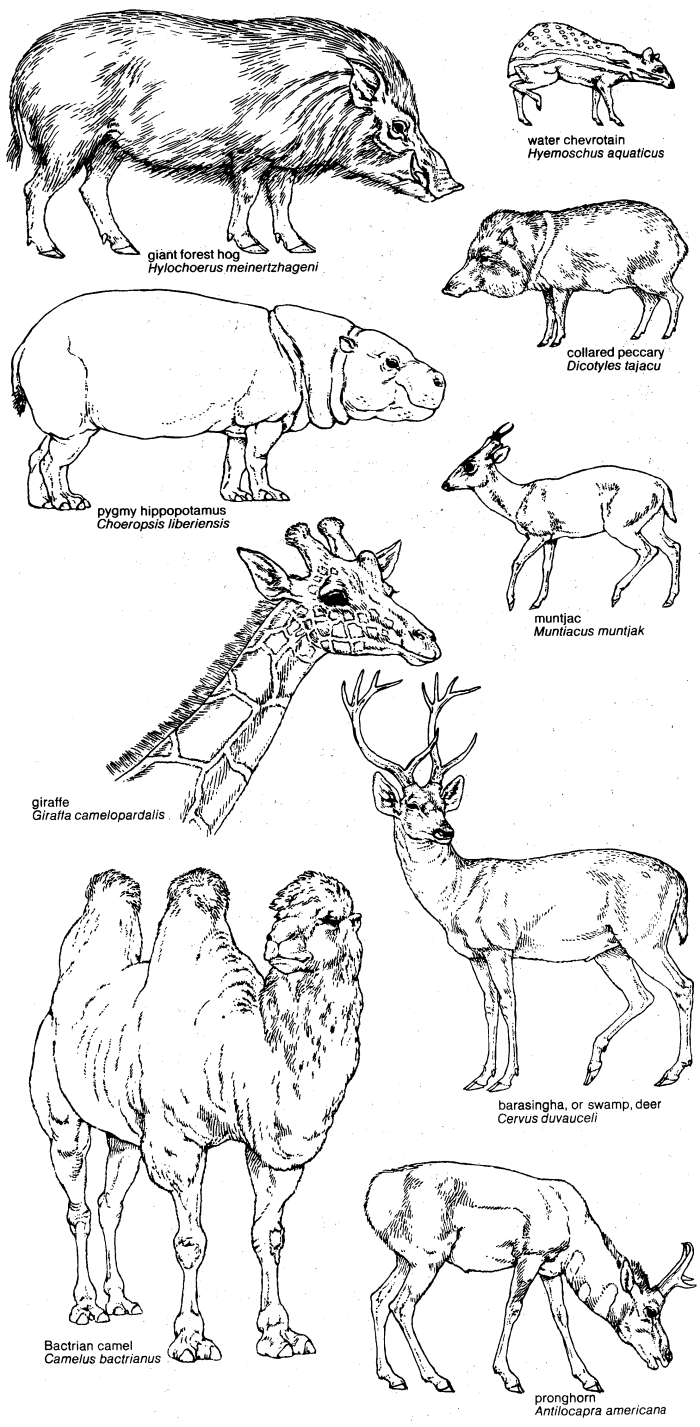


Figure 51: Representative non-bovid artiodactyls.
Drawing by R. Keane

trated in South America and tropical Asia. The giraffe and the okapi (Giraffidae), two distinctive African species, are closely related to deer. The pronghorn (Antilocapridae), although sometimes called pronghorn antelope, is not a true antelope; it is the only survivor of a stock of ruminants that was very successful in the later part of the Tertiary Period in North America (about 2,500,000 to 65,000,000 years ago). The family Bovidae is primarily African and Eurasian, with a few members in North America. Bovids are advanced artiodactyls, many of which live in open grassland and semi-arid areas.

Importance to man. Artiodactyls have long been exploited by man for economic purposes. At Olduvai Gorge in East Africa there is clear evidence of the use of antelopes for food almost 2,000,000 years ago. In Europe during Paleolithic times (about 30,000 years ago) Cro-Magnon man depended heavily on the reindeer. By this time the use of animals other than as food had become established; skins were used as clothing and footwear, and bones were used as tools, weapons, and accessories.

The domestication of animals was a major advance in human history. Domestication of herd animals probably arose gradually, perhaps before agriculture. Domesticated goats and sheep are first known from the Near East at some date close to 7000 BC. Cattle and pigs were domesticated at some subsequent date but certainly before 3000 BC. In South America the llama, now used for transport, and the alpaca, which provides a source of wool, were developed from guanacos by the Incas or their predecessors. The dromedary (*Camelus dromedarius*), domesticated in Arabia, was introduced into the Southwestern United States, southwestern Africa, and inland Australia in the 19th century. A large feral population now exists in Australia.

In addition to providing meat, milk, hides, and wool, artiodactyls have served man in a number of other ways. In Kashmir, the underfleece, or pashm, of the Siberian ibex (*Capra ibex*) and of local domesticated goats has been used as the basis for the manufacture of cashmere shawls. In southwestern France, pigs have been used to locate underground truffles (the fruiting bodies of certain edible fungi).

No group of mammals is more extensively hunted than the artiodactyls. Sport hunting of various deer supports a multimillion-dollar industry in North America and Europe. In many cultures hunting has been reserved for monarchs or the aristocracy. In the centuries after the Norman Conquest of England, the forest law provided severe punishment for the slaughter of deer and boars. Père David's deer (*Elaphurus davidianus*) of China now survives only because it was preserved first in the hunting park of the emperors of China and later by the Duke of Bedford after the slaughter of the Chinese herds at the end of the 19th century.

Wild ungulates were the primary source of meat for human populations long before the appearance of modern man. Prehistoric man hunted the large mammals of his environment with an ever increasing effectiveness that was certainly instrumental in his survival. The extent to which man was involved in the extinction of some of the larger Pleistocene animals (*i.e.*, those that were abundant 10,000 to 2,500,000 years ago) is still being investigated. There is now known to have been a wave of late Pleistocene extinction of large mammals, including artiodactyls; in North America this wave reached its zenith about 9000 BC. Many animals also became extinct in Africa, where long-horned buffalo and large relatives of hartebeests survived until very recently. More of the large mammals have survived in Africa than elsewhere, but the reason for their survival is not known. A second, probably final, wave of extermination of the larger mammals has taken place with the spread of European culture and firearms in the past 300 years. It has been marked by wanton slaughter and has ultimately produced an interest in conservation. It now seems, however, that the unprecedented demands on the environment being made by rapidly expanding human populations will result in a nearly complete extinction of large wild mammals.

NATURAL HISTORY

Behaviour. *Migration.* Many artiodactyls undertake seasonal migrations between their breeding grounds and feeding areas or between different feeding areas. They can then take advantage of the seasonal changes in different areas. This means that larger populations, and hence a larger biomass (*i.e.*, the total weight of all individuals in an area), can be supported than if all passed their lives in one area. The North American mule deer (*Odocoileus hemionus*) comes from its summer pastures at high altitudes as the first snow falls and returns at the end of winter, several weeks after the snow has melted.

Seasonal changes

Social behaviour. Although the popular image of artiodactyls is one of great herds numbering thousands of individuals, some species are solitary, and many others form only small family groups. The maternal family unit, in fact, is the most cohesive one, providing the basis for

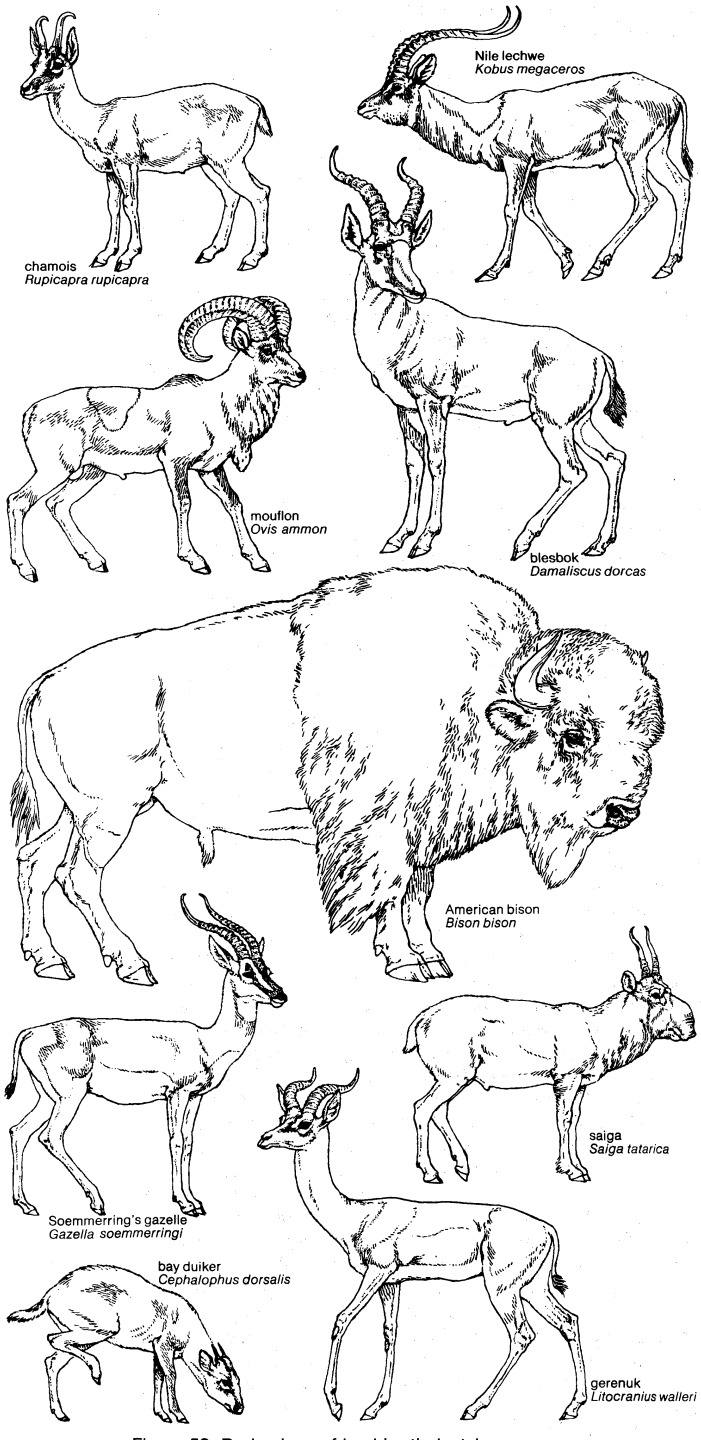


Figure 52: Body plans of bovid artiodactyls.
Drawing by R. Keane

herd formation. Most artiodactyls are more or less social, and grazing forms may be found in especially large aggregations. It appears that the practice of aggregating gives protection, favouring those members of the species that are the most active contributors to the gene pool (thus the most available to natural selection), since the individuals most frequently taken by predators are old, solitary males, males maintaining territories, and animals of either sex separated from the herd.

The advantage of herd behaviour

Social facilitation (the instigation of collective behaviour) takes place in herds. After one animal flees, all of the others flee, and the predator may thus not catch any. Social facilitation may also promote a restricted season for births; this helps survival of the young by denying these easy-prey individuals to predators through much of the year, and keeps the predator population lower than if young were available throughout the year. Another advantage of herding is that the older generation in a herd can guide migrations to water, feeding areas, or mating grounds.

Females and young are usually in herds separate from those of the younger males, but territorial (the older, proven) males may accompany the females. There are some variations of this behaviour. In the Eurasian roe deer (*Capreolus capreolus*), for example, the basic unit includes the doe, her litter of two, and often the young of the previous year. During the rutting (mating) season males associate with females in heat but do not gather harems. The female herds of red deer (*Cervus elephas*) are separate from the males except in the breeding season, when the stag will defend his female herd against other males. Among cattle and related species, the males associate with the females and young, but the bulls are ranked below a so-called master bull, each defending its place within the rank order. Female hippopotamuses and their young form a group in water and have a favourite resting and basking sandbank. The males have their resting places around this area. Each male's rank in the social hierarchy determines how close to the females he may be.

There can be some flexibility of social organization within a species. During the rutting season the male Rocky Mountain goat (*Oreamnos americanus*) makes little effort to herd females within a fixed area if there is little snow, but he does drive off other males. When there is much snow, he neither fights other males nor defends individual females.

Forest-dwelling artiodactyls often live singly, as does the okapi (*Okapia johnstoni*) of central Africa; individuals meet only for mating. Female moose (*Alces alces*) with calves are intolerant of their own young of the previous year and of adults, so even small herds do not form.

Territoriality

The territory of an animal is an area from which the possessor attempts to exclude other individuals of the same species (and occasionally other species). An animal in an area lacking its own scent is more timid and ready to flee. Among solitary artiodactyls the territory holder defends an area sufficient to meet his needs for food and shelter. Among social artiodactyls the territorial system is interwoven with breeding activities, and territories are normally defended only by certain males. Other males are driven off, and a percentage of males are prevented from mating.

The most simple territorial organization among artiodactyls is that of the common wild pig (*Sus scrofa*), which lives within a home range including resting, feeding, drinking, and wallowing places. There is little sign of territorial defense, and the herd (called the sounder) may move to a new area. At the other extreme, male Uganda kob antelopes (*Kobus kob*) hold territories, for breeding only, that are as small as 15 to 30 metres (50 to 100 feet) in diameter. There are 30 to 40 territories on the breeding ground of a herd, and groups of females and young move about the territories despite the efforts of individual males to detain them. The semi-arid Serengeti plains of northern Tanzania contain nomadic aggregations of blue wildebeest (*Connochaetes taurinus*), males of which defend temporary territories only while an aggregation remains stationary.

In territorial defense an aggressive encounter between males is generally preceded by visual signalling of intentions. Chital deer (*Cervus axis*), for example, have several sorts of threatening displays. When sharp, potentially lethal horns appeared in early ruminants, intimidating displays rather than combats would doubtless have been favoured. Horns or antlers eventually functioned to maintain head contact during struggles rather than to bruise, slash, or gore. This stylized fighting, in which the competing males interlock horns or antlers and try to "outwrestle" each other, minimizes the danger of killing an opponent of the same species (conspecific). It evolved in two ways: further development of the wrestling, found in stags and some of the antelopes, and ramming, as in sheep. In sheep the horns are the sole organs of display. They increase in size throughout life and parallel the dominance order of the males, so that unnecessary fighting is minimized. Ramming may have intermediate forms; goats, for example, butt with a sideways hooking motion. In the fighting of hornless artiodactyls, such as pigs, the combatants may be

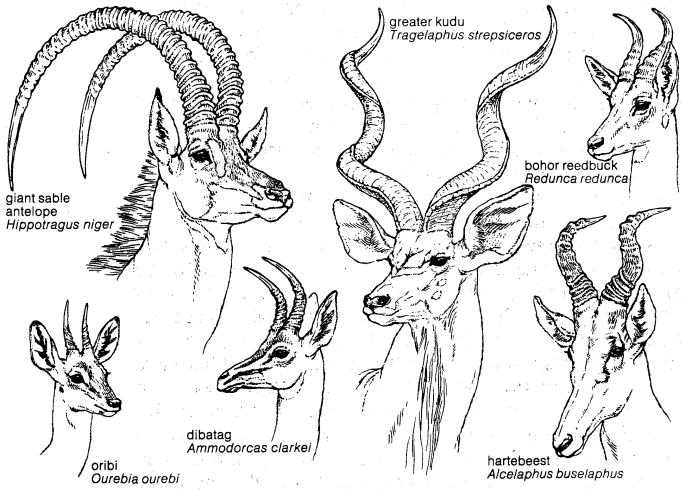


Figure 53: Horn variation in African antelopes.
Drawing by R. Keane

badly mauled or even killed. The fighting behaviour of camels retains primitive elements of biting, kicking, and neck wrestling.

Reproduction. Many advanced artiodactyls have elaborate courtship behaviour, a regular component of which is for the male to sniff or lick the female's urine, and afterward to raise his head slightly with upcurled lips. This behaviour, which has been called *flehmen,* apparently enables the male to recognize females in heat. In the mating ceremonies of tragelaphine antelopes (kudus, bushbucks, and others) the male follows the female, nuzzling her neck several times. When he mounts, he lays his neck along hers so that their heads touch. In Thomson's gazelle (*Gazella thomsoni*), following the *flehmen* behaviour, the male runs close behind the female and finally taps her hindleg with his foreleg. Similar leg contact also occurs in some other antelopes. Its function could be to test the female's readiness to mate, to habituate her to contact, or to heighten her readiness to mate. It appears to be equivalent to the neck contact of tragelaphines. During mounting, the male Thomson's gazelle holds his head high and does not touch the female's flanks with his forelegs; the pair may continue walking. This is probably a more advanced pattern of events than that in tragelaphines. The kob antelope has elaborate displays after mating. These and the specialized sexual displays seem to be a consequence of this species' tightly clustered territories on the mating grounds. Another pattern occurs in the normally solitary Indian hog deer (*Cervus porcinus*); as many as 20 or 30 aggregate loosely in a certain area, then females and males leave in pairs and usually remain together until they have mated. Mating in artiodactyls often intensifies toward dawn and dusk.

Gestation periods

Gestation periods vary and are related in part to the size of the animal. They range from four months in the small chevrotain to 14 months in the Bactrian camel (*Camelus bactrianus*) and over 14 months in the giraffe. Females of normally gregarious species become solitary a few days before giving birth. The female chital, or axis deer, for example, remains near a patch of dense bush and high grass to which she can retreat if endangered. The female collared peccary (*Dicotyles tajacu*) withdraws to a burrow. The European wild pig gives birth in a rough nest.

In temperate regions, birth takes place in spring or early summer, and in tropical areas there are often more births during or just after the rainy season. The absence of a well-defined breeding season in a species may indicate less rigorous environmental conditions, which sometimes vary in different parts of a species' range. Warthogs have one restricted breeding season in most of eastern and southern Africa, while elsewhere two seasons or year-round breeding have been recorded. The breeding season of the waterbuck (*Kobus ellipsiprymnus*) is continuous in Uganda, but in Zambia its breeding season shows a sharp peak at the height of the rains.

Most modern artiodactyls have one young at each birth, but there are some well-known exceptions among ruminants. The Chinese water deer (*Hydropotes inermis*) bears twins or triplets, but during gestation carries even more fetuses; early records (now known to be incorrect) of large litters were based on observations of dead pregnant females containing the large number of fetuses. The mule deer, white-tailed deer (*Odocoileus virginianus*), roe deer, pronghorn (*Antilocapra americana*), nilgai (*Boselaphus tragocamelus*), four-horned antelope (*Tetracerus quadricornis*), and saiga (*Saiga tatarica*) commonly bear twins. In the white-tailed and mule deer and in the saiga, a higher percentage of twins are borne by the older females; this is probably true in other species. The number of young is usually three in the warthog, five in the European wild pig, and two in peccaries.

The female wild pig almost ignores her young, which free themselves from their birth membranes and seek a teat. Female camels show comparatively little maternal attention and do not eat the afterbirth (the fetal membranes and placenta). Ruminants generally eat the afterbirth, as well as the dung and urine of the young, thus helping to prevent discovery of the young by predators. Licking of the young tends to facilitate its recognition by the mother.

Precocity in offspring

An artiodactyl is normally precocious (well developed) at birth and may weigh one-tenth as much as its mother. An extreme example of precocity is the wildebeest calf, which rises within five minutes of birth, follows its mother within another five minutes, and can move as fast as an adult in 24 hours. Young deer fawns "freeze" during danger but rejoin the herd when the danger is long past or when retrieved by the mother.

Pigs and hippopotamuses are weaned after a few months, but among higher artiodactyls, lactation lasts longer. Wildebeest, for example, suckle for almost a year, although they start to eat grass when only a few days old. This may either maintain a bond between parent and offspring and form the base for larger social groupings or help to "develop" the four-chambered stomach. Higher artiodactyls eat soil when they begin to eat solid food, probably to establish a normal flora and fauna in the rumen (the first of the four stomach chambers).

Locomotion. Artiodactyls are preyed upon by carnivores and therefore need speed and agility to escape death. They have an added disadvantage in the sheer weight of their very large stomachs, which they need in order to digest plant food. Running ability reaches an extreme in advanced artiodactyls living in open country. The hippopotamus, with an adult weight of 2,500 to 3,000 kilograms (5,500 to 6,600 pounds), is the only living artiodactyl big enough to need heavy, pillar-like limbs for support.

In the normal walking of artiodactyls the legs move in the following order: (a) left front, (b) right rear, (c) right front, (d) left rear. This basic pattern is masked in faster walking or trotting by each foot being lifted off the ground before the one ahead of it in the sequence reaches the ground, resulting in telescoping the first (a and b) and second (c and d) pairs of movements. In galloping or fast running the two front legs leave the ground one immediately after the other, then the two back legs. The chief propulsive force in locomotion comes from the back legs, except in the giraffe (*Giraffa camelopardalis*), in which the front legs provide the main propulsive power.

Camels often amble, both legs of each side moving together, and the giraffe and the okapi always use this walking gait. Here the middle two (b and c) and the first and last (a and d) actions of the normal walking pattern occur together. The giraffe, having a short body and great height, could not adopt the normal ruminant gait without tripping. The long neck moves back and forth in time with the strides and helps smooth the movement. Galloping by the giraffe is of the normal ungulate type.

Artiodactyls living among bush or rocky cover may develop a bounding sort of gait in which the legs are pulled up very sharply during each stride. Deer and some antelopes are examples. When walking, species in such habitats are supported by the diagonally opposite legs for a greater length of time in each stride than are fast-running, open-country ruminants. This is a more primitive stable

position and allows an easier leap from hidden danger. Some bovids, notably goats in Eurasia and the klipspringer (*Oreotragus oreotragus*) of Africa, are especially agile on rocky slopes and precipitous ground.

Maximum speeds of locomotion

The maximum speeds of some artiodactyls are: warthog, 48 kilometres (30 miles) per hour; camel, 14–16 kmph (9–10 mph); giraffe, a little over 48 kmph (30 mph); Cape buffalo (*Syncerus caffer*), 56 kmph (35 mph); Thomson's gazelle, 80 kmph (50 mph).

Ecology. *Food habits.* Most artiodactyls are closely tied to the resources of their environment. They are dependent, for example, on feeding areas not being covered by too much snow or shrivelled under a drought, and on the regulating effects of fire or other herbivores on the seasonal succession of vegetation. Various grazing species feed on grass at different heights. Browsers, those that feed on the foliage of shrubs and trees, show more extreme variation in feeding height, the maximum being that of the giraffe.

Herbivorous animals need less initiative and intelligence to collect food than do the meat-eating, hunting carnivores, but digestion is more difficult. Advanced artiodactyls have evolved the ability to bolt food and to ruminate it (chew it more thoroughly) at a later time or while resting in an area where they may be less obvious to predators and can conserve energy. Tropical artiodactyls frequently have adaptations for water conservation, having developed to a high degree internal physiological regulation (homeostasis).

Primitive artiodactyls were probably omnivorous but favoured plant foods, a characteristic still found in pigs. The latter dig with the snout and, to a lesser extent, with the front legs and upper tusks (canine teeth). The wart-hog of Africa (*Phacochoerus aethiopicus*) has a modified method of gathering food. When food is scarce it forages for young grass shoots under very low bushes; its tusks and localized thickening on its skin protect the eyes and muscles from thorn damage, and small incisors enable it to pluck food.

Hippopotamuses (*Hippopotamus amphibius*), although they spend a great deal of time submerged in lakes or rivers, do not feed in the water. They graze at night, wandering over well-used trails, sometimes far from water, often damaging crops.

Most members of the camel family are found in arid habitats. The vicuña (*Lama vicugna*) of the South American Andes lives at high altitudes where it grazes on soft grasses and herbs. It has much the same food requirements as domestic sheep.

Chevrotains live in dense undergrowth close to water or in marshes, where they browse on soft vegetation, roots, and tubers, following a way of life probably not unlike that of their ancestors.

The other ruminants browse or graze. They may take many plant species in the course of the year, but at any one season a large part of the diet consists of only five or six plants. Some ruminants are strongly specialized. The reindeer of the Arctic (*Rangifer tarandus*), for example, eats a variety of sedges, grasses, and herbaceous plants in summer but, as the long winter approaches, gradually shifts to a diet of lichens. It uses its front feet to scrape snow away from lichens to a depth of about 60 centimetres (two feet). The females are unique among deer in possessing antlers, which are thought to help them get scarce food in late winter by driving off the males that have by then shed their antlers. Reindeer may eat lemmings. The red deer, on the other hand, has catholic feeding habits. In woods it browses on lichens, berries, fungi, and the leaves of most deciduous trees; in open country it eats grass, heather, berries, and lichens. Shrubs and trees are used more in winter. When the red deer lives in the same areas as other ruminants it can be a serious competitor for food.

The importance of grasses in the ruminant diet

Grasses form a substantial part of the diet of many ruminants. Young grass consists of about 5 percent protein, 1 percent fat, 3 percent minerals, and 20 percent carbohydrates; the remaining percentage is water. The most noticeable changes as grass ages are an increase in carbohydrate content to 75 percent and a large decrease in the amount of water. Such food, especially when coated with silica, as are many grasses, or when covered with dust, would be impossible for nearly all nonruminant herbivores to eat or digest. The major evolutionary trend in ruminants has been to make use of grasses and grasslands, and the higher ruminants have evolved largely in adaptive balance with one another. This adaptive balance was shown during a study of the change from plains to thickets of scrub growth in an area in the eastern Congo over a period of about ten years. There was an accompanying decrease in numbers of antelopes and warthogs, no change in buffalo, and an increase in elephants and hippopotamuses.

There is not usually a one-to-one dependence of any artiodactyl species on one plant. The plant species that constitute the major part of the diet may vary with the season, and similar parts of different plants may be eaten in preference to other parts of the same plant. Food resources in an area are thus parcelled out among the various artiodactyls present. Sometimes behavioral differences minimize competition between closely related species in the same area. A study has shown that in central Africa the roan antelope (*Hippotragus equinus*), a grazer, favours open areas with taller, ranker perennial grasses and is more or less sedentary within a small area; the sable antelope (*H. niger*), also a grazer, prefers savanna woodland or the edges of open areas, and herds follow a more or less cyclic annual route over an area of about 500 square kilometres (200 square miles). When pasturage is restricted, sheep will cut grass very short, and goats will damage trees and bushes. An American zoologist, George B. Schaller, has observed that, in Kanha Park in central India in the hot season, blackbuck (*Antilope cervicapra*) continue to graze on grass shoots in open areas; chital deer seek out tender grass blades, especially along forest edges, and also feed on leaves and fruits; barasingha (*Cervus duvauceli*) eat dry and moderately coarse grass along ravines; sambar deer (*Cervus unicolor*) browse on leaves and crop coarse grasses in the forest; and gaur (*Bos gaurus*) graze on tall, coarse grass and break down saplings to get at the leaves. The choice of habitat also varies: chital avoid steep terrain and forests with an unbroken canopy; blackbuck require less water than the others and thus remain in drier regions; sambar and gaur are less specialized in habitat requirements, and both are active primarily at night; barasingha prefer reed beds but also enter forests and climb hills.

It has also become evident that grazing successions are one of the mechanisms that enable the maximum use to be made of environmental resources. On the Serengeti plains, for example, the wildebeest grazes on ground already covered by the zebra and leaves the grazed grass in a condition suitable for the Thomson's gazelle. Interactions take place between artiodactyls and some plant species. It has been noted in the Tarangire area of northern Tanzania that *Acacia* seedlings germinate only where the impala (*Aepyceros melampus*) has left its dung. In parts of southern Peru plants growing on or close to the dung of the vicuña are different from those of the surrounding pasture.

Artiodactyls often favour the boundary zone between habitats. In Rhodesia, Lichtenstein's hartebeest (*Alcelaphus lichtensteini*) is usually found at the edge of clearings adjacent to woodland.

Areas of distribution. Some artiodactyls have surprisingly small ranges; Hunter's hartebeest (*Beatragus hunteri*) and the dibatag (*Ammodorcas clarkei*), for example, are found in two very restricted areas in eastern Africa. Others have extremely large ranges, such as the roe deer, which lives from the western shores of Europe to the eastern shores of Asia, or the red deer, which is found in a similar band across Eurasia and is regarded by many as conspecific with the North American wapiti or elk (otherwise called *Cervus canadensis*). Sometimes a considerable area may be occupied by a chain of related species, an example being the oryxes; the beisa and gemsbok (races of *Oryx gazella*) occur in South and East Africa, the scimitar-horned oryx (*O. dammah*) in West Africa, and the Arabian oryx (*O. leucoryx*) in Arabia.

The influence of climate on the distribution of artiodactyls

It is well known that climate is one of the factors limiting the ranges of artiodactyls. A number of South African antelopes differ, at the species level, from their ecological counterparts farther north in Africa. The bontebok and blesbok, races of *Damaliscus dorcas,* are found in the south and the sassaby (*D. lunatus*) farther north; the black

wildebeest (*Connochaetes gnou*) occurs in the south and the blue wildebeest (*C. taurinus*) farther north. This probably is a result of climatic or climatically influenced factors; each species evidently functions best in a certain temperature and aridity range. Wide distributions can occur more easily along lines of latitude than they can by spanning the tropics to temperate or polar regions. Species that cross lines of latitude are often associated with mountain chains, examples being the Rocky Mountain goat, with its wide latitudinal range in western North America, and the goral (*Nemorhaedus goral*), found from Indochina to the Amur River. Climatic effects on distributions sometimes occur with regard to altitude. In Central Asia, the goa (*Gazella picticaudata*) is found in valleys from 3,000 to 3,660 metres (10,000 to 12,000 feet) above sea level, the chiru (*Pantholops hodgsoni*) and the yak (*Bos mutus*) are on the very high steppe between 5,500 and 6,100 metres (18,000 and 20,000 feet).

South America has a more impoverished artiodactyl fauna than Africa, being limited to deer and camelids. This arises in part from the late arrival of the artiodactyls (deer in middle to late Pliocene, about 4,000,000 years ago, camelids perhaps a little later) and in part because a number of large rodents compensate for the shortage of large herbivores. The cervids in South America have not shown the same capacity for radiation in open country as have bovids in the Old World.

Influence of pests on distribution

The areas of distribution and numbers of individuals are determined by complicated interweaving of effects not yet completely understood. Bloodsucking flies are thought to be the main reason that red deer in Scotland ascend to higher feeding grounds in June, and reindeer are afflicted by horse flies (*Tabanus*) and other dipteran pests. It is questionable whether the level of artiodactyl populations is controlled by predation, by availability of food, by reproductive rate, by disease, by climate, or by competition, insofar as these can be regarded as separate factors. It is known that undernourishment increases the susceptibility of an animal to the effects of parasites. If such an infected animal, say a pig, is caught by a leopard, it would be an oversimplification to assign a single reason for its death; it could have died from starvation, parasites, or predation. There is no evidence that artiodactyls are affected more than marginally by predators during most of their mature lives. Mortality is greatest among juvenile and aged animals. In a study of central African warthogs, it was estimated that a 60 percent loss occurred during the first six months of life in an expanding population and 95 percent in a declining one. Although predation was thought to be the main cause, another was the fact that the piglets had only limited control over their body temperatures and were thus more at the mercy of environmental temperature change. Food supply may sometimes be decisive, either directly or through the indirect action of intermediate agencies such as drought. The year 1961 lacked long rains, causing a severe shortage of forage in the Nairobi Game Park in Kenya. Many antelopes died of starvation, populations fell, and those of the blue wildebeest had not recovered nine years later, perhaps for reasons unconnected with the initial drought. Disease has generally been considered to have only a secondary importance in regulating numbers.

Thickness of the snow cover in winter is a very important factor for Asian artiodactyls. The saiga, for example, cannot move in snow deeper than about 40 centimetres (16 inches), and the wild sheep *Ovis ammon* in snow deeper than 60 centimetres (24 inches), at the most. The snow may have other effects; a layer of ice on top of snow may damage an animal's legs and weaken the animal to the extent that it is caught by a predator. Saiga may be unable to dig through even a shallow layer of compacted snow. Hoarfrost on vegetation is especially dangerous when prolonged or when it occurs in consecutive winters, though elk may escape the worst effects by feeding in winter on bark and high shoots. Massive periodic mortalities among Palearctic (Eurasian) ungulates in winter have been known since ancient times. The saiga has adapted to these crises by migrating great distances in a short time away from snowstorms or from areas where fodder is short. It also has a very rapid maturation to a reproductive state, ensuring that populations will build up after heavy mortalities.

Population density over the range of a species is affected by social behaviour, such as the effects of territoriality, dispersal of the young, and whether the species lives in herds. Fecundity may be reduced in overcrowded conditions by effects on reproductive control mechanisms, reduced viability of the young, or retarded maturation.

FORM AND FUNCTION

General structure. Artiodactyls have larger stomachs and longer intestines than carnivorous animals because plant food is less easily digested than meat. The necessity of escaping predators and the handicap of a heavy digestive system have resulted in limb bone adaptations.

Comparison of odd- and even-toed ungulates

In all artiodactyls the main weight-bearing axis of the leg passes through the third and fourth toes together. This has been called paraxonic support and is contrasted with the mesaxonic limb support of the other great order of herbivorous mammals, the perissodactyls (rhinoceros, horse, tapir), in which the weight-bearing axis passes through the third or central toe alone. As artiodactyls evolved there was increasing development of the third and fourth toes and a parallel decline of the second and fifth toes flanking them. Progressive simplification of limb extremities has characterized the evolution of the artiodactyls, and even in the earliest known artiodactyls, the pollex and hallux (corresponding to the big toe and thumb of man) were already rare.

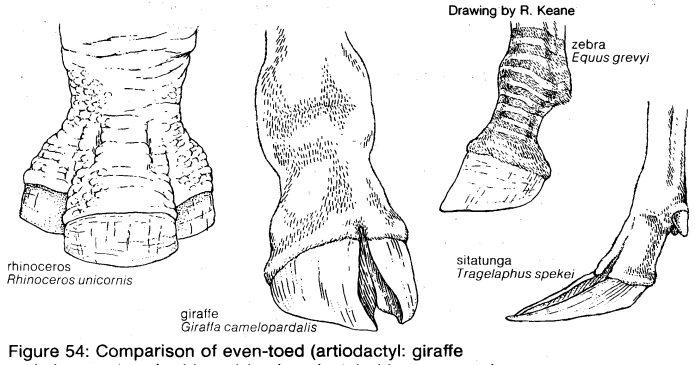


Figure 54: Comparison of even-toed (artiodactyl: giraffe and sitatunga) and odd-toed (perissodactyl: rhinoceros and zebra) ungulate feet.

The other main morphological characteristic of artiodactyls is that the astragalus, one of the bones in the ankle, has upper and lower rounded articulations (areas of contact of bones) and no constricted neck, instead of simply one rounded articulation above a neck, as in other mammals. This character is so basic to artiodactyls that it has not developed very much within the known history of the order, having already been present in long extinct members. The artiodactyl astragalus also has an articulation on its rear surface for the calcaneum (heel bone). The three articulations are in nearly parallel planes, allowing the astragalus to rotate vertically.

Other features of the limbs, skull, and dentition distinguish artiodactyls. The ulna (posterior forearm bone) and fibula (posterior bone of the lower leg) have become reduced. The humerus, the upper bone of the foreleg, is large and has a large protrusion, the greater trochanter, to which muscles are attached. The femur, the upper bone of the hindleg, has a large greater trochanter and a second, lesser trochanter, but lacks the third trochanter characteristic of perissodactyls. There are typically 19 thoracic and lumbar (upper and lower back) vertebrae. The separate lumbar region of the spine is retained with its forwardly directed transverse processes (lateral projections on the vertebrae). There is no clavicle, or collarbone, in the shoulder girdle. The hip girdle shows fore-and-aft elongation and a well-developed ischium (upper anterior bone of the pelvis). There is never a penis bone.

The large tongue is very mobile and can be thrust forward. The brain is moderately developed, with folding of the surface of the cerebral hemispheres variably developed, often less in small artiodactyls than in large ones.

The olfactory region of the brain is well developed and hearing is acute. The brains of earlier artiodactyls, such as the extinct entelodonts, were smaller than those of later forms. There are often scent glands on the head and body.

Specializations of the head. The skulls of pigs and peccaries lack a complete bony bar behind the eye (postorbital bar) as in most suiform artiodactyls and the early camels. The hippopotamuses, most camels, all ruminants, and two fossil suiform groups (entelodonts and oreodonts) have a complete postorbital bar. Any surface exposure of the periotic bone (bone around the ear) on the skull is called the mastoid, and skulls without such a surface exposure are described as being amastoid. Amastoid skulls are found in most suiform groups (including entelodonts, anthracotheres, and all living suiform groups); mastoid skulls occur in some early suiform groups, oreodonts, and all remaining artiodactyls that have lived since the end of the Eocene Epoch (about 38,000,000 years ago). Hippopotamuses have many modifications for aquatic life—large lungs, eyes and nostrils on top of the head, nostrils that can be closed by muscular control, and small ears. They are able to remain submerged for at least five minutes.

Horns and antlers Pigs, peccaries, hippopotamuses, camels, and chevrotains have no horns or antlers. In the early Miocene, Old World ruminants related to giraffes and deer first developed such appendages. The majority of deer have antlers, defined as solid, bony, branched outgrowths of the frontal bones, present only in the males (but also in female reindeer) and shed seasonally. They are not covered by a horny sheath but, during a growth period of about four months, have a fine-haired skin or "velvet." The antlers have two basic branches, the anterior or brow tine, and the posterior branch or beam. The brow tine is unbranched, except in Père David's deer, in which both it and the beam are branched, the brow tine forming the dominant part of the antler. Antlers are specialized sex characters used for fighting by males in the rutting season and to scrape or slash at trees and bushes for territorial marking.

The distinction between horns and antlers

A study of the chital deer showed that antlers increase in size up to the seventh year, remain at a constant size until the ninth year, then decline. The horn of bovids consists of a hollow, unbranched horny sheath (formed of modified skin like fingernails and toenails) that fits over a bony core; horns are often present on both sexes. If such a horn is accidentally lost it is not regenerated; this is unlike the situation in deer, in which normal shedding is followed by regrowth. In the giraffe, but not in the okapi, horn growth is mainly from the parietal bone. The pronghorn has horns in both sexes. The sheaths are shed each year after the breeding season, and new ones develop under the old ones. The sheath is two pronged, but the underlying bony core is unbranched.

Teeth. There is a complete set of teeth in early artiodactyls and in modern pigs of the genus *Sus,* consisting on each side of three upper and lower incisors, an upper and lower canine, four upper and lower premolars, and three upper and lower molars. There has been a tendency toward reduction of the front teeth and development of a gap (diastema) between them and the back teeth. There has been very little tendency for the premolars to molarize, and the first premolar often disappears. Early forms had five-cusped upper molars, but the fifth cusp (protoconule) disappeared early.

Members of the suborder Suiformes have the full complement of incisors and canines, except for peccaries, which lack the lateral pair of upper incisors. Hippopotamuses have continuously growing incisors and canines, the lower canines being very large.

The canines of pigs grow continuously. In this group the canines are weapons for offense and defense, the sharp cutting edges of the lower canines being maintained by wear against the uppers. Young camels retain the full complement of front teeth, with three incisors and one canine in the upper and lower jaws; the upper incisors are extremely small. In the upper jaw of the adult only the rear incisor and canine are present. The vicuña has continuously growing lower incisors.

The molars of pigs are low crowned (except those of the warthog) and have many cusps; those of peccaries are more simple. Peccaries have one less premolar than pigs; camels also have reduced premolars. Chevrotains have rather flattened lower premolars but have incipiently selenodont molars; *i.e.,* in which the cusps are drawn out into longitudinal crescents. Premolars of ruminants are wider, and the molars definitely selenodont. In many bovids and the pronghorn, but not in giraffes or deer, the molars are markedly high crowned.

Limb adaptations for fast running. Adaptations for fast running reach an extreme in advanced artiodactyls living in open country. In addition to the increased rotation of the astragalus, which increases the propulsive thrust at the ankle and enables a quicker recovery at the end of a stride before starting the next one, there are other features that help to increase the speed of striding. The legs of most camels and ruminants have lengthened, especially in the lower parts; the number of toes, or digits, in the feet is reduced from the original mammalian five, and ruminants walk on the tips of their toes. The muscles are inserted high on the legs; only tendons pass lower, so that a large mass is not concentrated near the tip of the limb, where its inertia would restrict speed of movement. Muscle contraction is fast. The movement of each leg is almost limited to a fore-and-aft plane. Emphasis on the fore-and-aft articulations between the limb bones is especially pronounced in many bovids, the alternating bones in the wrist (carpus) and ankle (tarsus) taking the strain of impact on uneven ground.

Pigs have four toes on each foot, but only two of them touch the ground. Their limbs are short and not very advanced. Peccaries have lost the outer accessory hind hoof in the back leg. All four toes of each foot of hippopotamuses touch the ground, and the terminal phalanges have nail-like hoofs. The toe bones of camels are completely enclosed in hardened, horny hoofs, and lateral toes spread across the broad pad which aids in walking on desert sands. Chevrotains have four hoofed toes on each foot; deer often retain the first and second phalanges (sections) of their lateral toes; but all bovids have lost the bones of their lateral toes.

The fibula bone in the back leg and the ulna in the front leg have been reduced in different artiodactyl lineages. Both are still complete in pigs and hippopotamuses, although the fibula is slender. In most other artiodactyls, the lower end of the fibula has survived, and the upper end is occasionally found, but always less noticeably. In camels the ulna has fused with the radius. Pigs, hippopotamuses, and camels have separate navicular and cuboid bones in the ankle, and magnum and trapezoid bones in the wrist; other artiodactyls have a fused naviculo-cuboid and magnum-trapezoid. In chevrotains and some deer, the adjacent ectocuneiform is sometimes joined with the naviculo-cuboid.

Modifications of the foot

The artiodactyl method of limb support through the third and fourth toes, with the attendant lengthening of lower limb bones, has frequently led to a fusion of the two principal metacarpal and metatarsal (midfoot) bones in the forelegs and hindlegs, respectively, forming cannon bones. The nearest approach to a cannon bone in the living Suiformes is the proximal fusion (*i.e.,* at the upper ends) of the two central metatarsals in peccaries. Camels have front and rear cannon bones, but the fusion does not extend right to the bottom, the lower articular surfaces being less pulley-like than in ruminants. There is a hind cannon bone in all chevrotains and, in addition, a front one in Asiatic species (*Tragulus*). All other living artiodactyls have front and rear cannon bones. Lateral metatarsals and metacarpals survive in chevrotains; splints of lateral metacarpals often survive in bovids; and either upper or lower splints of metacarpals in deer.

Modifications of the skin. *Hair and coloration.* Pigs are covered with rather sparse, coarse hairs, and peccaries with a denser coat of coarse hairs. Except for those of the warthog and the babirusa (*Babyrousa babirussa*), piglets have longitudinal stripes or flecks. Hippopotamuses are naked. Tragulids have light-coloured flecks and stripes in their fur. The coats of camelids and deer are much thicker in species living toward the polar regions, at great heights,

or in deserts, but are not noted for striking colours or patterns. Many young deer and the adults of a few species have pale flecks and stripes, and some South American deer have reddish fur. Antelopes have a wider range of coat colours, and some are strikingly marked; *e.g.,* the oryxes, bontebok, and blesbok of southern Africa.

Scent glands. External glands occur in various places on artiodactyls. Preorbital glands, immediately in front of the eyes, are present in the giant forest hog (*Hylochoerus meinertzhageni*), in all cervids except the roe deer, and, among the bovids, in duikers, many neotragines, gazelles and their allies, and the hartebeest group. These glands are apparently required in small forest forms and have disappeared in many, but not all, open-country forms. In some, the glands are definitely connected with territorial marking; a firm object is marked by rubbing, soft vegetation by swinging the head gently from side to side. Foot, or pedal, glands are present in the African bush pig (*Potamochoerus porcus*), camels, tragulids, the pronghorn, some bovids, and on the back legs only of most American deer.

Inguinal (belly) glands are found in bovids, there being two in sheep, saiga, chiru, gazelles, duikers, and blackbuck, and four in members of the tribes Reduncini and Tragelaphini. Carpal (wrist) glands are present in some pigs, some gazelles and allies, and the oribi (*Ourebia ourebi*). Glands in other positions are rather less frequent, but postcornual ones (behind the horns) occur in the Rocky Mountain goat, the pronghorn, and the chamois (*Rupicapra rupicapra*), supraorbital ones in muntjacs (several species of *Muntiacus*). There are jaw glands in the pronghorn; neck glands in camels; dorsal glands on the back of peccaries, pronghorn, and springbok; and preputial glands (in front of the genital region) in several pigs, grysbok (*Raphicerus melanotis*), and the musk deer. Tail glands are found in musk deer, pronghorn, and goats; tarsal glands in pronghorn and American deer; and metatarsal glands in camels, some deer, and the impala. Pronghorn, blackbuck, gazelles, and oribi are thus particularly well equipped with glands. The use of such glands, apart from the use of preorbital glands in some species for territorial marking, is a matter for conjecture. Chital deer, when alarmed, thump the ground several times with their hind feet, which possess glands; the scent remaining on the ground may function as a danger signal. In general, mammals often mark with their glands when they are threatening other individuals of their own species.

Digestive system. The higher artiodactyls feed only on plant matter, which consists largely of cellulose and other carbohydrates and water. This necessitates adaptations of the structure and functioning of the stomach and intestines. Even pigs have enlarged stomachs—they have a pouch near the cardiac orifice (the upper opening) of the stomach—and in peccaries the stomach is more complicated. In hippopotamuses the stomach is divided into four compartments, and micro-organisms ferment food as part of the digestive process. Unlike pigs, hippopotamuses have lost the cecum (a blind pouch) further on in the gut.

Ruminant stomach

In the most advanced ruminants, the much enlarged stomach consists of four parts. These include the large rumen (or paunch), the reticulum, the omasum (psalterium or manyplies)—which are all believed to be derived from the esophagus—and the abomasum (or reed), which corresponds to the stomach of other mammals. The omasum is almost absent in chevrotains. Camels have a three-chambered stomach, lacking the separation of omasum and abomasum; the rumen and reticulum are equipped with glandular pockets separated by muscular walls having sphincters (valves) and glands. The esophagus opens into the rumen, not into the area between rumen and reticulum; these and other differences suggest that camels evolved the ruminating habit independently of the true ruminants. The total stomach of the domestic ox (*Bos taurus*) occupies nearly three-quarters of the abdominal cavity, and, even in medium-sized cattle, the rumen alone can have a capacity of 95 to 285 litres (25 to 75 gallons), having undergone a tremendous growth in early life, with the changeover from a milk diet.

Food taken into the rumen is later regurgitated into the mouth and completely masticated, then swallowed again and passed to the reticulum, omasum, and abomasum. The regurgitation and chewing in the mouth is called rumination.

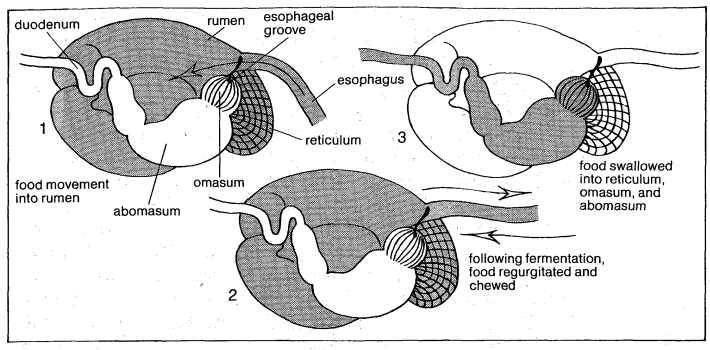


Figure 55: Stages of digestion in the four-part stomach of a representative ruminant.
Drawing by R. Keane

In the rumen many different species of minute protozoans (animals) and bacteria live without free oxygen. The digestion of the cellulose of plant cell walls is the main function of the fauna and flora in the rumen, since mammalian digestive juices are incapable of digesting cellulose. The contents of the plant cells are thus released for digestion. Large volumes of saliva are secreted into the rumen to help digestion. Soluble products of microbial action, mainly fatty acids, are absorbed through the rumen wall. In the omasum, some fatty acids and 60–70 percent of the water are absorbed; in the abomasum gastric juice containing hydrochloric acid is secreted, as in an ordinary mammalian stomach.

In the rumen any ingested protein is degraded into fatty acids and ammonia; the ammonia and other simple nitrogen-containing substances are used by the microorganisms for their own cell-protein synthesis. These organisms are ultimately digested in the abomasum and small intestine, thus providing the ruminant with protein.

Many artiodactyls are adapted to living in conditions of water shortage. The best known and one of the most spectacular examples of this is the camel. Its body temperature can fluctuate according to the outside temperature, thus minimizing water loss through sweating; it excretes rather dry dung and a concentrated urine (*i.e.,* high in urea and low in water) and is not seriously weakened by as much as a 25 percent dehydration in its body, since water is not withdrawn from the bloodstream and the continuing circulation avoids any buildup of excessive internal temperatures. The thick coat hinders the inward transference of heat from the environment (the temperature of which may often exceed the animal's body temperature); a thirsty camel can take in water very rapidly. Oryxes and gazelles are antelopes noted for needing little water, the dorcas gazelle (*Gazella dorcas*) in the Sudan depending on leaves of *Acacia* bushes for its water. The zebu (a form of domesticated cattle) needs less water than most temperate climate breeds.

Reproductive specializations. The testes of male artiodactyls descend outside the body cavity but may regress into the abdomen in the nonbreeding season. Female pigs have many teats, but ruminants have only two to four (although domestic cattle occasionally have as many as six). Among the bovids, the alcelaphines (hartebeests, wildebeests, and relatives), gazelles, and some caprines (sheep, goats, and relatives) have two, the rest have four.

The epitheliochorial and syndesmochorial placentas

The unborn mammal within its mother breathes, feeds, and excretes through an organ called the placenta, which is connected with the tissues of the mother's uterus (womb) wall. Hippopotamuses and pigs have an epitheliochorial placenta, a layer of fetal tissue merely pressed close against the uterus wall, but camels and ruminants possess a syndesmochorial placenta, in which the epithelium of the maternal tissues is eroded to facilitate intercommunication. This is an advance over the epitheliochorial placenta, but the artiodactyls are not particularly advanced, when compared with other mammals, in which there may be still closer association of maternal and fetal blood vessels

(endothelial and hemochorial placentas). Even in many syndesmochorial placentas the uterus lining may be wholly or partly restored before the end of pregnancy. Although there is no erosion of maternal tissues in the epitheliochorial placenta, the capillaries beneath the fetal and maternal surface layers may pass just beneath the surface layers, making them thin. The actual fingerlike processes (villi), through which the placenta contacts the uterus, are evenly distributed ("diffuse" placentas) in hippopotamuses, pigs, camels, and tragulids; in higher artiodactyls they are in pockets or groups called cotyledons ("cotyledonary" placentas). It is interesting that there are few of these cotyledons in deer—for instance only five in Père David's deer—but many in giraffes and bovids (up to 160 or 180 in giraffes and goats). The musk deer (*Moschus moschiferus*) is exceptional among deer in retaining a diffuse placenta.

EVOLUTION AND PALEONTOLOGY

The artiodactyls can be traced back to a probable descent from a group of early generalized mammals called condylarths, and were certainly distinct by the Eocene Epoch, which ended about 38,000,000 years ago. Fossil artiodactyls can be more or less convincingly classified in three suborders; the more primitive Suiformes, centred around pigs, the Tylopoda, centred on camels, and the Ruminantia or ruminants. The most primitive artiodactyls are the suiform group Palaeodonta, which had four functional toes on each foot, primitive, low-cusped cheek teeth, and the typical artiodactyl astragalus. The artiodactyls became more prominent in the Oligocene (between about 38,000,000 and 26,000,000 years ago) with a decline of the then dominant perissodactyls, and the later history of artiodactyls appears as successive waves of groups, each better adapted than its predecessors to the changing environment. In the suiform line, the earlier palaeodonts are succeeded by other groups such as the entelodonts, giant "pigs" of the European and North American Oligocene, characterized by very large skulls (some nearly a metre [three feet] long), very small brains, and a large, bony flange below the eyes. The functionally two-toed ruminants succeeded four-toed suiforms in the Miocene, and within the Old World ruminants of the bovid subfamily Caprinae, the zenith of the tribe Caprini, for example, followed that of the mainly Pliocene tribe Ovibovini.

The earliest artiodactyls

The artiodactyls had an interesting history in North America through the Tertiary Period. Some forms, such as the entelodonts, were shared with the Old World, but others were characteristic of North America. One very prominent New World family was the merycoidodonts (or oreodonts), which lasted until the early Pliocene (about 6,000,000 years ago). They had somewhat piglike proportions, short faces, a large upper canine and a caniniform first lower premolar, and selenodont molars. A close relative, *Agriochoerus,* had clawed feet, the function of which remains uncertain.

Camels and horses in North America

Camelids evolved in North America and, at or toward the end of the Tertiary, spread into South America and into the Old World. By the end of the Pleistocene they all became extinct in their homeland, just as horses did. The hypertragulids were a mainly Oligocene group of chevrotain-like forms related to the Protoceratidae. The latter had horns above their noses, a position unique among artiodactyls, as well as in the usual position. The North American Miocene (26,000,000 to 7,000,000 years ago) produced some ruminants, such as *Blastomeryx,* that are hard to distinguish from the early palaeomerycine relatives of giraffes and deer in the Old World, which, with the North American groups, constitute the family Palaeomerycidae. Some developed horns, and the dromomerycine *Cranioceras* even had a third horn above the back of its skull. During the Miocene and Pliocene there finally appeared relatives of the surviving pronghorn, an example being *Merycodus.* Many of these North American groups have parallels with Old World groups, and the subject of North American artiodactyl evolution is of great interest. Only further finds will indicate whether *Blastomeryx,* the dromomerycines, *Merycodus,* and the pronghorns evolved from hypertragulids already in North America or sprang from some immigrant ruminant and, if the latter, whether the supposed hypertragulid *Leptomeryx* could be such an immigrant ruminant. It is uncertain whether the hypertragulids are nearer the tragulines or the camels, and how close the oreodonts are to the anthracotheres. Of the great New World radiation there survived after the Pleistocene only three or four camelid species and the pronghorn (deer and bovids in the Americas are immigrants), whereas in the Old World as little as 200 years ago, Eurasia and Africa had abundant deer and antelopes.

Until the Miocene there were some archaic artiodactyls in Europe, the xiphodonts, which have cautiously been taken as tylopods, and the cainotheres and anoplotheres, which are classified near anthracotheres.

A possible ruminant ancestor was *Archaeomeryx* from the upper Eocene of China, a small animal that already had a fused naviculo-cuboid bone in the ankle. Tragulids occurred in Africa and Eurasia back to the Miocene, and the more advanced gelocids are known from the upper Eocene and lower Oligocene. At the end of the Oligocene, the first ruminants began to appear with teeth more advanced than those of tragulids. From early in the Miocene they began to be recognizable as giraffes, deer, or antelopes, although the last were relatively uncommon before the late Miocene. Much remains to be learned about the detailed early history of these groups. Several different giraffids lived in later Miocene and early Pliocene times, but the group has since declined to only two species. Deer gradually acquired more complicated antlers, which became very large in some lineages. Different subfamilies of bovids originated in Eurasia and Africa, and it is of zoogeographic interest that representatives of African subfamilies have been found as fossils in northern India and Pakistan.

CLASSIFICATION

Annotated classification. The following classification is principally based on that of American paleontologist George Gaylord Simpson, with alterations in the bovid subfamilies, in the placing of early relatives of giraffes and deer in a giraffoid subfamily Palaeomerycinae, and in the placing of hypertragulids and protoceratids with camels. Groups indicated by the dagger (†) are known only as fossils.

ORDER ARTIODACTYLA

Cloven-hoofed ungulates, the major group of herbivorous mammals. Weight supported mainly through 3rd and 4th toes; astragalus with upper and lower articulations rounded. Stomach compound and, with intestines, enlarged for plant digestion. About 150 species.

Suborder Suiformes

Complete dentition, bunodont (low-cusped) molars, short legs, 4-toed feet are among their important characteristics.

†Infraorder Palaeodonta

Eocene to lower Miocene. Primitive, small-brained artiodactyls. Two superfamilies, Dichobunoidea and Entelodontoidea, with 5 and 3 families, respectively, and collectively about 30 genera.

Infraorder Suina

Lower Oligocene to present. Includes the living pigs, peccaries, and their likely ancestors and extinct relatives.

Family Suidae (pigs). Lower Oligocene to present; Old World. Small to moderate size; shoulder height to about 100 cm (39 in.). Coarse hair. Omnivorous, with sharp-edged tusks. Five Recent and about 22 fossil genera.

Family Tayassuidae (peccaries). Differ from pigs by having 1 fewer incisor and premolar, smaller canines, less advanced cheek teeth; hindleg with a cannon bone, more complicated stomach and more densely haired coat.

Infraorder Ancodonta

†Family Anoplotheriidae. Eocene and Oligocene; Europe; uncertain relationships.

†Family Anthracotheriidae. Eocene to Pleistocene. Mainly Old World, a few in North American Oligocene. Large, with cheek teeth showing beginnings of selenodonty.

Family Hippopotamidae (hippopotamuses). Middle Pliocene to present. Thought to be derived as late as the Pliocene from anthracotheres. Old World, now restricted to Africa. One large (shoulder height to 170 cm [67 in.]; weight to 3,000 kg [6,600

lb]) and one small species (height to 90 cm [35 in.]). Feed on land but frequently resort to water.

†*Family Cainotheriidae.* Mainly Oligocene; small European forms of uncertain relationships.

†*Infraorder Oreodonta*

†*Family Merycoidodontidae* (North American oreodonts). Eocene to early Pliocene. No suppression of upper incisors, an incisiform lower canine, selenodont cheek teeth, short faces and short limbs.

†*Family Agriochoeriidae.* Eocene to lower Miocene. Close to the above family but with clawed feet.

Suborder Tylopoda

Some reduction of upper incisors, reduced premolars, selenodont cheek teeth; cannon bones present. Feet became 2-toed early in the geological history of the group.

†*Family Hypertragulidae.* Upper Eocene to lower Miocene; North America. Like Old World Tragulina (see below) but with a canine-like first lower premolar. Fused naviculo-cuboid in the ankle.

†*Family Protoceratidae.* Oligocene to lower Pliocene; North America. Some with horns on the top and at the front of the skull. Later ones with hindleg cannon bone. Generally considered close to hypertragulids, but failed to fuse navicular and cuboid bones.

Family Camelidae (camels and lamoids). Upper Eocene to present; now a relict group, represented in southern South America, in Asia, and in North Africa. Red blood corpuscles oval. Gallbladder absent. The hump of the 2 Old World camels is composed of fibrous connective tissue and fat.

†*Family Xiphodontidae.* Eocene and lower Oligocene of Europe. Already 2-toed, despite their antiquity, and tentatively placed with the camels.

Suborder Ruminantia (ruminants)

Upper incisors lacking; lower canine incisor-like; cheek teeth selenodont. Fused magnum-trapezoid bone in the wrist. Two-toed feet evolved within suborder.

Infraorder Tragulina

†*Superfamily Amphimerycoidea*

†*Family Amphimerycidae.* European Eocene and Oligocene of Europe; poorly known. Asian *Archaeomeryx,* usually placed in the Hypertragulidae, but may fit here; it retained upper incisors and had a fused naviculo-cuboid in the hind leg.

†*Family Gelocidae.* Eocene-Oligocene of Europe and Asia.

Superfamily Traguloidea

Family Tragulidae (chevrotains). Miocene to present. Sabre-like upper canines in males; incipiently selenodont molars. Bony carapace often develops above the pelvic girdle in males.

Infraorder Pecora

Mostly with horns or antlers and without upper canines. Hollowed auditory bullae. Four-chambered stomach.

Superfamily Giraffoidea

†*Family Palaeomerycidae.* Upper Oligocene to upper Pliocene; North America, Europe, Asia, Africa. Three subfamilies, the Palaeomerycinae (Old World basal stock for giraffes and deer), Blastomerycinae, and Dromomerycinae (the last two New World).

Family Giraffidae (giraffes and okapi). Miocene to present; Old World, now confined to Africa. Living giraffes long-necked and long-legged; the okapi more compact. Giraffes may be up to 5.5 m (18 ft) in total height. Extinct relatives including the large, short-legged, and grotesquely horned sivatheres. Living species have no gallbladder.

Superfamily Cervoidea

Family Cervidae (deer).

Subfamily Moschinae (musk deer). Pleistocene and present; Asia. One recent species, the musk deer (*Moschus moschiferus*) with sabre-like upper canines, gallbladder present (lacking in most other deer).

Subfamily Muntiacinae (muntjacs). Lower Miocene to present; southern Asia. Includes the living muntjacs (*Muntiacus*), with small, 2-pronged antlers above a long, skin-covered base; and tufted deer (*Elaphodus cephalophus*) with tiny antlers. Eurasian fossil genus *Dicrocerus* had larger, 2-pronged antlers. All have large, curved upper canines.

Subfamily Odocoileinae. Pliocene to present. Characterized by the persistence of the lower ends of the lateral metacarpals, thickly haired skin between the hoofs, and in all except the moose a large interdigital gland in at least the hindfoot. The vomer is fused far back posteriorly with the palate in the American deer and in the reindeer. Includes moose (Eurasian elk), reindeer, roe deer, perhaps the Chinese water deer (which has long canines but no antlers), and the deer of North and South America other than the wapiti (American elk), a cervine.

Subfamily Cervinae. Pliocene to present. Top ends of the lateral metacarpals persist. Smooth skin between the hoofs; interdigital glands lacking. Branching of beam of antler differs from that in New World deer; red deer and wapiti, fallow deer, chital, sika, Père David's deer.

Superfamily Bovoidea

Family Antilocapridae (pronghorn and merycodonts). Miocene to present; North America. Smooth, branched horns consisting of hollow sheaths over bony cores; only sheaths shed annually; in fossil Merycodontinae, 1 or more burrs at base of cores. Teeth high-crowned.

Family Bovidae (cattle, sheep and goats, antelopes). Wild cattle, sheep, and goats were ancestral to domestic livestock not differing in any fundamental characters from antelopes. Great variety in horn shape. Many with high-crowned teeth.

Subfamily Bovinae (cattle and some antelopes). Miocene to present. African tribe Tragelaphini, with keeled, spiral horns and not very advanced teeth, includes eland, kudu, nyala, and bushbuck. Tribes Boselaphini and Bovini, mainly Eurasian, former including the Indian nilgai, the 4-horned antelope, and some extinct forms, the latter including cattle, bison, and buffaloes. Bovini are descended from extinct Boselaphini; large size, up to 180 cm (71 in.) at the shoulder. Teeth specialized for grazing. Animals wallow frequently.

Subfamily Cephalophinae (antelopes). Mostly small, with tiny horns set toward the back of the head. Gallbladder lacking. Generally forest-living. African duikers.

Subfamily Hippotraginae (antelopes). Moderate-sized, stocky, mainly grazing antelopes; shoulder height 60–160 cm (24–63 in.). High-crowned teeth. Tribe Hippotragini includes the roan, sable, oryx, and addax antelopes; and Reduncini the reedbuck, kobs, lechwes, and waterbucks. Oryx and addax inhabit arid areas; others near water with adjacent cover or high grass. All except nearly extinct Arabian oryx are now native only to Africa, but the subfamily formerly occurred in India.

Subfamily Alcelaphinae (antelopes). African wildebeests, hartebeests, topis, and several extinct lineages. Long-faced, plains-living, grazing antelopes. Sometimes included in the Hippotraginae.

Subfamily Antilopinae (antelopes). Tribe Neotragini includes some small African antelopes, and Antilopini include gazelles, springbok, and the Indian blackbuck. Graceful, long-legged antelopes of arid, open country. Subfamily may also include the saiga and the Tibetan chiru (tribe Saigini), hitherto classified with Caprinae.

Subfamily Caprinae (sheep, goats, and relatives). Mainly Eurasian. Moderate-sized, shoulder height often 90–100 cm (35–39 in.). High-crowned teeth. Agile animals; many species in mountains or on steep, rocky slopes. Tribe Caprini comprises sheep and goats; Ovibovini the musk-ox (*Ovibos*), the Chinese takin, and some bizarre extinct forms; Rupicaprini the chamois, serow, goral, and Rocky Mountain goat.

Critical appraisal. The great 18th-century classifier Carolus Linnaeus recognized the camels and ruminants as associated but placed some nonartiodactyls with them. It was the French naturalist Henri de Blainville who, at the beginning of the 19th century, first recognized the complete order of artiodactyls as it is accepted today. Nine discrete groups exist among the living forms: pigs, peccaries, hippopotamuses, camels, chevrotains, deer, giraffes, pronghorn, and bovids; their classification presents no great problems, apart from a few genera. Fossils, however, bring confusion to various schemes.

The relationship of North American Tertiary artiodactyls to those of the Old World is a basic question in the study of their zoogeography, history, and classification. It can be agreed that the camels evolved in North America and are as old as the tragulines, which in the Old World were ancestral to ruminants. Most paleontologists today believe that the hypertragulids, protoceratids, and oreodonts were related to the camels, but others have linked the first two groups with the tragulines and the oreodonts with the anthracotheres. Other questions affecting the higher levels of artiodactyl classification are the placing of the North American native ruminants and whether the earliest Old World pecorans should be taken as giraffoids or cervoids.

The subdivisions of the Bovidae remain controversial. Modifying a classification proposed by German zoologist Max Schlosser, some authorities have grouped the Bo-

vinae, Cephalophinae, and Hippotragineae as the Boödontia and the Alcelaphinae, Antilopinae, and Caprinae as the Aegodontia, to indicate phyletic lines believed to have arisen early in bovid history. Boödonts and aegodonts have evolved differently in Africa and Eurasia, and there is much to be said for developing a classification that reflects these geographical relationships. Until additional evidence of phylogenetic relationships is available, however, the modified version of Simpson's classification above will remain favoured by most taxonomists. (A.W.G.)

BIBLIOGRAPHY

Mammalia. The literature on the biology of mammals is vast. The following list includes only a sample of available sources, with emphasis on those in the English language. Standard general references on the mammals are: F.E. BEDDARD, *Mammalia* (1902); W.H. FLOWER and R. LYDEKKER, *An Introduction to the Study of Mammals, Living and Extinct* (1891); and J.Z. YOUNG, *The Life of Mammals* (1957).

Less technical accounts include: F. BOURLIERE, *Vie et moeurs des mammifères* (1951; Eng. trans., *The Natural History of Mammals,* 1954); and L.H. MATTHEWS, *The Life of Mammals* (1970). Widely used textbooks in mammalogy are: E.L. COCKRUM, *Introduction to Mammalogy* (1962); and D.E. DAVIS and F.B. GOLLEY, *Principles in Mammalogy* (1963). The only thorough up-to-date treatment in English of the families of living mammals is S. ANDERSON and J.K. JONES, JR., (eds.), *Recent Mammals of the World: A Synopsis of Families* (1967). E.P. WALKER *et al., Mammals of the World,* 3 vol. (1964), is a semitechnical work including illustrations of representatives of most living genera; the third volume is a classified bibliography. Perhaps the most comprehensive reference on the biology of mammals is P.P. GRASSE (ed.), *Traité de zoologie,* vol. 16–17 (1967–68, 1955).

The morphology and classification of early mammals is reviewed by J.A. HOPSON and A.W. CROMPTON, "Origin of Mammals," *Evolutionary Biol.,* 3:15–72 (1969); and J.A. HOPSON, "The Classification of Nontherian Mammals," *J. Mammal.,* 51:1–9 (1970). A standard reference on the classification of mammals is G.G. SIMPSON, "The Principles of Classification and a Classification of the Mammals," *Bull. Am. Mus. Nat. Hist.,* 85:1–350 (1945); the evolution of major mammalian groups is treated by A.S. ROMER in *Vertebrate Paleontology,* 3rd ed. (1966), and *Notes and Comments on Vertebrate Paleontology* (1968). See also JASON A. LILLEGRAVEN, *et al.,* (eds.), *Mesozoic Mammals: The First Two-Thirds of Mammalian History* (1979).

Basic references on the status of endangered species of mammals are: F. HARPER, *Extinct and Vanishing Mammals of the Old World* (1945); and G.M. ALLEN, *Extinct and Vanishing Mammals of the Western Hemisphere with the Marine Species of All the Oceans* (1942).

Summaries of specialized topics in mammalogy include: H.T. ANDERSEN (ed.), *The Biology of Marine Mammals* (1969); S.A. ASDELL, *Patterns of Mammalian Reproduction,* 2nd ed. (1964); L.S. CRANDALL, *The Management of Wild Mammals in Captivity* (1964); R.F. EWER, *Ethology of Mammals* (1968); C.P. LYMAN and A.R. DAWE (eds.), *Mammalian Hibernation* (1960); W.V. MAYER and R.G. VAN GELDER (eds.), *Physiological Mammalogy,* 2 vol. (1963–64); A.G. SEARLE, *Comparative Genetics of Coat Colour in Mammals* (1968); GEORGE G. SIMPSON, *Splendid Isolation* (1980), a study of South American mammals.

Some widely known journals that publish papers dealing exclusively with mammals are: *Journal of Mammalogy, Mammalia,* and *Säugetierkundliche Mitteilungen* (all issued quarterly), and *Zeitschrift für Säugetierkunde* (6/yr.). Additional technical literature is published by university and public museums and government agencies. Semitechnical and popular accounts of mammals appear in a wide variety of publications of state game departments, museums, and zoological gardens.

Monotremata. The following are popular works on monotremes: H. BURRELL, *The Platypus* (1927); D. FLEAY, *We Breed the Platypus* (1944), and "Observations on the Breeding of the Platypus in Captivity," *Victorian Nat.,* 61:8–14 (1944); M. GRIFFITHS, *Echidnas* (1968). Works of a more technical nature are: M.L. AUGEE, E.H.M. EALEY, and H. SPENCER, "Biotelemetric Studies of Temperature Regulation and Torpor in the Echidna, *Tachyglossus aculeatus,*" *J. Mammal.,* 51:561–570 (1970); J.H. CALABY, "The Platypus (*Ornithorhynchus anatinus*) and Its Venomous Characteristics," in W. BUCHERL, E.E. BUCKLEY, and V. DEULOFEU (eds.), *Venomous Animals and Their Venoms,* 1:15–29 (1968); W.K. GREGORY, "The Monotremes and the Palimpsest Theory," *Bull. Am. Mus. Nat. Hist.,* 88:1–52 (1947); K.W. ROBINSON, "Heat Tolerances of Australian Monotremes and Marsupials," *Aust. J. Biol. Sci.,* 7:348–360 (1954); K. SCHMIDT-NIELSEN, T.J. DAWSON, and E.C. CRAWFORD, JR., "Temperature Regulation in the Echidna (*Tachyglossus aculeatus*)," *J. Cell. Comp. Physiol.,* 67:63–72 (1966); H.M. VAN DEUSEN and G.G. GEORGE, "Results of the Archbold Expeditions, no. 90, notes on Echidnas (Mammalia, Tachyglossidae) of New Guinea," *Am. Mus. Novit.* no. 2383 (1969).

Marsupialia. Among general works are the following: CHARLES L. BARRETT, *Wild Life of Australia and New Guinea* (1954); CHARLES W. BRAZENOR, *The Mammals of Victoria and the Dental Characteristics of Monotremes and Australian Marsupials* (1950); ALBERT S. LE SOUEF and HARRY BURRELL, *The Wild Animals of Australasia, Embracing the Mammals of New Guinea and the Nearer Pacific Islands* (1926); BASIL J. MARLOW, *Marsupials of Australia* (1962); ELLIS TROUGHTON, *Furred Animals of Australia,* 8th ed. rev. (1966); ERNEST P. WALKER *et al., Mammals of the World,* 3 vol. (1964; 2nd ed., vol. 1, 1968); F. WOOD JONES, *The Mammals of South Australia,* 3 vol. (1923–25).

Journal and magazine articles include: J. PEARSON, "Some Problems of Marsupial Phylogeny," *Rep. Meet. Aust. N.Z. Ass. Advmt. Sci.,* 25:71–102 (1947); H.C. REYNOLDS, "Studies on Reproduction in the Opossum (*Didelphus virginiana virginiana*)," *Univ. Calif. Publs. Zool.,* 52:223–284 (1952); G.G. SIMPSON, "The Affinities of the Borhyaenidae," *Am. Mus. Novit.,* no. 1118 (1941); and "The Beginning of the Age of Mammals in South America," *Bull. Am. Mus. Nat. Hist.,* vol. 91, art. 1 (1948); G.H.H. TATE, "On the Anatomy and Classification of the Dasyuridae (Marsupialia)," *Bull. Am. Mus. Nat. Hist.,* vol. 88, art. 3 (1947); "Studies on the Anatomy and Phylogeny of the Macropodidae (Marsupialia)," vol. 91, art. 2 (1948); and "Studies in the Peramelidae (Marsupialia)," vol. 92, art. 6 (1948); D. FLEAY, "Strange Animals of Australia," *Natn. Geogr. Mag.,* 124:388–411 (1963); J.H. CALABY, "Australia's Threatened Mammals," *Wildlife,* 1:15–18 (1963); H.H. FINLAYSON, "Mitchell's Wombat in South Australia," *Trans. R. Soc. S. Aust.,* 85:207–215 (1961); A.G. LYNE, "Australian Mammals," *Aust. Mus. Mag.,* 12:121–125 (1956); J. MCNALLY, "Koala Management in Victoria," *Wildl. Circ. Vict.,* no. 4 (1957); W.E. POOLE and P.E. PILTON, "Reproduction in the Grey Kangaroo, *Macropus canguru,* in Captivity," *C.S.I.R.O. Wildl. Res.* 9:218–234 (1964); G.B. SHARMAN, "Studies on Marsupial Reproduction. III. Normal and Delayed Pregnancy in *Setonix brachyurus,*" *Aust. J. Zool.,* 3:56–70 (1955); G.B. SHARMAN and J.H. CALABY, "Reproductive Behaviour in the Red Kangaroo, *Megaleia rufa,* in captivity," *C.S.I.R.O. Wildl. Res.,* 9:58–85 (1964); E.M.O. LAURIE and J.E. HILL, "List of Land Mammals of New Guinea, Celebes and Adjacent Islands, 1758–1952," *Br. Mus. Nat. Hist.* (1954).

Insectivora. A. CABRERA, *Genera mammalium,* vol. 1, *Insectivora,* vol. 2, *Galeopithecia* (1919–25), a Spanish classic—a still unsurpassed illustrated review of living insectivora; PETER CROWCROFT, *The Life of the Shrew* (1957), a review of the biology of the common Eurasian *Sorex araneus;* J.F. EISENBERG and E. GOULD, "The Behaviour of *Solenodon paradoxus* in Captivity with Comments on the Behavior of Other Insectivora," *Zoologica,* 51:49–58 (1966); "The Tenrecs: A Study in Mammalian Behavior and Evolution," *Smithson. Contr. Zool.,* no. 27 (1970), a technical but easily read review of the biology of this family; F.G. EVANS, "The Osteology and Relationships of the Elephant Shrews (Macroscelididae)," *Bull. Am. Mus. Nat. Hist.,* 80:85–125 (1942), a technical comparison of elephant ws, tree shrews, erinaceids, and lemuroids; GILLIAN GODFREY and PETER CROWCROFT, *The Life of the Mole* (*Talpa europaea Linnaeus*) (1960), a readable, semi-popular account of the common European mole; E. GOULD, N.C. NEGUS, and A. NOVICK, "Evidence for Echolocation in Shrews," *J. Exp. Zool.,* 156:19–38 (1964); K. HERTER, "Das Verhalten der Insektivoren," *Handb. Zool.,* vol. 8, sect. 9, pp. 1–50 (1957), a German language review of the behaviour of all insectivores; M.W. LYON, "Treeshrews: An Account of the Mammalian Family Tupaiidae," *Proc. U.S. Natn. Mus.,* vol. 45 (1913); S.B. MCDOWELL, "The Greater Antillean Insectivores," *Bull. Am. Mus. Nat. Hist.,* 115:117–214 (1958), a technical review of the relationships of solenodons to other insectivores.

Chiroptera. G.M. ALLEN, *Bats* (1939, reprinted 1962), a comprehensive view by a naturalist; K.C. ANDERSEN, *Catalogue of the Chiroptera in the Collection of the British Museum,* 2nd ed., vol. 1, *Megachiroptera* (1912), definitive taxonomy of the Megachiroptera; R.W. BARBOUR and W.H. DAVIS, *Bats of America* (1970), comprehensive coverage of North American bats, distribution, natural history, photographs, and taxonomy; A. BROSSET, *La Biologie des Chiroptères* (1966), the most complete, up-to-date review of all bats; R.G. BUSNEL (ed.), *Animal Sonar Systems,* 2 vol. (1967), the most recent important symposium on echolocation and related systems; R.B. DAVIS, C.F. HERREID II, and H.L. SHORT, *Mexican Free-Tailed Bats in Texas* (1962), the most complete published survey of a given species of bats; M. EISENTRAUT, *Aus dem Leben der Fledermäuse und Flughunde* (1957), a review of European bats by an important student of bat natural history and physiology; K. FAEGRI and L. VAN

DER PIJL, *The Principles of Pollination Ecology* (1966), the role of bats in pollination; D.R. GRIFFIN, *Listening in the Dark: The Acoustic Orientation of Bats and Men* (1958), the most recent authoritative summary of echolocation; *Echoes of Bats and Men* (1959), a paperback introduction to echolocation; A. NOVICK and N. LEEN, *The World of Bats* (1970), a recent review of bat behaviour and ecology with action photographs; L. VAN DER PIJL, *Principles of Dispersal in Higher Plants* (1969), the role of bats in dispersal of seeds; D.R. ROSEVEAR, *The Bats of West Africa* (1965), an important compilation on the bats of a particular range; H. SAINT GIRONS, A. BROSSET, and M.C. SAINT GIRONS, "Contribution à la connaissance du cycle annuel de la Chauve-souris *Rhinolophus ferrum-equinum* (Schreber, 1774)," *Mammalia,* 33:357–470 (1969), an extensive study of the biology of one European bat species; J. VERSCHUREN, *Ecologie, biologie, et systématique des Cheiroptères* (1957), the primary work on the ecology of bats; B. VILLA-R., *Los murcielagos de Mexico* (1966), an important compilation of knowledge on the bats of Mexico; W. WIMSATT (ed.), *The Biology of Bats* (1971), the definitive current work on bat biology.

Primates. *General works:* J.R. and P.H. NAPIER, *A Handbook of Living Primates: Morphology, Ecology and Behaviour of Nonhuman Primates* (1967), a comprehensive, technical book with one or more photographs of each genus; W.C. OSMAN HILL, *Primates: Comparative Anatomy and Taxonomy,* 8 vol. (1953–70), also technical and heavily illustrated, treating the primates in greater depth; W.E. LE GROS CLARK, *The Antecedents of Man,* 3rd ed. (1971), a readable scientific account of nonhuman primates; F. WOOD JONES, *Arboreal Man* (1916), a classic work that contains much useful information; I.T. SANDERSON, *The Monkey Kingdom* (1957), an extensively illustrated popular book; I. DEVORE (ed.), *Primate Behavior* (1965); and C.R. CARPENTER, *Naturalistic Behavior of Nonhuman Primates* (1964), two technical works on primate ethology; W.W. HOWELLS, *Mankind in the Making,* rev. ed. (1967), an informative book on human evolution for the general reader; S. ZUCKERMAN, *Functional Affinities of Man, Monkeys, and Apes* (1933), a technical dissertation on primate structure; IAN TATTERSALL, *The Primates of Madagascgar* (1982), a survey of many lower primates; and G.A. DOYLE, *The Study of Prosimian Behavior* (1979).

Tree shrews: W.E. LE GROS CLARK, "The Myology of the Tree-Shrew (*Tupaia minor*)," 1924:461–497; "On the Brain of the Tree-Shrew (*Tupaia minor*)," 1924:1053–1074; "On the Skull of *Tupaia,*" 1925:559–567; and "On the Anatomy of the Pen-Tailed Tree Shrew (*Ptilocercus lowii*)," 1926: 1179–1309, all in the *Proc. Zool. Soc. Lond.* (1924–26), four important papers on tree shrews; R.D. MARTIN, "Treeshrews: Unique Reproductive Mechanism of Systematic Importance," *Science,* 152:1402–1404 (1966); and M.W. SORENSON and C.H. CONAWAY, "Observations on the Social Behaviour of Tree Shrews in Captivity," *Folia Primatologica,* 4:124–145 (1966), two technical papers.

Tarsioids: W.E. LE GROS CLARK, "Notes on the Living Tarsier (*Tarsius spectrum*)," *Proc. Zool. Soc. Lond.,* 1924:217–223 (1924); and H.H. WOOLLARD, "The Anatomy of *Tarsius spectrum,*" *ibid.,* 1925:1071–1184 (1925), technical papers concerned with the relationships of tarsiers; H. SPRANKEL, "Untersuchungen an *Tarsius.* I. Morphologie des Schwanzes nebst ethologischen Bermerkungen," *Folia Primatologica,* 3:153–188 (1965), a summary, in German, of information on the taxonomy of tarsioids.

Lemuroids: J.J. PETTER, *Recherches sur l'écologie et l'éthologie des Lémuriens malgaches* (1962); "Ecological and Behavioral Studies of Madagascar Lemurs in the Field," *Ann. N.Y. Acad. Sci.,* 102:267–281 (1962); and "The Lemurs of Madagascar," in I. DEVORE (ed.), *Primate Behavior* (1965), technical works on a variety of lemurs; A. JOLLY, *Lemur Behavior* (1967), a readable, semipopular account, based on personal observations; A. WALKER, "Patterns of Extinction Among the Subfossil Madagascan Lemuroids," in P.S. MARTIN and H.E. WRIGHT (eds.), *Pleistocene Extinctions* (1967), a technical discussion of lemur evolution.

New World monkeys: R.I. POCOCK, "On the External Characters of the South American Monkeys," *Proc. Zool. Soc. Lond.,* 1920:91–113 (1920); and C.R. CARPENTER, "Behavior of Red Spider Monkeys in Panama," *J. Mammal.,* 16:171–180 (1935), two technical articles; L.A. ROSENBLUM and R.W. COOPER (eds.), *The Squirrel Monkey* (1968), a detailed account of the biology of one species; A. CABRERA, *Catalogo de los mamiferos de America del Sur* (1957), a broad treatment, in Spanish, of the taxonomy and distribution of South American monkeys; P. HERSHKOVITZ, "Mammals of Northern Colombia; Preliminary Report No. 4: Monkeys (Primates), with Taxonomic Revisions of Some Forms," *Proc. U.S. Natn. Mus.,* 98:323–427 (1951), a scientific treatment of some South American forms.

Old World monkeys: R.I. POCOCK, "The External Characters of the Catarrhine Monkeys and Apes," *Proc. Zool. Soc., Lond.,* 1925:1479–1579 (1925); and "The Monkeys of the Genera *Pithecus* (or *Presbytis*) and *Pygathrix* Found to the East of the Bay of Bengal," *ibid.,* 1934:895–961 (1934), two articles providing much technical information on the langur; C. G. HARTMAN and W.L. STRAUSS (eds.), *The Anatomy of the Rhesus Monkey, Macaca mulatta* (1933 and 1961), a detailed account of the morphology of this important species; J.R. and P.H. NAPIER (eds.), *Old World Monkeys* (1970), a technical but comprehensive account of the catarrhines; N.C. TAPPEN, "Problems of Distribution and Adaptation of the African Monkeys," *Curr. Anthrop.,* 1:91–120 (1960), a scientific account of the zoogeography of the catarrhines.

Anthropoid apes: C.F.M. SONNTAG, *Morphology and Evolution of Apes and Man* (1924), a comprehensive and technical book, with many illustrations; R.M. and A.W. YERKES, *The Great Apes* (1929), a classic work, written for the general reader; C.R. CARPENTER, *A Field Study in Siam of the Behavior and Social Relations of the Gibbon (Hylobates lar)* (1941, reprinted 1967); G.B. SCHALLER, *The Mountain Gorilla* (1963); and J. VAN LAWICK-GOODALL, *In the Shadow of Man* (1971), three readable accounts of the lives of individual species, based on extensive field work; V. REYNOLDS, *The Apes* (1967); and R. and D. MORRIS, *Men and Apes* (1966), well-illustrated popular books; C.P. GROVES, "Population Systematics of the Gorilla," *J. Zool.,* 161:287–300 (1970); and *Gorillas* (1970), respectively, a technical and popular work on this interesting ape; W.K. GREGORY, *The Anatomy of the Gorilla: The Studies of Henry Cushier Raven, and Contributions by William B. Atkinson and Others* (1950), a carefully illustrated, technical book; TED CRAIL, *Apetalk and Whalespeak: The Quest for Interspecies Communication* (1982).

Edentata. The following works, although broad in scope, are particularly rich in information on this group: J.C. BARLOW, "Edentates and Pholidotes," in S. ANDERSON and J.K. JONES, JR. (eds.), *Recent Mammals of the World* (1967), a discussion of the taxonomy of this order and its families; J. DORST, *South America and Central America: A Natural History* (1967), many fine photographs of edentates; E.R. HALL and K.R. KELSON, *The Mammals of North America,* 2 vol. (1959), descriptions and distributional information on North and Central American members of this order; R. HOFFSTETER, "Xenarthra," and R. SABAN, "Palaeanodonta," in J. PIVETEAU (ed.), *L'Origine des Mammifères evolution,* 2 vol., pt. 6 of *Traité de paléontologie* (1958), detailed accounts of early edentates; and E.P. WALKER, *Mammals of the World,* 2nd ed., vol. 1 (1968), an account of the natural history and distribution of every genus in the order.

Lagomorpha. J.N. LAYNE, "Lagomorphs," in *Recent Mammals of the World,* ed. by S. ANDERSON and J.K. JONES (1967), general review of characters, distribution, and habits of living lagomorphs; E.P. WALKER *et al., Mammals of the World,* 2nd ed. (1968), photographically illustrated articles on living genera of lagomorphs; R.M. LOCKLEY, *The Private Life of the Rabbit: An Account of the Life History and Social Behaviour of the Wild Rabbit* (1964), semipopular work dealing with the rabbit in Britain and the disease myxomatosis; H.E. BROADBOOKS, "Ecology and Distribution of the Pikas of Washington and Alaska," *Am. Midl. Nat.,* 73:299–335 (1965), discussion of habits and habitats of North American pikas; M.R. DAWSON, "Lagomorph History and the Stratigraphic Record," in *Essays in Paleontology and Stratigraphy,* ed. by C. TEICHERT and E.L. YOCHELSON (1967), review of fossil lagomorphs and their usefulness for stratigraphic studies; A.A. GUREEV, "Lagomorpha," in *Mammals,* vol. 3, no. 10, *Fauna S.S.S.R.,* Zoologicheskii Institut Akademii Nauk S.S.S.R., New series no. 87 (1964), comprehensive treatment of the morphology and relationships of fossil and living lagomorphs (in Russian); E.R. HALL, "A Synopsis of the North American Lagomorpha," *Univ. Kans. Publs. Mus. Nat. Hist.,* 5:121–202 (1951), on the taxonomy and distribution of North American forms; A.S. LOUKASHKIN, "On the Pikas of North Manchuria," *J. Mammal.* 21:402–405 (1940), on the behaviour and habitats of Asian pikas.

Rodentia. J.R. ELLERMAN, *The Families and Genera of Living Rodents,* 2 vol. (1940, reprinted 1966), a scientific catalog of existing rodents; C.S. ELTON, *Voles, Mice and Lemmings* (1942, reprinted 1965), a popular, historical record of plagues of various rodents, followed by an analysis of the lemming population fluctuations in northern Labrador; P.P. GRASSE and P.L DEKEYSER, "Ordre des Rongeurs," in *Traité de zoologie,* vol. 17, pt. 2, pp. 1321–1525 (1955), a section from a zoological reference work dealing with the anatomy, ecology, habits, and classification of rodents, with brief descriptions down to the generic level (in French); T.G. HULL, *Diseases Transmitted from Animals to Man,* 5th ed. (1963), descriptions of diseases, with discussion of causative factors and carriers, many of which are rodents; D. MACCLINTOCK, *Squirrels of North America* (1970), an account of the ecology, habits, and relationships of North American squirrels; H.G.Q. ROWETT, *The Rat as a Small*

Mammal, 2nd ed. (1965), a student laboratory manual with detailed notes on the dissection of the rat; M. SHORTEN, *Squirrels* (1954), a study of the red and gray squirrels in Britain for the general reader; G.D. SNELL (ed.), *Biology of the Laboratory Mouse,* 2nd ed. (1966), a standard reference work on all of the aspects of mouse raising; B.S. VINOGRADOV and A.I. ARGIROPULO, *Key to Rodents* (1968; orig. pub. in Russian, 1941), a handbook on the rodents in the U.S.S.R. and their economic importance; E.P. WALKER *et al., Mammals of the World,* 3 vol. (1964), descriptions at the generic level of all living mammals, including every recognized genus of rodent, usually with pictures; L. WILSSON, *Bäver* (1964; Eng. trans., *My Beaver Colony,* 1968), a popular account of beaver behaviour; S.E. WOODS, *The Squirrels of Canada* (1980), a guide to 22 species.

Carnivora. For general information on the biology of the carnivores, see E.P. WALKER *et al., Mammals of the World,* 3 vol. (1964), in which each genus is described and illustrated, along with a brief summation of its biology. The taxonomy of carnivores is discussed in G.G. SIMPSON, "Principles of Classification and a Classification of Mammals," *Bull. Am. Mus. Nat. Hist.,* vol. 85 (1945), a classic work on classification, followed by most recent mammalogists; and H.J. STAINS, "Carnivores and Pinnipeds," in S. ANDERSON and J.K. JONES, JR. (eds.), *Recent Mammals of the World: A Synopsis of Families* (1967). F.E. BEDDARD, *Mammalia* (1902), is a definitive early work on mammalian anatomy. The paleontology of the Carnivora is summarized in two works by A.S. ROMER: *Vertebrate Paleontology,* 3rd ed. (1966), basic to an understanding of fossil forms, and *Notes and Comments on Vertebrate Paleontology* (1968), containing additional information not found in the general text. Books devoted to particular subgroups of the carnivores include: A. DENIS, *Cats of the World* (1964), an excellent summary of the status of all members of the Felidae; C.J. HARRIS, *Otters* (1968), a fine summary of the status of the otters of the world; H.E. HINTON and A.M.S. DUNN, *Mongooses: Their Natural History and Behavior* (1967), which contains much interesting information difficult to find, except in scattered literature; R.J. HARRISON *et al.* (eds.), *The Behavior and Physiology of Pinnipeds* (1968), an excellent summation of recent knowledge of these aquatic carnivores; and V.B. SCHEFFER, *Seals, Sea Lions, and Walruses: A Review of the Pinnipedia* (1958), a major work on the taxonomy of the Pinnipedia.

Cetacea. G.M. ALLEN, "The Whalebone Whales of New England," *Mem. Bost. Soc. Nat. Hist.,* 8:107–322 (1916), contains much valuable information on the baleen whales; HARALD T. ANDERSEN (ed.), *Biology of Marine Mammals* (1969), a modern review of physiology, anatomy, and behaviour of cetaceans; F.C. FRASER, *Report on Cetacea Stranded on the British Coasts from 1927 to 1932* (1934), *... 1933 to 1937* (1946), and *... 1938 to 1947* (1953), information on North Atlantic cetaceans; *Handbook of R.H. Burne's Cetacean Dissections* (1952), a good source on whale anatomy; RICHARD J. HARRISON and JUDITH E. KING, *Marine Mammals* (1965), a simple guidebook on most aspects of cetaceans; R. KELLOGG, "The History of Whales—Their Adaptation to Life in the Water," *Q. Rev. Biol.,* 3: 29–76, 174–208 (1928); and "A Review of the Archaeoceti," *Publs. Carnegie Instn. 482* (1936), important papers on the evolution and adaptation of cetaceans; W.N. KELLOGG, *Porpoises and Sonar* (1961), a preliminary synthesis of studies on porpoise echolocation; N.A. MACKINTOSH and J.F.G. WHEELER, "Southern Blue and Fin Whales," *'Discovery' Rep.,* 1: 257–540 (1929); L. HARRISON MATTHEWS, "The Humpback Whale, *Megaptera nodosa,*" *ibid.,* 17:7–92 (1937); and "The Sei Whale, *Balaenoptera borealis,*" *ibid.,* 17:183–289 (1938), three papers that deal with economically important species; J.R. NORMAN and F.C. FRASER, *Giant Fishes, Whales, and Dolphins,* new ed. (1948), an excellent review of the cetacea; K.S. NORRIS (ed.), *Whales, Dolphins, and Porpoises* (1966), a readable account of the overall biology of the group; DALE W. RICE and VICTOR B. SCHEFFER, "A List of the Marine Mammals of the World," *Spec. Scient. Rep. U.S. Fish Wildl. Serv. (Fisheries No. 579)* (1968), a systematic listing of all modern cetaceans; P.F. SCHOLANDER, "Experimental Investigations on the Respiratory Function in Diving Mammals and Birds, Hvalräd Skr., 22:1–131 (1940); and P.F. SCHOLANDER and W.E. SCHEVILL, "Counter-Current Vascular Heat Exchange in the Fins of Whales," *J. Appl. Physiol.,* 8:279–282 (1955), technical papers that deal with special adaptations of cetaceans; E.J. SLIJPER, *Die Cetaceen* (1936, in Dutch) and *Walvissen* (1958; Eng. trans., *Whales,* 1962), comprehensive popular books on whales; V.B. SCHEFFER, *The Year of the Whale* (1969), a hypothetical narrative of the life history of the sperm whale, with factual information on other species; A.G. TOMILIN, *Cetacea,* vol. 9 of *Mammals of the U.S.S.R. and Adjacent Countries* (1967; orig. pub. in Russian, 1957); FREDERICK W. TRUE, "Contributions to the Natural History of the Cetaceans: A Review of the Family *Delphinidae,*" *Bull. U.S. Natn. Mus. 36* (1889), an early synthesis of cetacea that still has much validity; ERNEST P. WALKER *et al., Mammals of the World,* vol. 2, *Cetacea,* 2nd ed. (1968), photographs and natural history data of most modern cetaceans. LYALL WATSON, *Sea Guide to Whales of the World* (1981), is a reference and field guide to whales, dolphins, and porpoises.

Proboscidea. L.S. DE CAMP, *Elephant* (1964), suitable for specialist and nonspecialist alike; RICHARD CARRINGTON, *Elephants: A Short Account of Their Natural History, Evolution and Influence on Mankind* (1958), an excellent study of living as well as extinct forms; P.E.P. DERANIYAGALA, *Elephas maximus, the Elephant of Asia* (1951), with much information on the Asian elephant; and *Some Extinct Elephants, Their Relatives, and the Two Living Species* (1955), reviewed by G. Gaylord Simpson, a remarkable miscellany of elephant lore and observation, ancient and recent; HENRY F. OSBORN, *Proboscidea* (1939), a comprehensive pioneer work; A.S. ROMER, *Vertebrate Paleontology,* 3rd ed. (1966), a *sine qua non* for students of proboscidean evolution; IVAN T. SANDERSON, *The Dynasty of Abu: A History and Natural History of the Elephants and Their Relatives, Past and Present* (1962), a good all-around work.

Sirenia. G.M. ALLEN, "Extinct and Vanishing Mammals of the Western Hemisphere, with the Marine Species of All the Oceans," *Spec. Publs. Am. Comm. Int. Wildl. Prot.,* no. 11 (1942), provides information on the causes of extinction and the basis for conservation of sirenians; C. BERTRAM, *In Search of Mermaids: The Manatees of Guiana* (1963), is a readable popular account of the biology of manatees; C.K. and G.C.I. BERTRAM, "The Sirenia as Aquatic Meat-Producing Herbivores," *Symp. Zool. Soc. (London),* no. 21, pp. 385–391 (1968), contains an extensive bibliography of the literature on the Sirenia.

Perissodactyla. Information on the ungulates as a group is contained in J. SIDNEY, "The Past and Present Distribution of Some African Ungulates," *Trans. Zool. Soc. Lond.,* vol. 30 (1965), a comprehensive account; and J.R. ELLERMAN and T.C.S. MORRISON-SCOTT, *Checklist of Palaearctic and Indian Mammals, 1758 to 1946,* 2nd ed. (1966).

The following papers are concerned with the biology of certain perissodactyls. W. VON RICHTER, "Untersuchungen über angeborene Verhaltensweisen des Schabrackentapirs (*Tapirus indicus*) und des Flachlandtapirs *(Tapirus terrestris),*" *Zool. Beitr.* Neue Folge, 12:67–159 (1966), one of the few detailed studies of tapirs; C.A.W. GUGGISBERG, *S.O.S. Rhino* (1966), a readable general account of the living rhinoceroses, their biology and conservation; H. KLINGEL, "Soziale Organisation und Verhalten freilebender Steppenzebras," and "Soziale Organisation und Verhaltensweisen von Hartman und Bergzebras," *Z. Tierpsychol.,* 24:580–624 and 25:76–88 (1967–68), descriptions of the behaviour and social organization of zebras; G.G. SIMPSON, *Horses: The Story of the Horse Family in the Modern World and through 60 Million Years of History* (1961), an interesting account of the natural history and evolution of the Equidae. MYRON J. SMITH, *Equestrian Studies* (1981), is a comprehensive, classified bibliography of more than 4,600 equestrian studies in English.

Artiodactyla. I.W. CORNWALL, *Bones for the Archaeologist* (1956), on the morphology and identification of bones including artiodactyls; J. DORST and P. DANDELOT, *A Field Guide to the Larger Mammals of Africa* (1970), summarized descriptions, habits, ecology, and distribution maps for all African artiodactyls; R.F. EWER, *Ethology of Mammals* (1968), a comprehensive text with much recent information on artiodactyl behaviour; V. GEIST, "The Evolution of Horn-Like Organs," *Behaviour,* 27:175–214 (1966), on the functional implications of horn shape; G.G. SIMPSON, "The Principles of Classification and a Classification of Mammals," *Bull. Am. Mus. Nat. Hist.,* vol. 85 (1945), forms the basis for most modern classifications of artiodactyls; T. HALTENORTH in W. KUKENTHAL and T. KRUMBACH, *Handbuch der Zoologie,* vol. 8, pp. 1–167, *Klassifikation der Säugetiere: Artiodactyla* (1963), a weighty classification of artiodactyls (in German); V.G. HEPTNER, A.A. NASIMOVIC, and A.G. BANNIKOV (eds.), *Die Säugetiere der Sowjetunion,* vol. 1, *Paarhufer und Unpaarhufer* (1966), a massive work on Eurasian ungulates, most of it dealing with artiodactyls, originally published in Russian in 1961; A. KEAST, "Comparisons of the Contemporary Mammalian Faunas of the Southern Continents," *Q. Rev. Biol.,* 44:121–167 (1969), a review of zoogeography and adaptations, with further references; P.S. MARTIN and H.E. WRIGHT (eds.), *Pleistocene Extinctions: The Search for a Cause* (1967), a collection of essays on Pleistocene extinctions involving artiodactyls; D. MORRIS, *The Mammals* (1965), a general account with illustrations; G.B. SCHALLER, *The Deer and the Tiger* (1967), a study of the life of Indian artiodactyls; C.A. SPINAGE, *The Book of the Giraffe* (1968), much information well presented for the general reader; W.P. TAYLOR (ed.), *The Deer of North America* (1956), all aspects of the life of North American deer; J.Z. YOUNG, *The Life of Vertebrates,* 2nd ed. (1962), contains a chapter on artiodactyls; F.E. ZEUNER, *A History of Domesticated Animals* (1963), a useful source for much information difficult to find elsewhere.

Manchester

Manchester remains an important regional city of Great Britain, dominating much of northwestern England, but it has lost that extraordinary vitality and unique influence that made it such a phenomenon of the Industrial Revolution. For Manchester was an urban prototype: in many respects it could claim to be the first of the new generation of huge industrial cities created in the Western world during the past 250 years. In 1717 it was merely a market town of 10,000 people, but by 1851 its textile (chiefly cotton) industries had so prospered that it had become a manufacturing and commercial city of more than 300,000 inhabitants, already spilling out its suburbs and absorbing its industrial satellites. By the beginning of the 20th century, salients of urban growth linked Manchester to the ring of cotton-manufacturing towns—Bolton, Rochdale, and Oldham, for example—that almost surround the city, and a new form of urban development, a conurbation, or metropolitan area, was evolving. By 1911 it had a population of 2,350,000. In the following years, however, the pace of growth slowed dramatically. If the 19th century—when it was indisputably Britain's second city—was Manchester's golden age, the 20th century has been marked by increasing industrial problems associated with the decline of the textile trades (the result of foreign competition and technological obsolescence).

This article is divided into the following sections:

Physical and human geography 503
The landscape 503
The city site
Climate
Architecture and the face of the city
The people 504
The economy 504
Industry
Trade and transportation
Administrative and social conditions 505
Government
Education and social services
Cultural life 505
History 506
Early settlement and medieval growth 506
Evolution of the modern city 506
Bibliography 506

Physical and human geography

THE LANDSCAPE

The city site. Manchester occupies a featureless plain made up of river gravels and the glacially transported debris known as drift. It lies at a height of 133 feet (40 metres) above sea level, enclosed by the slopes of the Pennine range on the east and the upland spur of Rossendale on the north. Much of the plain is underlain by coal measures; mining was once widespread, but only a single large pit remains, at Agecroft. Within this physical unit, known as the Manchester embayment, the conurbation of southeastern Lancashire and northeastern Cheshire (often referred to by the acronym SELNEC) evolved. Manchester, the central city (now combined with other districts to form the metropolitan area of Greater Manchester), is situated on the east bank of the River Irwell and has an elongated north–south extent, the result of late 19th- and early 20th-century territorial expansion. In 1930 the city extended its boundaries far to the south beyond the River Mersey, to annex nine square miles (23 square kilometres) of northern Cheshire. Two large districts adjoin the district of Manchester on the west and southwest: the district of Salford and the district of Trafford. Together these three administrative units form the chief concentration of commercial employment. From this core, suburbs have spread far to the west and south, chiefly within the county of Cheshire. To the north and east of Manchester smaller industrial towns and villages, mixed with suburban development, merge into one another and extend as a continuous urban area to the foot of the encircling upland. Close to the upland margin lies a ring of large towns, which were traditionally the major centres of the cotton-spinning industry—Bolton, Bury, and Rochdale on the north and Oldham, Ashton-under-Lyne, and Stockport on the east.

The structure of the conurbation

The urban structure of metropolitan Manchester is largely determined by its industrial zones. By far the most important of these is the one bisecting it from east to west. This contains most of the heavier industry—petrochemicals on the Ship Canal near Irlam, electrical engineering in Trafford Park and Salford, and machine tools and metal fabrication in eastern Manchester. Industry in the south is confined to a few compact, largely planned factory estates, notably at Altrincham and Wythenshawe. North and east of Manchester ribbons of long-established industry follow every railway, river valley, and abandoned canal. The electrochemical industries of the Irwell Valley, the dyestuffs of the Irk, and, everywhere, the old textile mills (many converted to new industrial uses) are the dominant features.

Climate. Manchester's climate is most kindly described as mild, moist, and misty. The temperate climate is without extremes: winters are mild, with a January mean temperature of 39° F (4° C), and summers are cool, with a July mean temperature of 59° F (15° C). Occasional high-pressure systems produce cold, clear spells in winter or hot droughts in summer, but these rarely persist. Winds from the west and south prevail, and these bathe the city in frequent gentle rain derived from the almost constant succession of Atlantic weather systems. Annual rainfall is not notably high by the standards of western Britain (32 inches [818 millimetres]), but it occurs on no less than half of the days in an average year. There is little reliable seasonal variation, but the months of March through May offer the best chance of prolonged dry spells.

The wet Atlantic air banked against the Pennine slopes to the east of the city produces extreme cloudiness; on about 70 percent of the days of the year the afternoon sky is at least half covered by cloud. This limits sunshine, which was further reduced during the decades of the city's industrial prosperity by air pollution. Up to about 1960 the city centre recorded the abnormally low total of only 970 sunshine hours annually. Foul fogs were another problem of the man-made industrial climate. Manchester then had an average of 55 days of serious fog in a typical year, and the death rate from respiratory diseases surged following these fog episodes. But the city's peculiarly sunless and fog-bound winter climate was transformed by effective air-pollution control. Annual hours of bright sunshine have risen to about 1,300, and serious fogs have been reduced to about 20 days each year. This has been a major factor in reducing the incidence of two formerly endemic diseases, bronchitis and tuberculosis, which had given the city an unenviably high death rate.

Architecture and the face of the city. Manchester's extraordinary 19th-century wealth left a permanent record in an architectural variety and virtuosity that makes the city centre an outdoor museum of styles from Greek classical to early tall steel-framed structures. Commercial firms vied to commission the best architects to design offices and warehouses of ornate splendour, and the public buildings were intended to outshine London's. Thus, banks occupied Greek temples or turreted Gothic castles and warehouses were given the facades of Venetian palaces. The offices of the Ship Canal Company were given a Grecian colonnade perched high above street level, and the Town Hall, de-

signed by Alfred Waterhouse, is regarded as perhaps the ultimate in Victorian Gothic fantasies.

Conserving this priceless architectural heritage has presented great problems. Many of the buildings are protected landmarks but are unsuited to modern commercial needs, though some imaginative conversions have taken place. The Royal Exchange, once the hub of the textile trade, contains as the old trading floor the largest room in Europe; it now houses a freestanding theatre-in-the-round. The old Central Station, a huge glazed train shed, has been converted into an exhibition centre. A complex of buildings at Castlefield, including the world's oldest railway station, has been developed as a regional museum of science and industry.

A wave of office redevelopment in the 1960s and '70s added many steel-and-glass structures to the Manchester skyline. One of the earliest is Manchester's tallest building, the Co-operative Insurance Society tower, at 400 feet (122 metres).

As new shopping centres began to develop in outlying areas, the level of retail trade in the city centre suffered. This led to the development of a large enclosed shopping precinct, the Arndale Centre, which contains a significant proportion of the total retail activity in the city centre. As it grew, however, older shopping streets suffered by the shift of businesses, so that parts of the city core have a run-down, half-abandoned appearance; but this is part of the process by which the Victorian central business district is reshaping itself to meet modern needs.

THE PEOPLE

Greater Manchester is one of the world's most compact and crowded metropolitan areas. The overcrowded conditions explain the chief demographic trend of recent years, that of population loss by out-migration. Manchester city itself lost almost one-third of its population to migration between 1961 and 1981, which was one of the highest rates of migrational loss among all British cities. Natural increase is below the national average, for the migration is chiefly of young families of child-bearing age, leaving an older population in the core cities. Thus overall population decline is serious. This trend is also widespread in the other old industrial towns of the conurbation.

Much of this migration is to suburban areas, though there is also an interregional loss of population to more prosperous areas of Britain, and the "dormitory" districts of the fringes (and especially Cheshire to the south) are growing strongly. Thus, the metropolitan area is decentralizing quickly, and its overall population trend is more favourable than those of its major constituent cities. Total metropolitan population has been virtually stable since 1961, with the low rate of natural increase being entirely offset by net out-migration.

Increasingly, families living in decaying substandard housing have been rehoused. Manchester has exported population to "overspill" estates at Middleton and Hyde, and Salford families have moved to Worsley. All of these are large schemes, involving population transfers of at least 10,000, and all lie within the conurbation. There also has been movement to the New Town project at Warrington, a major development point on the Ship Canal, 18 miles (29 kilometres) west of Manchester. Within the city there has been massive redevelopment. The Hulme scheme of the early 1970s involved the rehousing of a population of almost 60,000.

Like many British cities Manchester experimented in the 1960s with high-rise housing to accommodate families from the slum clearance zones. In the past, row houses had been the traditional housing form in low-income areas of the inner city, and the new high-rise schemes proved to be a social failure—some were demolished within a decade of construction. The emphasis of city planning was shifted from total clearance and replacement of old housing to its conservation and improvement through Housing Action Areas. Thus, old housing is given new life and the community is kept together: where new dwellings are built they are the modern equivalent of the traditional row houses.

This out-migration has been partly counterbalanced by an in-migration from Commonwealth countries, particularly of persons from the West Indies and the Indian subcontinent. Manchester itself has a multiracial immigrant community, which is chiefly concentrated in the Moss Side area. Some of the textile towns, too, have attracted Commonwealth immigrants, chiefly Indian and Pakistani textile workers. The metropolitan area as a whole has been one of the main magnets to Commonwealth immigrants in Britain.

Commonwealth in-migration

THE ECONOMY

Industry. There has long been a contrast between the economies of the core city (Manchester itself, together with the industrial areas of Salford and Stretford) and the textile towns that form the northern and eastern margins of the urban cluster. Until the 1960s the latter had narrowly based economies largely dependent on the textile trade, which still provided more than half the employment of women. The former, however, had an economy of greater diversity: manufacturing was varied (including printing and the production of engineering and electrical products, chemicals, and clothing), and a broad range of service activities gave stability to the economy. This old pattern of contrast was breaking down in the late 20th century, as the core city lost factory employment at a rapid rate and became increasingly dependent on services, while the peripheral towns acquired greater industrial diversity and thus a securer (and locally expanding) manufacturing base.

The entire metropolitan area of Greater Manchester has undergone major economic changes. The textile industry has been reduced to a mere vestige of the enormous manufacture that once underpinned the economy of the city. It continues to decline, despite diversification from cotton to man-made fibres and resultant close links with the chemical industry. The surviving mills have been reequipped for high productivity, but this, too, has had the effect of reducing labour demand. The clothing industry has declined with the textile industry but has remained a significant employer of women, chiefly in many small workshops in the inner city. Much more serious has been the sharp contraction of more modern industries that, until the 1970s, had served as replacements for the old industries. The decline in engineering, one of the main sources of jobs for men, is especially serious. Within the chemical industry, the main growth has been in the production of fine chemicals and pharmaceuticals, with research laboratories located in parkland at Alderley, on the southern fringe. The paper and printing industry is stable, reflecting Manchester's status as the second centre, after Greater London, of newspaper production in England.

Decline of manufacturing

Manchester's economy has been moving from an industrial to a postindustrial nature. Services have become the chief employers, with the "thinking" rather than the manual services undergoing expansion. Some services, such as transport and distribution, are declining, but the professions, finance and banking, administration, and general personal services are growing with explosive force. Most of these growth points require well-qualified workers: the declining demand for manual skills and the shift to mental skills have caused selective unemployment, which is clearly a persistent social problem.

The conversion of Manchester into a service city is not an entirely new trend, since the city has been the regional capital of northwestern England for two centuries. The process, however, has been quickened by the rapid decline of industry in the inner city. Clearance of the slum tracts and their subsequent redevelopment have removed entire urban districts that once housed many hundreds of small firms. Nearly half of the employment once available in manufacturing in the inner areas has disappeared. In these districts, a disadvantaged and ethnically mixed community experiences unemployment rates that are at least twice the city average.

Part of this loss of factory work in the inner city has been the result of the movement of firms to the fringe of the urban area, not only to planned industrial estates but also to the cotton mills left empty by the decline of the textile trades. Hundreds of mills have been converted to other uses, thereby providing the cheap factory floorspace necessary to young and struggling firms, so that the textile

towns have in some degree replaced the inner city as an industrial nursery in which it is possible for new firms to become established.

Trade and transportation. Apart from its massive volume of retail and wholesale trade, Manchester has a number of distinctions as a regional service centre. It houses a branch of the Bank of England and the Northern Stock Exchange, the headquarters of the Co-operative Wholesale Society, and one of the major provincial crown courts. Its airport at Ringway, 10 miles south of the city, is the leading British terminal outside of London in the volume of international traffic handled and in the diversity of both its European and transatlantic services. Ringway is owned by the city and is the country's second airfreight terminal.

From 1894 to 1986 Manchester was a seaport, with a group of docks at the head of the 37-mile Ship Canal. The growth in the size of shipping, together with changes in the pattern of maritime trade, led to a slow decline in the use of the waterway, and by the mid-1980s the upper parts had been closed to traffic. The lower reaches of the canal remained open and busy, serving the needs of bankside industries, especially the huge oil-refining and chemicals complex at Ellesmere Port. New industrial and commercial uses for the derelict terminal docks have been developed.

Public transport in Greater Manchester is coordinated by a Passenger Transport Executive, and it relies heavily on an integrated system of bus routes. The system faces private competition, however, especially from flexible minibus services. A dense network of commuter rail services, largely electrified, is managed in partnership with British Rail.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. Though the conurbation of metropolitan Manchester is a single cohesive socioeconomic unit, its local government was fragmented until 1974. The dominant unit was the county borough of Manchester, which assumed the financial burden of supplying central facilities (major museums and libraries, the airport) for the area as a whole. There were six other county boroughs, each independent and able to develop its own social, educational, and planning policies. Another 45 lesser authorities (municipal boroughs and urban districts) shared the administration of this fragmented conurbation, parts of which lay within the areas of the three counties of Lancashire, Cheshire, and Derbyshire.

Greater Manchester

The Local Government Act of 1972 (in effect from 1974) created a metropolitan county of Greater Manchester, defining an area rather larger than the conurbation by including the mining town of Wigan to the west. This new urban entity was divided into the 10 metropolitan districts of Manchester, Bolton, Bury, Oldham, Rochdale, Salford, Stockport, Tameside, Trafford, and Wigan. The county administered a number of general services (*e.g.*, strategic planning, transport, and recreation), while the districts handled the main range of services (*e.g.*, education, housing, and most personal and household services). The metropolitan county of Greater Manchester was abolished in 1986, and some of the general services it had provided were taken over by specialist successor authorities.

Education and social services. Of all Manchester's pioneer cultural achievements, none has prospered more than the Victoria University of Manchester. After its foundation in 1851 at a site in Quay Street, the college received a charter in 1872 and began growth on its present site in 1873. By 1880 it had combined with member colleges in both Leeds and Liverpool to form a federal institution. Since becoming a separate body again in 1903, the university has grown to become one of the largest in Britain. The faculty of technology has become autonomous as an Institute of Science and Technology, and, with the establishment of the University of Salford in 1967 and the growth of a large polytechnic, there are now four institutions of higher learning in and near the city.

The city provides the complete range of social and welfare services within the British system, but its special strength lies in health services and medical education. The Victoria University of Manchester has the largest medical school in western Europe; it is linked to three large groups of teaching hospitals that provide specialist treatment. One of the most distinguished of these is the Christie Hospital, a major centre for cancer research.

CULTURAL LIFE

In some respects the cultural life to be found in modern Manchester is less rich than formerly. For example, its prestigious newspaper, *The Guardian,* has (in the Mancunian view) fled to London and dropped the city's name from its title. Music, on the other hand, maintains its strength. The Hallé concerts reached their centenary in 1958, and the orchestra continues to maintain its international reputation.

The city has a large number of private, public, and specialized libraries. The municipal library, with more than 25 branches, has its headquarters at St. Peter's Square. Manchester also houses the notable John Ryland University Library (now part of the Victoria University of Manchester library) and Chetham's Library, one of the first free public libraries in Europe. Libraries

Among the galleries and museums, the Whitworth Art Gallery and the Manchester City Art Gallery are partic-

The Guardian, Manchester and London



Manchester skyline viewed from the mathematics building of Manchester University Institute of Science and Technology.

ularly well known. The latter contains a fine collection of paintings, sculpture, silver, and pottery and is supplemented by several branch galleries. The Manchester Museum has special exhibits of Egyptian and Japanese objects, as well as natural history collections and an aquarium.

There are two major association football (soccer) clubs, and at the grounds of the Old Trafford Cricket Club, test matches are played against overseas cricket teams visiting Great Britain.

History

EARLY SETTLEMENT AND MEDIEVAL GROWTH

Early in the Roman conquest of Britain a fort was established (AD 78–86) on a low sandstone plateau at the confluence of the Rivers Medlock and Irwell. In its first form, in the late 1st and early 2nd centuries, the fort was a simple field fortification of shallow ditches, earth banks, and timber palisades. By the early 3rd century the fort had been rebuilt in stone and contained a number of substantial buildings. Despite the difficulties of excavating the site—later covered by Victorian buildings and partly destroyed by railway construction—evidence of substantial activity of the period has been discovered. A *vicus* (Latin: "row of houses") of merchants and craftsmen had grown outside the walls, along the well-made road to York. But Roman occupation left no permanent imprint, except to give the modern city its name, derived from Mamucium ("Place of the Breastlike Hill"). There is no evidence of occupation after the 4th century, and the site seems to have lain empty for 500 years. In 919 the West Saxon king Edward the Elder sent a force to repair the Roman site as a defense against the Norsemen, and some traces of this reoccupation have been discovered. By then, however, the growth of Manchester had recommenced almost a mile from the fort, at the junction of the Rivers Irk and Irwell near the present cathedral.

Early growth

The Norman barony of Manchester was one of the largest landholdings in Lancashire, and its lords, the Grelley family, built a fortified hall close to the church. During the 13th century Manchester began its transition from village to town; corn and fulling mills (the latter for cloth) were established, and both a weekly market and an annual fair were held. Sometime before 1301 a charter was granted; there were then about 150 burgesses (adult male citizens) in the town. Although the parish served by the church was a huge one and the infant Manchester was acquiring regional importance, it was subordinate to its near neighbour, Salford, which was the capital manor of the hundred (district) and which had an earlier borough charter. The full development of the medieval borough followed the establishment in 1421 of a college of priests to take charge of the church. Part of the college survives as Chetham's Hospital, while a free church school set up in 1506 became the Manchester Grammar School in 1515, founded by Hugh Oldham, bishop of Exeter.

EVOLUTION OF THE MODERN CITY

By the 16th century Manchester was a flourishing market borough important in the wool trade, exporting cloth to Europe via London during the reign of Elizabeth I. The antiquary John Leland, who visited the town in 1538, thought it one of the "best builded" in the country. By 1620 a new industrial era had begun with the weaving of fustian, a cloth with a linen warp but a cotton weft. This was the origin of the cotton industry that was to transform southern Lancashire after 1770. As the trade grew, Manchester expanded and "improvements" were added, including the fine square and church of St. Ann (1712).

From the 1760s onward, growth quickened with the onset of the Industrial Revolution. The first canal, bringing cheap coal from Worsley, reached the town in 1762; later extended, it linked Manchester with the Mersey and Liverpool by 1776, and so served the import–export needs of the cotton industry. Manchester's first cotton mill was built in the early 1780s. By 1800 Manchester was said to be "steam mill mad," and by 1830 there were 99 cotton-spinning mills. The world's first modern railway, the Liverpool and Manchester, was opened in 1830, and by the 1850s the greater part of the present railway system of the city was complete. Despite its growth to a population of more than 70,000 by 1801, the town had no system of government and was still managed, like a village, by a manorial court leet (a court held semiannually by the lord of the manor or his steward to conduct local government). A police force was established in 1792, but not until 1838 did a charter of incorporation set up an elected council and a system of local government.

The industrial era

Manchester's economic history during the second half of the 19th century was one of growth and diversification. The city became less important as a cotton-manufacturing centre than as the commercial and financial nucleus of the trade; on the floor of the Royal Exchange the yarn and cloth of the entire industry was bought and sold. From an early textile-machinery industry many specialized types of engineering developed. Products included steam engines and locomotives, armaments, machine tools, and, later, those of electrical engineering. The opening of the 37-mile Manchester Ship Canal (1894) linked Manchester, via the Mersey Estuary at Eastham, to the Irish Sea and the world markets beyond. By 1910 Manchester had become the fourth port of the country, and alongside the docks, at Trafford Park, the first (and still the largest) industrial estate in Britain was developed. New industries also took sites there, with a prominent role played by such U.S. companies as Westinghouse and Ford, the latter moving to Essex in 1929. At its height more than 50,000 workers were accommodated within factories of the estate, though that number later declined.

The Manchester of the 19th century was a city of enormous vitality not only in its economic growth but also in its political, cultural, and intellectual life. Its prosperous and influential middle class pioneered the development of the amenities and institutions of a great modern city. *The Manchester Guardian* became Britain's leading provincial newspaper, achieving international influence, while the Hallé Orchestra was its equal in the world of music. Owens College (now known as Victoria University of Manchester) became the nucleus of the first and largest of the great English civic universities, while the academic success of the Manchester Grammar School made it something of a model in the development of selective secondary education in England. Politically, Victorian Manchester often led the nation: in the agitation for parliamentary reform and for free trade its influence was crucial. The Peterloo Massacre of 1819 arose from a peaceful political assembly, held on fields near the city, to demand parliamentary reform. A cavalry charge to disperse the assembly left 11 dead and hundreds wounded. In the period 1842–44 the German social philosopher Friedrich Engels lived in Manchester, and his influential book *Condition of the Working Class in England* (1845) was based on his experiences there. Among its other intellectual achievements were John Dalton's development of the atomic theory as the foundation of modern chemistry and the work of the "Manchester school" in the application of economic principles to the problems of commerce, industry, and government.

19th-century achievement

There was a price to be paid for this precocious growth; in its urban fabric, inner Manchester remained essentially a 19th-century city, and by the late 20th century it faced massive redevelopment problems. An industrial collar of obsolescent factory zones encircled the city centre, and huge areas of old slum housing survived with little renewal into the 1960s. Manchester, then, is a city in transition: its face is being transformed by redevelopment, and its dependence on the insecure base of the textile industries is declining with the growth of a broader economic structure.

20th-century problems

BIBLIOGRAPHY. Historical studies include ALAN J. KIDD and K.W. ROBERTS (eds.), *City, Class, and Culture: Studies of Social Policy and Cultural Production in Victorian Manchester* (1985); JOHN H.G. ARCHER (ed.), *Art and Architecture in Victorian Manchester* (1985); GARY S. MESSINGER, *Manchester in the Victorian Age* (1985), a popular work; NICHOLAS J. FRANGOPULO (ed.), *Rich Inheritance: A Guide to the History of Manchester* (1962, reissued 1969), particularly strong as a survey of sources for the city's history; D.A. FARNIE, *The Manchester Ship Canal and the Rise of the Port of Manchester, 1894–1975* (1980), on the development of the port and its regional impact; and SHENA D. SIMON, *A Century of City Government: Manchester,*

1838–1938 (1938), a readable and well-informed account of the development of local government in the city. For Manchester in its regional context, see C.F. CARTER (ed.), *Manchester and Its Region* (1962), a collection of articles on the geology, history, geography, industry, and cultural life; T.W. FREEMAN, H.B. RODGERS, and R.H. KINVIG, *Lancashire, Cheshire and the Isle of Man* (1966), a general geography of northwestern England; L.P. GREEN, *Provincial Metropolis* (1959), the city and conurbation as a major urban region; and H.P. WHITE (ed.), *The Continuing Conurbation: Change and Development in Greater Manchester* (1980), an analysis of social and economic conditions.

(H.B.Ro.)

Manila

Manila, the national capital and chief city of the Republic of the Philippines, is the centre of the country's economic, political, social, and cultural activity. It is located on Luzon Island and it spreads along the eastern shore of Manila Bay at the mouth of the Pasig River. The city's name, originally Maynilad, is derived from that of the *nilad* plant, a flowering shrub adapted to marshy conditions, which once grew profusely along the banks of the river; the name was shortened first to Maynila and then to its present form. The city proper encompasses an area of approximately 15 square miles (38 square kilometres). In 1975, by presidential decree, Manila and its contiguous cities and municipalities were integrated to function as a single administrative region, known as Metropolitan Manila (also called the National Capital Region), with an area of 246 square miles.

Manila has been the principal city of the Philippines for four centuries, and it is the centre of its industrial development as well as the international port of entry. It is situated on one of the finest sheltered harbours of the Far East, about 700 miles (1,100 kilometres) southeast of Hong Kong. The city has undergone rapid economic development since its destruction in World War II and its subsequent rebuilding; it is now plagued with the familiar urban problems of pollution, traffic congestion, and overpopulation. Measures have been taken, however, to ameliorate these problems.

This article is divided into the following sections:

Physical and human geography 507
  The landscape 507
    The city site
    Climate
    Plant and animal life
    The city layout
  The people 509
  The economy 509
    Industry
    Commerce and finance
    Transportation
  Administration and social conditions 509
    Government
    Public utilities
    Health and security
    Education
  Cultural life 510
History 510
Bibliography 510

Physical and human geography

THE LANDSCAPE

The city site. Manila lies on the eastern shore of Manila Bay, a large inlet with access to the sea through a channel 12 miles wide to the southwest. It occupies the low, narrow deltaic plain of the Pasig River, which flows northwestward to Manila Bay out of a large lake, Laguna de Bay, southeast of the city. Manila Bay lies to the west, the swampy delta of the southward-flowing Pampanga River to the north, the mountains of the Bataan Peninsula to the west, and Laguna de Bay to the southeast. Although the city's area is constricted, it is an excellent port site because of its sheltered harbour, its access to inland agricultural areas by way of the river, and its relative proximity to the Asian mainland.

Climate. The city is protected from extreme weather conditions by the hills of the Eastern Cordillera to the east and by the mountains of Bataan Peninsula, which lies west of Manila Bay. The tropical climate is characterized by a wet season that lasts from June to November and by a dry season lasting from December to May. High humidity and thunderstorms are common in July, August, and September, when more rain is received than in other months. The average annual rainfall totals about 82 inches (2,080 millimetres). There is little monthly variation from the mean annual temperature of 81° F (27° C).

Plant and animal life. The city is dotted with tropical trees, including the palm, banyan, and acacia; bamboo is common in the public parks. Water buffalo, horses, dogs, pigs, and goats are common. The wealth of birdlife includes shrikes, doves, and pigeons, and Manila Bay abounds with sardines, anchovies, mackerel, tuna, snappers, and barracuda. The city's natural beauty is marred, however, by air and water pollution caused by the expansion of industry and the growing number of motor vehicles.

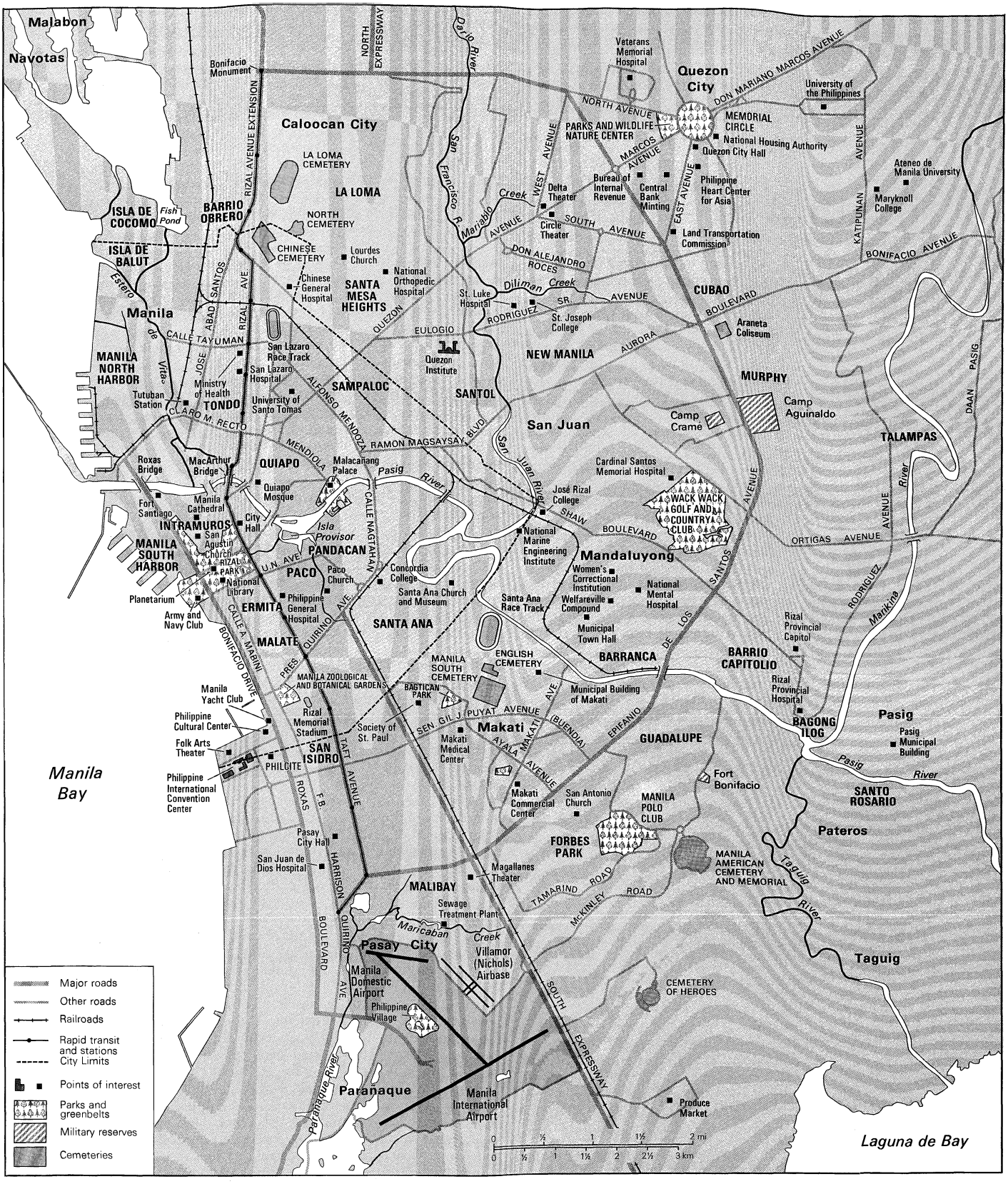
The city layout. The city is bisected by the Pasig River. It is divided into four administrative divisions comprised of 14 districts. The districts developed from the original fortress city of Intramuros (Within Walls) and the 13 villages located outside its walls. The districts of Tondo, Santa Mesa, Binondo, Santa Cruz, Quiapo, San Miguel, and Sampaloc lie to the north of the river and Singalong, Intramuros, Ermita, Malate, Paco, Pandacan, and Santa Ana to the south. The two sections of the city are connected by several bridges.

City districts

Although business areas are widespread, the districts of Tondo, Binondo, Santa Cruz, and Quiapo constitute the chief centres of trade and commerce. San Miguel is the site of Malacañang Palace, the presidential residence; several universities are located in Sampaloc. The heavily populated Tondo district on the northern shore is the site of Manila North Harbor, the local port, while the international port, Manila South Harbor, is on the southern shore. Intramuros is renowned for its 16th-century San Agustin church as well as for the ruins of its old walls and of Ft. Santiago. Ermita and Malate are choice residential districts and the sites of hotels and embassies; Paco, Pandacan, and Santa Ana are middle-income residential areas.

Metropolitan Manila includes the cities of Manila, Caloocan City to the north, Quezon City to the northeast, and Pasay City (located near the shore of Manila Bay) to the south and 13 municipalities. The municipalities include Makati, Mandaluyong, San Juan, Las Piñas, Malabon, Navotas, Pasig, Pateros, Parañaque, Marikina, Muntinlupa, Tagig, and Valenzuela. Metropolitan Manila was created in order to provide integrated services such as water supply, police and fire protection, and transport and to permit central planning for simultaneous and unified development.

Housing. The city has a chronic housing shortage, and tenement housing projects have been constructed by the government to help house the poor. One of these is the Bagong Lipunan Improvement of Sites and Services (BLISS), initiated by the governor of Metropolitan Manila. To provide homes for squatters, the government also has developed resettlement projects around Manila that are easily accessible by land motor transportation.



Manila with its adjoining municipalities.

Types of dwellings

Residential buildings include the single-family dwelling; the duplex for two independent households; the *accessoria,* whose dwelling units have individual entrances from the outside; the apartment building with common entrance; and the *barong-barong,* a makeshift shack built of salvaged materials (flattened tin cans, scrap lumber, cartons, or billboards) that is common in the poor area of Tondo.

Architecture. Architectural styles reflect American, Spanish, Chinese, and Malay influences. Rizal Park and a number of government buildings were designed by U.S. architect and city planner Daniel H. Burnham. Modern buildings—including multistoried commercial houses and public and private buildings—are commonly made of reinforced concrete and hollow cement blocks. Houses of

modern design—especially low, sprawling ranch houses with spacious lawns—are common in the districts of Ermita and Malate. Spanish-style houses, with tiled roofs, barred windows, and thick walls, were common before World War II and have remained popular. The churches of the city are American, Spanish, or European in character. The Manila cathedral was rebuilt in the 1950s and is an important landmark. It succeeds five earlier cathedrals—the first dating from the mid-16th century—that were destroyed either by earthquakes or during wartime.

THE PEOPLE

Metropolitan Manila is densely populated and contains a significant proportion of the population of the country. This concentration of people has been brought about by a constant rural–urban migration. The strain on municipal services has an adverse effect on the quality of life in the urban area. In an effort to ameliorate the situation, the government has adopted a "back-to-the-farm" policy and has established resettlement projects outside of the central urban area.

Almost all the residents of Manila are Filipinos. The largest single foreign community, representing less than 10 percent of the population, is made up of Chinese. The population of the city is predominantly Roman Catholic, although there are some Protestants, Muslims, and Buddhists. The two national Christian churches—the Iglesia ni Kristo and the Philippine Independent, or Aglipayan, Church—have small congregations.

THE ECONOMY

Industry. The diverse manufacturing activities of Manila include textile production, publishing and printing, food processing, and the production of paints, drugs, aluminum articles, rope and cordage, shoes, cigars and cigarettes, coconut oil, soap, and lumber. Factories generally are small and are located mostly in the congested districts of Tondo (which also has the railroad and truck terminals), Binondo, and Santa Cruz. Heavy industries are located in the districts of Paco, Pandacan, and Santa Ana.

Commerce and finance. Manila is the centre of commerce and finance in the Philippines. Trade flourishes within the metropolitan area and between the city and the provinces and other countries. Most of the Philippines' imports and exports pass through the port of Manila. Financial institutions headquartered in Manila include such establishments as the Development Bank of the Philippines, the Philippine National Bank, the Philippine Veterans Bank, the Government Service Insurance System, the Social Security System, and many private commercial and developmental banks. Private insurance companies and the Manila Stock Exchange also contribute to the mobilization of savings for investment.

Transportation. Within the area of Metropolitan Manila, public transportation is provided principally by buses, jeepneys (small buses built on the chassis of jeeps), and taxis. An elevated rail line, linking Caloocan City and the city of Baclaran (to the south of Pasay City), was completed in 1984. It is the first phase of a transit system, called Metrorail, that eventually will extend throughout the metropolitan area. Bus services operate routes to northern and southern Luzon; railroad services operated by the Philippine National Railways also connect the city with northern and southeastern Luzon. Interisland and international transportation is provided by domestic and foreign airlines and by shipping.

Traffic problems

Traffic congestion is serious, traffic tending to pile up at the bridges during the morning and evening rush hours. Adjacent towns serve as dormitory suburbs, and many people commute to the city, adding to the traffic problem.

The main international port is Manila South Harbor, enclosed by a low breakwater. The customhouse and several warehouses and sheds are enclosed. There are no railway lines within the port area, and cargo is transported from the piers by trucks or barges. Several heavy industries that depend upon imported raw materials are located within the port area.

The piers of the local port, Manila North Harbor, are congested with heavy traffic from all ports in the Philippines. It has several warehouses for storage of goods and equipment. Additional port facilities for international shipping have been built, partially on reclaimed land, in the area between the two harbours.



Monument (centre) to José Rizal, national hero of the Philippines, in Rizal Park, Manila.
Tsutomu Nakayama—Shostal Assoc.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. With the organization of Metropolitan Manila in 1975, the mayor–council type of government was superseded by a manager–commission type. Metropolitan Manila, which is vested with the powers and attributes of a corporation (including the power to make contracts, sue and be sued, hold, transfer, and dispose of property, and similar powers), is administered by the Metropolitan Manila Commission. It is headed by a governor as the general manager, who executes policies and measures approved by the commission and is responsible for the discharge of management functions. The other members are the vice governor, as deputy general manager, and three commissioners or board members, one each for planning, finance, and operations. All members are appointed by the president of the Philippines.

The Metropolitan Manila Commission

Supplanting the city or municipality boards or councils is the Sangguniang Bayan (Municipal Assembly), created for each city or municipality, which helps the metropolitan government in administration and legislation. It is composed of the mayor, vice mayor, councillors, captains of *barangays* (neighbourhoods), and representatives from other sectors, appointed by the president upon recommendation of the local unit. The mayors are the presiding officers in their respective areas of jurisdiction.

Public utilities. Potable water comes from a supply network managed by the Metropolitan Waterworks and Sewerage System. Satisfactory sanitation conditions are maintained by constant surveillance of markets, restaurants, motion-picture houses, recreation halls, and slaughterhouses. Insecticides are sprayed regularly on open sew-

ers, uncollected garbage, and standing water; garbage is collected by a fleet of trucks that operate night and day. Moreover, workers maintain cleanliness in the metropolitan area and are also responsible for the beautification of the city as directed by the governor of Metropolitan Manila.

Health and security. Health facilities in Manila are among the best in the Far East. The city government maintains numerous health centres as well as San Lazaro Hospital, where patients are treated free of charge, and subsidizes a number of government hospitals. There are also many missionary and private hospitals in the city.

Police and fire services

Police and fire services are well organized, improved equipment and techniques are used, and personnel are comparatively well paid. To intensify the campaign against crime, more police precincts were created, and the Police Community Relations District was organized. *Barangay* brigades and *barangay tanods* (guards) were also created in each *barangay* of every city and municipality in Metropolitan Manila. They are volunteers and selected leaders of the *barangay* whose duty is to maintain peace and order in their community.

Education. More than 96 percent of the population of 10 years of age or older is literate. More than 100 free public schools are maintained, in addition to the night vocational and secondary schools and the city-supported University of Manila. Educational opportunities are also provided for handicapped children, orphans of school age, and adults. As the education centre of the country, Metropolitan Manila houses many of the major institutions of higher education of the country, including the University of the Philippines (with its main campus in Quezon City), the Philippine Normal College, and the Technological University of the Philippines. There are several universities sponsored by religious bodies, including the University of Santo Tomas (founded in 1611), as well as nonsectarian institutions such as the University of the East and the Far Eastern University.

Manila's universities

CULTURAL LIFE

The centre of the performing arts in the country is the Philippine Cultural Center. There is also the Folk Arts Theater, facing Manila Bay, the renovated historic Metropolitan Theatre, and an open-air theatre in Rizal Park. The many libraries and museums include the National Library and the National Museum, known for its anthropological and archaeological exhibits; the National Institute of Science and Technology, with a scientific reference library and large collections of plants and animals; the geological museum of the Bureau of Mines and Geosciences; the Planetarium; Ft. Santiago, which houses original works of the Philippine patriot José Rizal; and the Kamaynilaan (Manila City) Library and Museum, which contains valuable carvings, paintings, and archives.

The city's parks

The foremost outdoor recreational area is Rizal Park, with a Japanese garden, a Chinese garden, an open-air theatre, a playground, a grandstand, and a long promenade adjacent to Manila Bay. Other areas include the Manila Zoological and Botanical Gardens, the Mehan Garden, and Paco Park. Athletic facilities include the Rizal Memorial Stadium and the Jai-Alai Fronton, both located in Manila, and the Araneta Coliseum in Quezon City. Annual festivals and carnivals are held in the sunken garden fronting the City Hall of Manila.

History

In the late 16th century, Manila was a walled Muslim settlement whose ruler levied customs duties on all commerce passing up the Pasig River. Spanish conquistadors under the leadership of Miguel López de Legazpi—first Spanish governor general of the Philippines—entered the mouth of the river in 1571. They destroyed the settlement and founded the fortress city of Intramuros in its place. Manila became the capital of the new colony. Outside the city walls stood some scattered villages, each ruled by a local chieftain and each centred on a marketplace. As Spanish colonial rule became established, churches were built near the marketplaces, where the concentration of population was greatest. Manila spread beyond its walls, expanding north, east, and south, linking together the market–church complexes as it did so.

The Spanish period

The propagation of Roman Catholicism began with the Augustinian friar Andrés de Urdaneta, who accompanied the expedition of 1571. He was followed by Franciscan, Dominican, Jesuit, and other Augustinian priests, who founded churches, convents, and schools. In 1574 Manila was baptized under the authorization of Spain and the Vatican as the "Distinguished and Ever Loyal City" and became the centre of Catholicism as well as of the Philippines. At various periods Manila was seriously threatened, and sometimes occupied, by foreign powers. It was invaded by the Chinese in 1574 and raided by the Dutch in the mid-17th century. In 1762, during the Seven Years' War, the city was captured and held by the British, but the Treaty of Paris (1763) resulted in its restoration to Spain. It was opened to foreign trade in 1832, and commerce was further stimulated by the opening of the Suez Canal in 1869.

The Manila area became the centre of anti-Spanish sentiment in the 1890s, and the execution of Filipino patriot José Rizal in the city in December 1896 sparked a year-long insurrection. During the Spanish-American War the Spanish fleet was defeated at Manila Bay on May 1, 1898, and on August 13 the city surrendered to U.S. forces. It subsequently became the headquarters for the U.S. administration of the Philippines.

The U.S. period and independence

The U.S. period was one of general social and economic improvement for the city. U.S. policy encouraged gradual Filipino political autonomy, and to help achieve this goal public schools were established in Manila and throughout the archipelago. The University of the Philippines, founded in 1908, became the apex of the educational system. The city developed into a major trading and tourist centre.

Upon the outbreak of World War II, Manila was declared an open city and was occupied by the Japanese in January 1942. The city suffered little damage during the Japanese invasion but was levelled to the ground during the fight for its recapture by U.S. forces in 1945.

Manila was in shambles when in 1946 it became the capital of the newly independent Republic of the Philippines. The city was rapidly rebuilt, however, with U.S. aid. A significant change in its appearance was brought about by industrialization. In 1948 suburban Quezon City was chosen as the site of a new national capital, but in 1976 Manila again became the capital and the permanent seat of the national government.

BIBLIOGRAPHY. General discussions on the history, government, economy, and living conditions in Manila include TEODORO A. AGONCILLO and MILAGROS C. GUERRERO, *History of the Filipino People,* 5th ed. (1977); ALFONSO J. ALUIT, *The Galleon Guide to Manila,* 3rd ed. (1973); MIGUEL ANSELMO BERNAD, *The Western Community of Manila: A Profile* (1974); ROBERT E. HUKE, *Shadows on the Land: An Economic Geography of the Philippines* (1963); LUNING B. IRA, *Streets of Manila* (1977); *The Philippine Atlas,* vol. 1, *A Historical, Economic and Educational Profile of the Philippines* (1975); DOMINGO C. SALITA, *Geography and Natural Resources of the Philippines* (1974).

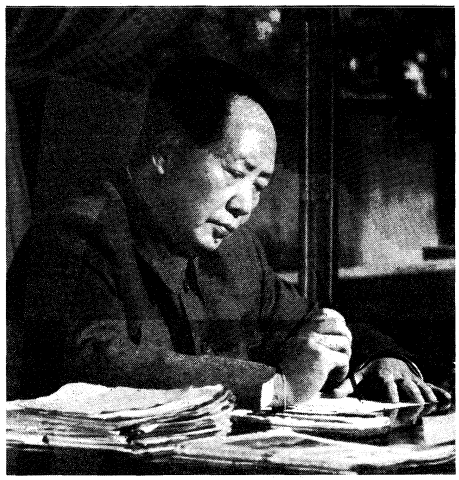
Specific treatment on population, housing, charter provisions, needs, and resources of the capital city may be found in *The Fookien Times Philippine Yearbook 1975,* "Towards a Metropolitan Manila"; NATIONAL CENSUS AND STATISTICS OFFICE, *1975 Integrated Census of the Population and Its Economic Activities* (1975); *Philippine Yearbook 1977; Philippine Law Report,* vol. 11, no. 2 (February 1976); *Metro Manila,* vol. 1 (February 1977–); METRO MANILA RESEARCH TEAM, *Restructuring Government in Metropolitan Manila* (1973). See also ROBERT R. REED, *Colonial Manila: The Context of Hispanic Urbanism and Process of Morphogenesis* (1978); ERNESTO M. PERNIA and CAYETANO W. PADERANGA, *Urbanization and Spatial Development in the Philippines* (1980); VISITACION R. DE LA TORRE, *Landmarks of Manila, 1571–1930* (1981); EMILY HAHN, *The Islands: America's Imperial Adventure in the Philippines* (1981).

(D.C.S.)

Mao Zedong

As China emerged from a half century of revolution as the world's most populous nation and launched itself on a path of economic development and social change, Mao Zedong (Wade–Giles: Mao Tse-tung), its principal revolutionary thinker and for many years its unchallenged leader, occupied a critical place in the story of the country's resurgence. To be sure, he did not play a dominant role throughout the whole struggle. In the early years of the Chinese Communist Party, he was a secondary figure, though by no means a negligible one, and even after the 1940s (except perhaps during the Cultural Revolution) the crucial decisions were not his alone. Nevertheless, looking at the whole period from the foundation of the Chinese Communist Party in 1921 to Mao's death in 1976, one can fairly regard Mao Zedong as the principal architect of the new China.

Eastfoto



Mao Zedong, 1966.

Early years. Born on Dec. 26, 1893, in the village of Shao-shan, Hunan Province, Mao was the son of a former poor peasant who had become affluent as a farmer and grain dealer. He grew up in an environment in which education was valued only as training for keeping records and accounts. From the age of eight he attended his native village's primary school, where he acquired a basic knowledge of the Confucian Classics. At 13 he was forced to leave and begin working full time on his family's farm. Rebelling against paternal authority, Mao left his family to study at a higher primary school in a neighbouring county and then at a secondary school in the provincial capital, Ch'ang-sha. There he came in contact with new ideas from the West, as formulated by such political and cultural reformers as Liang Ch'i-ch'ao and the Nationalist revolutionary Sun Yat-sen. Scarcely had he begun studying revolutionary ideas when a real revolution took place before his very eyes. On Oct. 10, 1911, fighting against the Manchu dynasty broke out in Wu-ch'ang, and within two weeks the revolt had spread to Ch'ang-sha.

Enlisting in a unit of the revolutionary army in Hunan, Mao spent six months as a soldier. While he probably had not yet clearly grasped the idea that, as he later put it, "political power grows out of the barrel of a gun," his first brief military experience at least confirmed his boyhood admiration of military leaders and exploits. In primary school days, his heroes had included not only the great warrior-emperors of the Chinese past but Napoleon and George Washington as well.

Enlistment in revolutionary army

The spring of 1912 saw the birth of the new Chinese Republic and the end of Mao's military service. For a year he drifted from one thing to another, trying, in turn, a police school, a law school, and a business school; he studied history in a secondary school and then spent some months reading many of the classic works of the Western liberal tradition in the provincial library. This period of groping, rather than indicating any lack of decision in Mao's character, was a reflection of China's situation at the time. The abolition of the official civil service examination system in 1905 and the piecemeal introduction of Western learning in so-called modern schools had left young people in a state of uncertainty as to what type of training, Chinese or Western, could best prepare them for a career or for service to their country.

Mao eventually was graduated from the First Provincial Normal School in Ch'ang-sha in 1918. While officially an institution of secondary level rather than of higher education, the normal school offered a high standard of instruction in Chinese history, literature, and philosophy as well as in Western ideas. While at the school, Mao also acquired his first experience in political activity by helping to establish several student organizations. The most important of these was the New People's Study Society, founded in the winter of 1917–18, many of whose members were later to join the Communist Party.

From the normal school in Ch'ang-sha, Mao went to Peking University, China's leading intellectual centre. The half year he spent in Peking working as a librarian's assistant was of disproportionate importance in shaping his future career, for it was then that he came under the influence of the two men who were to be the principal figures in the foundation of the Chinese Communist Party: Li Dazhao (Li Ta-chao) and Chen Duxiu (Ch'en Tu-hsiu). Moreover, he found himself at Peking University precisely during the months leading up to the May Fourth Movement of 1919, which was to a considerable extent the fountainhead of all of the changes that were to take place in China in the ensuing half century.

May Fourth Movement

In a limited sense, May Fourth Movement is the name given to the student demonstrations protesting against the Paris Peace Conference's decision to hand over former German concessions in Shantung Province to Japan instead of returning them to China. But the term also evokes a period of rapid political and cultural change, beginning in 1915, that resulted in the Chinese radicals' abandonment of Western liberalism for Marxism-Leninism as the answer to China's problems and the subsequent founding of the Chinese Communist Party in 1921. The shift from the difficult and esoteric classical written language to a far more accessible vehicle of literary expression patterned on colloquial speech also took place during this period. At the same time, a new and very young generation moved to the centre of the political stage. To be sure, the demonstration on May 4 was launched by Chen Duxiu, but the students soon realized that they themselves were the main actors. In an editorial published in July 1919, Mao wrote:

> The world is ours, the nation is ours, society is ours. If we do not speak, who will speak? If we do not act, who will act?

From then onward, his generation never ceased to regard itself as responsible for the nation's fate, and, indeed, its members remained in power, both in Peking and in Taipei, until the 1970s.

During the summer of 1919 Mao Zedong helped to establish in Ch'ang-sha a variety of organizations that brought the students together with the merchants and the workers—but not yet with the peasants—in demonstrations aimed at forcing the government to oppose Japan. His writings at the time are filled with references to the "army of the red flag" throughout the world and to the victory of the Russian Revolution, but it was not until January 1921 that he was finally committed to Marxism as the philosophical basis of the revolution.

Mao and the Chinese Communist Party. In September

1920 he became principal of the Lin Ch'ang-sha primary school, and in October he organized a branch of the Socialist Youth League there. That winter he married Yang Kaihui (Yang K'ai-hui), the daughter of his former ethics teacher. In July 1921 he attended the First Congress of the Chinese Communist Party, together with representatives from the other Communist groups in China and two delegates from the Moscow-based Comintern (Communist International). In 1923, when the young party entered into an alliance with Sun Yat-sen's Kuomintang (Nationalist Party), Mao was one of the first Communists to join the Kuomintang and to work within it. During the first half of 1924, he lived mostly with his wife and two infant sons in Shanghai, where he was a leading member of the Kuomintang Executive Bureau.

In the winter of 1924–25, Mao returned to his native village of Shao-shan for a rest. There, after witnessing demonstrations by peasants stirred into political consciousness by the shooting of several dozen Chinese by foreign police in Shanghai (May and June 1925), Mao suddenly became aware of the revolutionary potential inherent in the peasantry. Although born in a peasant household, he had, in the course of his student years, adopted the Chinese intellectual's traditional view of the workers and peasants as ignorant and dirty. His conversion to Marxism had forced him to revise his estimate of the urban proletariat, but he continued to share Marx's own contempt for the backward and amorphous peasantry. Now he turned back to the rural world of his youth as the source of China's regeneration. Following the example of other Communists working within the Kuomintang who had already begun to organize the peasants, Mao sought to channel the spontaneous protests of the Hunanese peasants into a network of peasant associations.

Peasants and the revolution

The Communists and the Kuomintang. Pursued by the military governor of Hunan, Mao was soon forced to flee his native province once more, and he returned for another year to an urban environment—this time to Canton, the main power base of the Kuomintang. But, though he lived in Canton, Mao still focused his attention on the countryside. He became the acting head of the propaganda department of the Kuomintang—in which capacity he edited its leading organ, the *Political Weekly,* and attended the Second Kuomintang Congress in January 1926—but he also served at the Peasant Movement Training Institute, set up in Canton under the auspices of the Kuomintang, as principal of the sixth training session. Chiang Kai-shek had become the leader of the Kuomintang after the death of Sun Yat-sen in March 1925; and although Chiang still declared his allegiance to the "world revolution" and wished to avail himself of Soviet aid, he was determined to remain master in his own house. He therefore expelled most Communists from responsible posts in the Kuomintang in May 1926. Mao, however, stayed on at the institute until October of that year. Most of the young peasant activists Mao trained were shortly at work strengthening the position of the Communists.

In July 1926, Chiang Kai-shek set out on what became known as the Northern Expedition, aiming to unify the country under his own leadership and to overthrow the conservative government in Peking as well as other warlords. In November Mao once more returned to Hunan; there, in January and February 1927, he investigated the peasant movement and concluded that in a very short time several hundred million peasants in China would "rise like a tornado or tempest—a force so extraordinarily swift and violent that no power, however great, will be able to suppress it." Strictly speaking, this prediction proved to be false. Revolution in the shape of spontaneous action by hundreds of millions of peasants did not sweep across China "in a very short time," or indeed at all. Chiang Kai-shek, who was bent on an alliance with the propertied classes in the cities and in the countryside, turned against the worker and peasant revolution, and in April he massacred the very Shanghai workers who had delivered the city to him. Stalin's strategy for carrying out revolution in alliance with the Kuomintang collapsed, and the Chinese Communist Party was virtually annihilated in the cities and decimated in the countryside. But in a broader and less literal sense, Mao's prophecy was justified. In October 1927 Mao led a few hundred peasants who had survived the autumn harvest uprising in Hunan to a base in the Ching-kang Shan (Ching-kang Mountains), on the Kiangsi–Hunan border, and embarked on a new type of revolutionary warfare in the countryside in which the Red Army, rather than the unarmed masses, would play the central role. But it was only because a large proportion of China's hundreds of millions of peasants sympathized with and supported this effort that Mao Zedong was able in the course of the civil war to encircle the cities from the countryside and thus defeat Chiang Kai-shek and gain control of the country.

Struggle with Chiang Kai-shek

The road to power. Mao Zedong's 22 years in the wilderness can be divided into four phases. The first of these is the initial three years when Mao and Zhu De (Chu Teh), the commander in chief of the army, successfully developed the tactics of guerrilla warfare from base areas in the countryside. These activities, however, were regarded even by their protagonists, and still more by the Central Committee in Shanghai (and by the Comintern in Moscow), as a holding operation until the next upsurge of revolution in the urban centres. In the summer of 1930 the Red Army was ordered by the Central Committee to occupy several major cities in south central China in the hope of sparking a revolution by the workers. When it became evident that persistence in this attempt could only lead to further costly losses, Mao disobeyed orders and abandoned the battle to return to the base in southern Kiangsi. During this year Mao's wife was executed by the Kuomintang and he married He Zizhen (Ho Tzu-chen), with whom he had been living since 1928.

The second phase (the Kiangsi period) centres on the founding of the Chinese Soviet Republic, November 1931, in a portion of Kiangsi Province, with Mao as chairman. Since there was little support for the revolution in the cities, the promise of ultimate victory now seemed to reside in the gradual strengthening and expansion of the base areas. The Soviet regime soon came to control a population of several million; the Red Army, grown to a strength of some 200,000, easily defeated large forces of inferior troops sent against it by Chiang Kai-shek in the first four of the so-called encirclement and annihilation campaigns. But it was unable to stand up against Chiang's own elite units, and in October 1934 the major part of the Red Army, Mao, and his pregnant wife abandoned the base in Kiangsi and set out for the northwest of China, on what is known as the Long March.

The Kiangsi period

There is wide disagreement among specialists as to the extent of Mao's real power, especially in the years 1932–34, and as to which military strategies were his or other party leaders'. The majority view is that, in the last years of the Chinese Soviet Republic, Mao functioned to a considerable extent as a figurehead with little control over policy, especially in military matters. In any case, he achieved de facto leadership over the party (though not the formal title of chairman) only at the Tsun-i Conference of January 1935 during the Long March.

When some 8,000 troops who had survived the perils of the Long March arrived in Shensi Province in northwestern China in the autumn of 1935, events were already moving toward the third phase in Mao's rural odyssey, which was to be characterized by a renewed united front with the Kuomintang against Japan and by the rise of Mao to unchallenged supremacy in the party. This phase is often called the Yen-an period (for the town in Shensi where the Communists were based), although Mao did not move to Yen-an until December 1936. In August 1935 the Comintern at its Seventh Congress in Moscow proclaimed the principle of an anti-Fascist united front, and in May 1936 the Chinese Communists for the first time accepted the prospect that such a united front might include Chiang Kai-shek himself, and not merely dissident elements in the Nationalist camp. The so-called Sian Incident of December 1936, in which Chiang was kidnapped by military leaders from northeastern China who wanted to fight Japan and recover their homelands rather than participate in civil war against the Communists, accelerated the evolution toward unity. By the time the Japanese

Yen-an period

began their attempt to subjugate all of China in July 1937, the terms of a new united front between the Communists and the Kuomintang had been virtually settled, and the formal agreement was announced in September 1937.

In the course of the anti-Japanese war, the Communists broke up a substantial portion of their army into small units and sent them behind the enemy lines to serve as nuclei for guerrilla forces that effectively controlled vast areas of the countryside, stretching between the cities and communication lines occupied by the invader. As a result, they not only expanded their military forces to somewhere between 500,000 and 1,000,000 at the time of the Japanese surrender but also established effective grass-roots political control over a population that may have totaled as many as 90,000,000. It has been argued that the support of the rural population was won purely by appeals to their nationalist feeling in opposition to the Japanese. This certainly was fundamental, but Communist agrarian policies likewise played a part in securing broad support among the peasantry.

Writings, 1936–40

During the years 1936–40, Mao had, for the first time since the 1920s, the leisure to devote himself to reflection and writing. It was then that he first read in translation a certain number of Soviet writings on philosophy and produced his own account of dialectical materialism, of which the best known portions are those entitled "On Practice" and "On Contradiction." More important, Mao produced the major works that synthesized his own experience of revolutionary struggle and his vision of how the revolution should be carried forward in the context of the united front. On military matters there was first *Strategic Problems of China's Revolutionary War,* written in December 1936 to sum up the lessons of the Kiangsi period (and also to justify the correctness of his own military line at the time), and then *On Protracted War* and other writings of 1938 on the tactics of the anti-Japanese war. As to his overall view of the events of these years, Mao adopted an extremely conciliatory attitude toward the Kuomintang in his report entitled *On the New Stage* (October 1938), in which he attributed to it the leading role both in the war against Japan and in the ensuing phase of national reconstruction. By the winter of 1939–40, however, the situation had changed sufficiently so that he could adopt a much firmer line, claiming leadership for the Communists. Internationally, Mao argued, the Chinese revolution was a part of the world proletarian revolution directed against imperialism (whether it be British, German, or Japanese); internally, the country should be ruled by a "joint dictatorship of several parties" belonging to the anti-Japanese united front. For the time being, Mao felt, the aims of the Communist Party coincided with the aims of the Kuomintang, and therefore Communists should not try to rush ahead to socialism and thus disrupt the united front. But neither should they have any doubts about the ultimate need to take power into their own hands in order to move forward to socialism. During this period, in 1939, Mao divorced He Zizhen and married a well-known film actress, Lan P'ing, later called Jiang Qing (Chiang Ch'ing).

The issues of Kuomintang–Communist rivalry for the leadership of the united front are related to the continuing struggle for supremacy within the Chinese Communist Party, for Mao's two chief rivals—Wang Ming, who had just returned from a long stay in Moscow, and Chang Kuo-t'ao, who had at first refused to accept Mao's political and military leadership—were both accused of excessive slavishness toward the Kuomintang. But perhaps even more central in Mao's ultimate emergence as the acknowledged leader of the party was the question of what he had called in October 1938 the "Sinification" of Marxism—its adaptation not only to Chinese conditions but to the mentality and cultural traditions of the Chinese people.

Mao could not claim the firsthand knowledge possessed by many other leading members of the Chinese Communist Party of how Communism worked within the Soviet Union nor the ability to read Marx or Lenin in the original, which some of them enjoyed. He could and did claim, however, to know and understand China. The differences between him and the Soviet-oriented faction in the party came to a head at the time of the so-called Rectification Campaign of 1942–43. This program aimed at giving a basic grounding in Marxist theory and Leninist principles of party organization to the many thousands of new members who had been drawn into the party in the course of the expansion since 1937. But a second and equally important aspect of the movement was the elimination of what Mao called "foreign dogmatism"—in other words, blind imitation of Soviet experience and obedience to Soviet directives.

Mao becomes chairman

In March 1943 Mao achieved for the first time formal supremacy over the party, becoming chairman of the Secretariat and of the Politburo. Shortly thereafter, the Rectification Campaign took, for a time, the form of a harsh purge of elements not sufficiently loyal to Mao. The campaign was run by Kang Sheng (K'ang Sheng), who was later to be one of Mao's key supporters in the Cultural Revolution. Exaggerating considerably this dimension of events, Soviet spokesmen have bitterly denounced the Rectification Campaign as an attempt to purge the Chinese Communist Party of all those elements genuinely imbued with "proletarian internationalism" (*i.e.,* devotion to Moscow). It is therefore not surprising that, as Mao's campaign in the countryside moved into its fourth and last phase—that of civil war with the Kuomintang—Stalin's lack of enthusiasm for a Chinese Communist victory should have become increasingly evident. Looking back at this period in 1962, when the Sino-Soviet conflict had come to a head, Mao declared:

> In 1945, Stalin wanted to prevent China from making revolution, saying that we should not have a civil war and should cooperate with Chiang Kai-shek, otherwise the Chinese nation would perish. But we did not do what he said. The revolution was victorious. After the victory of the revolution he [Stalin] next suspected China of being a Yugoslavia, and that I would become a second Tito.

This account of Stalin's attitude is substantiated by a whole series of public gestures at the time, culminating in the fact that when the People's Liberation Army took Nanking in April 1949, the Soviet ambassador was the only foreign diplomat to accompany the retreating Nationalist government to Canton. Stalin's motives were obviously those described by Mao in the above passage; he did not believe in the capacity of the Chinese Communists to achieve a clear-cut victory, and he thought they would be a nuisance if they did.

Formation of the People's Republic of China. Nevertheless, when the Communists did take power in China, both Mao and Stalin had to make the best of the situation. In December 1949 Mao, now chairman of the People's Republic of China—which he had proclaimed on October 1—traveled to Moscow, where, after two months of arduous negotiations, he succeeded in persuading Stalin to sign a treaty of mutual assistance accompanied by limited economic aid. Before the Chinese had time to profit from the resources made available for economic development, however, they found themselves dragged into the Korean War in support of the Moscow-oriented regime in P'yŏngyang. Only after this baptism of fire did Stalin, according to Mao, begin to have confidence in him and believe he was not first and foremost a Chinese nationalist.

Despite these tensions with Moscow, the policies of the People's Republic of China in its early years were based in very many respects, as Mao later said, on "copying from the Soviets." While Mao and his comrades had experience in guerrilla warfare, in mobilization of the peasants in the countryside, and in political administration at the grass roots, they had no firsthand knowledge of running a state or of large-scale economic development. In such circumstances, the Soviet Union provided the only available model. A five-year plan was therefore drawn up under Soviet guidance; it was put into effect in 1953 and included Soviet technical assistance and a number of complete industrial plants. Yet, within two years, Mao had taken steps that were to lead to the breakdown of the political and ideological alliance with Moscow.

The emergence of Mao's road to socialism. In the spring of 1949, Mao proclaimed that while in the past the Chinese revolution had followed the unorthodox path of "encircling the cities from the countryside," it would

in the future take the orthodox road of the cities leading and guiding the countryside. In harmony with this view, he had agreed in 1950 with Liu Shaoqi (Liu Shao-ch'i) that collectivization would be possible only when China's heavy industry had provided the necessary equipment for mechanization. In a report of July 1955, he reversed this position, arguing that in China the social transformation could run ahead of the technical transformation. Deeply impressed by the achievements of certain cooperatives that claimed to have radically improved their material conditions without any outside assistance, he came to believe in the limitless capacity of the Chinese people, especially of the rural masses, to transform at will both nature and their own social relations when mobilized for revolutionary goals. Those in the leadership who did not share this vision he denounced as "old women with bound feet." He made these criticisms before an ad hoc gathering of provincial and local party secretaries, thus creating a ground swell of enthusiasm for rapid collectivization such that all those in the leadership who had expressed doubts about Mao's ideas were soon presented with a fait accompli. The tendency thus manifested to pursue his own ends outside the collective decision-making processes of the party was to continue and to be accentuated.

Even before Nikita S. Khrushchev's secret speech of February 1956 denouncing Stalin's crimes, Mao Zedong and his colleagues had been discussing measures for improving the morale of the intellectuals in order to secure their willing participation in building a new China. At the end of April, Mao proclaimed the policy of "letting a hundred flowers bloom"—that is, the freedom to express many diverse ideas—designed to prevent the development in China of a repressive political climate analogous to that in the Soviet Union under Stalin. In the face of the disorders called forth by de-Stalinization in Poland and Hungary, Mao did not retreat but rather pressed boldly forward with this policy, against the advice of many of his senior colleagues, in the belief that the contradictions that still existed in Chinese society were mainly nonantagonistic. When the resulting "great blooming and contending" got out of hand and called into question the axiom of party rule, Mao savagely turned against the educated elite, which he felt had betrayed his confidence. Henceforth, he would rely primarily on the creativity of the rank and file as the agent of modernization. As for the specialists, if they were not yet sufficiently "red," he would remold them by sending them to work in the countryside.

Hundred Flowers speech

It was against this background that Mao, during the winter of 1957–58, worked out the policies that were to characterize the Great Leap Forward, formally launched in May 1958. While his economic strategy was by no means so one-sided and simplistic as commonly believed in the 1960s and '70s, and though he still proclaimed industrialization and a "technical revolution" as his goals, Mao displayed continuing anxiety regarding the corrupting influence of the fruits of technical progress and an acute nostalgia for the purity and egalitarianism that had marked the moral and political world of the Ching-kang Shan and Yen-an eras.

Great Leap Forward

Thus it was logical that he should endorse and promote the establishment of "people's communes" as part of the Great Leap strategy. As a result, the peasants, who had been organized into cooperatives in 1955–56 and then into fully socialist collectives in 1956–57, found their world turned upside down once again in 1958. Neither the resources nor the administrative experience necessary to operate such enormous new social units of several thousand households were in fact available, and not surprisingly the consequences of these changes were chaos and economic disaster.

By the winter of 1958–59, Mao himself had come to recognize that some adjustments were necessary, including decentralization of ownership to the constituent elements of the communes and a scaling down of the unrealistically high production targets in both industry and agriculture. He insisted, however, that in broad outline his new Chinese road to socialism, including the concept of the communes and the belief that China, though "poor and blank," could leap ahead of other countries, was basically sound. At the Lushan meeting of the Central Committee in July–August 1959, Peng Dehuai (P'eng Te-huai), the minister of defense, denounced the excesses of the Great Leap and the economic losses they had caused. He was immediately removed from all his party and state posts and remained in detention until his death during the Cultural Revolution. From that time, Mao regarded any criticism of his policies as nothing less than a crime of lèse-majesté, meriting exemplary punishment.

Retreat and counterattack. Though few spoke up at Lushan in support of Peng, a considerable number of the top leaders sympathized with him in private. Almost immediately, in 1960, Mao began building an alternative power base in the People's Liberation Army, which the new defense minister, Lin Biao (Lin Piao), had set out to turn into a "great school of Mao Zedong Thought." At about the same time, Mao began to denounce the emergence, not only in the Soviet Union but also in China itself, of "new bourgeois elements" among the privileged strata of the state and party bureaucracy and the technical and artistic elite. Under these conditions, he concluded, a "protracted, complex, and sometimes even violent class struggle" would continue during the whole socialist stage.

Split with Soviet Union

The open split with the Soviet Union, which had become public and irreparable by 1963—though it can be traced to Mao's resentment at Khrushchev's failure to consult him before launching de-Stalinization—resulted, above all, from the Soviet reaction to the Great Leap policies. Regarding Mao's claims for the communes as ideologically presumptuous, Khrushchev heaped ridicule upon them; he underlined his displeasure by withdrawing Soviet technical assistance in 1960, leaving many large plants unfinished. Khrushchev also tried to put pressure on China in its dealings with Taiwan and India and in other foreign-policy issues. Mao forgot neither the affront to his and China's dignity nor the economic damage.

As for class struggle in China itself, Mao's fear that revisionism might appear there was also heightened by the policies pursued in the early 1960s to deal with the economic consequences of the Great Leap Forward. The disorganization and waste created by the Great Leap, compounded by natural disasters and by the termination of Soviet economic aid, led to widespread famine in which, according to much later official Chinese accounts, millions of people died. The response to this situation by Liu Shaoqi (who had succeeded Mao as chairman of the People's Republic in 1959), Deng Xiaoping (Teng Hsiao-p'ing), and the economic planners was to make use of material incentives and to strengthen the role of individual households in agricultural production. At first Mao agreed reluctantly that such steps were necessary, but during the first half of 1962 he came increasingly to perceive the methods used to promote recovery as implying the repudiation of the whole thrust of the Great Leap strategy. It was as a direct response to this challenge that at the 10th Plenary Session of the Central Committee in September 1962 he issued the call, "Never forget the class struggle!"

During the next three years Mao waged such a struggle, primarily through the Socialist Education Movement in the countryside, and it was over the guidelines for this campaign that the major political battles were fought within the Chinese leadership. At the end of 1964, when Liu Shaoqi refused to accept Mao's demand to direct the main thrust of class struggle against "capitalist roaders" in the party, Mao decided that "Liu had to go."

The Cultural Revolution. The movement that became known as the Great Proletarian Cultural Revolution represented an attempt by Mao to go beyond the party rectification campaigns, of which there had been many since 1942, and to devise a new and more radical method for dealing with what he saw as the bureaucratic degeneration of the party. But it also represented, beyond any doubt or question, a deliberate effort to eliminate those in the leadership who, over the years, had dared to cross him. The victims, from throughout the party hierarchy, suffered more than mere political disgrace. All were publicly humiliated and detained for varying periods, sometimes under very harsh conditions; many were beaten and tortured, and not a few were killed or driven to suicide.

Elimination of Liu

Among the casualties was Liu, who died because he was denied proper medical attention.

The justification for these sacrifices was defined in a key slogan of the time: "Fight selfishness, criticize revisionism." When the Red Guards, who constituted the first shock troops of Mao's enterprise, burst upon the scene in the summer of 1966, with their battle cry, "To rebel is justified!" it seemed for a time that not only the power of the party cadres but also authority in all its forms was being questioned. It soon became evident that Mao, who in 1956 had justified decentralization as a means to building a "strong socialist state," still believed in the need for state power. When the Shanghai leftists Zhang Chunqiao (Chang Ch'un-ch'iao) and Yao Wenyuan (Yao Wen-yüan)—who were later to make up half the Gang of Four—came to see him in February 1967, immediately after setting up the Shanghai Commune, Mao asserted that the demand for the abolition of "heads" (leaders), which had been heard in their city, was "extreme anarchism" and "most reactionary"; in fact, he stated, there would "always be heads." Communes, he added, were "too weak when it came to suppressing counterrevolution" and in any case required party leadership. He therefore ordered them to dissolve theirs and to replace it with a "revolutionary committee."

These committees, based on an alliance of former party cadres, young activists, and representatives of the People's Liberation Army, were to remain in place until two years after Mao's death. At first, they were largely controlled by the army. The Ninth Congress of 1969 initiated the process of rebuilding the party; and the death of Lin Biao diminished, though it by no means eliminated, the army's role. Thereafter, it seemed briefly, in 1971–72, that a compromise, of which Zhou Enlai (Chou En-lai) was the architect, might produce some kind of synthesis between the values of the Cultural Revolution and the pre-1966 political and economic order.

Even before Zhou's death in January 1976, however, this compromise had been overturned. All recognition by Mao of the importance of professional skills was swallowed up in an orgy of political rhetoric, and all things foreign were regarded as counterrevolutionary. Mao's last decade, which had opened with manifestos in favour of the Paris Commune model of mass democracy, closed with paeans of praise to that most implacable of centralizing despots, Shih Huang-ti, the first Ch'in emperor. Mao Zedong died in Peking on Sept. 9, 1976.

Mao's death

Assessment. While the Cultural Revolution was an entirely logical culmination of Mao's last two decades, it was by no means the only possible outcome of his approach to revolution, nor need a judgment of his work as a whole be based primarily on this last phase.

Few would deny Mao Zedong the major share of credit for devising the pattern of struggle based on guerrilla warfare in the countryside that ultimately led to victory in the civil war and thereby to the overthrow of the Kuomintang, the distribution of land to the peasants, and the restoration of China's independence and sovereignty. These achievements must be given a weight commensurate with the degree of injustice prevailing in Chinese society before the revolution and with the humiliation felt by the Chinese people as a result of the dismemberment of their country by the foreign powers. "We have stood up," Mao said in September 1949. These words will not be forgotten.

Mao's record after 1949 is more ambiguous. The official Chinese view, defined in June 1981, is that his leadership was basically correct until the summer of 1957, but from then on it was mixed at best and frequently quite wrong. It cannot be disputed that Mao's two major innovations of his later years, the Great Leap and the Cultural Revolution, were ill-conceived and led to disastrous consequences. His goals of combating bureaucracy, encouraging popular participation, and stressing China's self-reliance were generally laudable, but the methods he used to pursue them, though bold and imaginative, were largely self-defeating.

Looking at Mao's whole career, it is not easy to put a figure on the positive and negative aspects. How does one weigh the good fortune of peasants acquiring land against millions of executions and deaths from civil war? How does one balance the real economic achievements after 1949 against the starvation that came in the wake of the Great Leap Forward or the bloody shambles of the Cultural Revolution? It is, perhaps, possible to accept the official verdict that, despite the "errors of his later years," Mao's merits outweighed his faults, while underscoring the fact that the account is very finely balanced.

BIBLIOGRAPHY. A detailed biography of Mao Zedong is ROSS TERRILL, *Mao* (1980, reissued 1981). Two earlier works that remain useful for the pre-1949 period are JEROME CH'EN, *Mao and the Chinese Revolution* (1965, reissued 1972); and STUART R. SCHRAM, *Mao Tse-tung,* rev. ed. (1967, reprinted 1974). See also Schram's *Mao Zedong, A Preliminary Reassessment* (1983). The most vivid account of Mao's youth is his autobiography as recounted in 1936 in EDGAR SNOW, *Red Star over China,* rev. ed. (1968, reissued 1974). JUI LI, *The Early Revolutionary Activities of Comrade Mao Tse-tung* (1977; originally published in Chinese, 1957), is also an important source.

Regarding Mao Zedong's thought, a substantial collection of source materials for the period before 1949 is available in *Selected Works of Mao Tse-tung,* 5 vol. (1961–77). A variorum in Chinese of the collected writings of Mao to 1949 is *Mao Tse-tung chi,* ed. by MINORU TAKEUCHI, 10 vol. (1970–72), completed by a set of supplements, *Mao Tse-tung chi pu chüan* (1983–85). Mao's talks and letters from 1956 to 1971 are found in STUART R. SCHRAM (ed.), *Mao Tse-tung Unrehearsed* (1974, U.S. title, *Chairman Mao Talks to the People,* 1975). See also JEROME CH'EN (ed.), *Mao Papers* (1970); and MAO TSETUNG, *A Critique of Soviet Economics* (1977), trans. from Chinese by MOSS ROBERTS. On Mao's thought in his early years, see BRANTLY WOMACK, *The Foundation of Mao Zedong's Political Thought, 1917–1935* (1982); FREDERIC WAKEMAN, JR., *History and Will: Philosophical Perspectives of Mao Tse-tung's Thought* (1973, reprinted 1975), which links Mao's ideas of the May Fourth period with those of the Cultural Revolution; and RAYMOND F. WYLIE, *The Emergence of Maoism* (1980). JOHN BRYAN STARR, *Continuing the Revolution: The Political Thought of Mao* (1979), is a comprehensive overview that accepts at face value the Chinese view of the Chairman during his lifetime. Among the older works, ARTHUR A. COHEN, *The Communism of Mao Tse-tung* (1964, reprinted 1971), stresses the Stalinist roots of Mao's thought; and JAMES HSIUNG, *Ideology and Practice: The Evolution of Chinese Communism* (1970), emphasizes the links between Mao's thought and Chinese tradition. See also STUART R. SCHRAM (ed.), *The Political Thought of Mao Tse-tung,* rev. ed. (1969). Finally, a series of useful if somewhat premature appreciations are in DICK WILSON (ed.), *Mao Tse-tung in the Scales of History* (1977).

(S.R.S.)

Mapping and Surveying

Maps and charts are representations of features on the Earth's surface drawn to scale. Globes are maps presented on the surface of a sphere. The definition can be extended to include representations of the Moon, Mars, and other bodies. Star charts and related celestial plats, long used for navigation and astronomical studies, might also be included; they are treated in the article STARS AND STAR CLUSTERS.

In order to imply the elements of accurate relationships, and some formal method of projecting the spherical subject to a map plane, further qualifications might be applied to the definition. The tedious and somewhat abstract statements resulting from attempts to formulate precise definitions of maps and charts are more likely to confuse than to clarify. The words map, chart, and plat are used somewhat interchangeably. The connotations of use,

however, are distinctive: charts for navigation purposes (nautical and aeronautical), plats (in a property-boundary sense) for land-line references and ownership, and maps for general reference.

Cartography is the art and science of making maps and charts. As such, it is allied with geography in its concern with the broader aspects of the Earth and its life. In early times cartographic efforts were more artistic than scientific and factual. As man explored and recorded his environment, the quality of his maps and charts improved. These lines of Jonathan Swift were inspired by early maps:

So geographers, in Afric maps,
With savage pictures fill their gaps,
And o'er unhabitable downs
Place elephants for want of towns.

Topographic maps are graphic representations of natural and man-made features of parts of the Earth's surface plotted to scale. They show the shape of land and record elevations above sea level, lakes, streams and other hydrographic features, and roads and other works of man. In short, they provide a complete inventory of the terrain and important information for all activities involving the use and development of the land. They provide the bases for specialized maps and data for compilation of generalized maps of smaller scale.

Nautical charts are maps of coastal and marine areas, providing information for navigation. They include depth curves or soundings or both; aids to navigation such as buoys, channel markers, and lights; islands, rocks, wrecks, reefs and other hazards; and significant features of the coastal areas, including promontories, church steeples, water towers, and other features helpful in determining positions from offshore.

The terms hydrography and hydrographer date from the mid-16th century; their focus has become restricted to studies of ocean depths and of the directions and intensities of oceanic currents; though at various times they embraced much of the sciences now called hydrology and oceanography. The British East India Company employed hydrographers in the 18th century, and the first hydrographer of the Royal Navy, Alexander Dalrymple (1737–1808), was appointed in 1795. A naval observatory and hydrographic office was established administratively in the United States Navy in 1854. In 1866 a hydrographic office was established by statute, and in 1962 it was renamed the U.S. Naval Oceanographic Office.

Interest in the charting of oceanic areas away from seacoasts developed in the second half of the 19th century, concurrently with the perfection of submarine cables. As knowledge of the configuration of the ocean basins increased, the attention of scientists was drawn to this field of study. A feature of marine science since the 1950s has been increasingly detailed bathymetric (water-depth measurement) surveys of selected portions of the seafloor. Together with collection of associated geophysical data and sampling of sediments, these studies assist in interpreting the geologic history of the ocean-covered portion of the Earth's crust.

Aeronautical charts provide essential data for the pilot and air navigator. They are, in effect, small-scale topographic maps on which current information on aids to navigation have been superimposed. To facilitate rapid recognition and orientation, principal features of the land that would be visible from an aircraft in flight are shown to the exclusion of less important details.

This article is organized into the following sections:

History of cartography 516
  Maps and geography in the ancient world 516
    Greek maps and geography
    The Roman period
  The Middle Ages 518
  The age of discovery and exploration 519
    Revival of Ptolemy
    Maps of the discoveries
  18th century to the present 520
    The rise of national surveys
    International Map of the World (IMW)
    World War II and after
Mapmaking 521
  Elements 521
    Map scales and classifications
    Map projections
    Development of reference spheroids
    Geographic and plane coordinate systems
    Basic data for compilation
    Symbolization
    Nomenclature
  Types and uses of maps and charts 525
    World status of mapping and basic data
    Types of maps and charts available
    Government and other mapping agencies
    International organizations
  Modern mapmaking techniques 527
    Compilation from existing materials
    Map production from original surveys
    Final steps in map preparation
    Automation in mapping
Surveying 529
  History 529
  Modern surveying 530
    Basic control surveys
    Global positioning
    Establishing the framework
    Detail surveying
    Aerial surveying
    Hydrography
    Height determination
Bibliography 537

History of cartography

Maps on clay tablets

Centuries before the Christian Era, Babylonians drew maps on clay tablets, of which the oldest specimens found so far have been dated about 2300 BC. This is the earliest positive evidence of graphic representations of parts of the Earth; it may be assumed that mapmaking goes back much further and that it began among nonliterate peoples. It is logical to assume that men very early made efforts to communicate with each other regarding their environment by scratching routes, locations, and hazards on the ground and later on bark and skins.

The earliest maps must have been based on personal experience and familiarity with local features. They doubtless showed routes to neighbouring tribes, where water and other necessities might be found, and the locations of enemies and other dangers. Nomadic life stimulated such efforts by recording ways to cross deserts and mountains, the relative locations of summer and winter pastures, and dependable springs, wells, and other information.

Markings on cave walls associated with paintings by primitive man have been identified by some archaeologists as attempts to show the game trails of the animals depicted, though there is no general agreement on this. Similarly, networks of lines scratched on certain bone tablets could possibly represent hunting trails, but there is definitely no conclusive evidence that the tablets are indeed maps.

Many nonliterate peoples, however, are skilled in depicting essential features of their localities and travels. During Capt. Charles Wilkes's exploration of the South Seas in the 1840s, a friendly islander drew a good sketch of the whole Tuamotu Archipelago on the deck of the captain's bridge. In North America the Pawnee Indians were reputed to have used star charts painted on elk skin to guide them on night marches across the plains. Montezuma is said to have given Cortés a map of the whole Mexican Gulf area painted on cloth, while Pedro de Gamboa reported that the Incas used sketch maps and cut some in stone to show relief features. Many specimens of early Eskimo sketch maps on skin, wood, and bone have been found.

Maps in the New World

MAPS AND GEOGRAPHY IN THE ANCIENT WORLD

The earliest specimens thus far discovered that are indisputably portrayals of land features are the Babylonian

tablets previously mentioned; certain land drawings found in Egypt and paintings discovered in early tombs are nearly as old. It is quite probable that these two civilizations developed their mapping skills more or less concurrently and in similar directions. Both were vitally concerned with the fertile areas of their river valleys and therefore doubtless made surveys and plats soon after settled communities were established. Later they made plats for the construction of canals, roads, and temples—the equivalent of today's engineering plans.

A tablet unearthed in Iraq shows the Earth as a disk surrounded by water with Babylon as its centre. Aside from this specimen, dating from about 1000 BC, there appear to have been rather few attempts by Babylonians and Egyptians to show the form and extent of the Earth as a whole. Their mapmaking was preoccupied with more practical needs, such as the establishment of boundaries. Not until the time of the Greek philosopher-geographers did speculations and conclusions as to the nature of the Earth begin to take form.

Greek maps and geography. The Greeks were outstanding among peoples of the ancient world for their pursuit and development of geographic knowledge. The shortage of arable land in their own region led to maritime exploration and the development of commerce and colonies. By 600 BC Miletus, on the Aegean, had become a centre of geographic knowledge, as well as of cosmographic speculation.

Hecataeus, a scholar of Miletus, probably produced the first book on geography in about 500 BC. A generation later Herodotus, from more extensive studies and wider travels, expanded upon it. A historian with geographic leanings, Herodotus among the other things recorded an early circumnavigation of the African continent by Phoenicians. He also improved on the delineation of the shape and extent of the then-known regions of the world, and he declared the Caspian to be an inland sea, opposing the prevailing view that it was part of the "northern oceans" (Figure 1).

Herodotus

Although Hecataeus regarded the Earth as a flat disk surrounded by ocean, Herodotus and his followers questioned the concept and proposed a number of other possible forms. Indeed, the philosophers and scholars of the time appear to have been preoccupied for a number of years with discussions on the nature and extent of the world. Some modern scholars attribute the first hypothesis of a spherical Earth to Pythagoras (6th century BC) or Parmenides (5th century). The idea gradually developed into a consensus over many years. In any case by the mid-4th century the theory of a spherical Earth was well accepted among Greek scholars, and about 350 BC Aristotle formulated six arguments to prove that the Earth was, in truth, a sphere. From that time forward, the idea of a spherical Earth was generally accepted among geographers and other men of science.

About 300 BC Dicaearchus, a disciple of Aristotle, placed an orientation line on the world map, running east and west through Gibraltar and Rhodes. Eratosthenes, Marinus of Tyre, and Ptolemy successively developed the reference-line principle until a reasonably comprehensive system of parallels and meridians, as well as methods of projecting them, had been achieved.

The greatest figure of the ancient world in the advancement of geography and cartography was Claudius Ptolemaeus (Ptolemy; AD 90–168). An astronomer and mathematician, he spent many years studying at the library in Alexandria, the greatest repository of scientific knowledge at that time. His monumental work, the *Guide to Geography* (*Geōgraphikē hyphēgēsis*), was produced in eight volumes. The first volume discussed basic principles and dealt with map projection and globe construction. The next six volumes carried a list of the names of some 8,000 places and their approximate latitudes and longitudes. Except for a few that were made by observations, the greater number of these locations were determined from older maps, with approximations of distances and directions taken from travelers. They were accurate enough to show relative locations on the very small-scale, rudimentary maps that existed.

The eighth volume was a most important contribution, containing instructions for preparing maps of the world and discussions on mathematical geography and other fundamental principles of cartography. Ptolemy's map of the world as it was then known marked the culmination of Greek cartography as well as a compendium of accumulated knowledge of the Earth's features at that time (Figure 2).

The Roman period. Although Ptolemy lived and worked at the time of Rome's greatest influence, he was a Greek and essentially a product of that civilization, as was the great library at Alexandria. His works greatly influenced the development of geography, which he defined in mapmaking terms: "representation in picture of the whole known world, together with the phenomena contained therein." This had considerable influence in directing scholars toward the specifics of map construction and away from the more abstract and philosophical aspects of geography.

One fundamental error that had far-reaching effects was attributed to Ptolemy—an underestimation of the size of the Earth. He showed Europe and Asia as extending over half the globe, instead of the 130 degrees of their true extent. Similarly, the span of the Mediterranean ultimately was proved to be 20 degrees less than Ptolemy's estimate. So lasting was Ptolemy's influence that 13 centuries later Christopher Columbus underestimated the distances to Cathay and India partly from a recapitulation of this basic error.

Ptolemy's fundamental error

A fundamental difference between the Greek and Roman philosophies was indicated by their maps. The Romans were less interested in mathematical geography and tended toward more practical needs for military campaigns and provincial administration. They reverted to the older concepts of a disk-shaped world for maps of great areas

By courtesy of the Library of Congress, Washington, D.C.

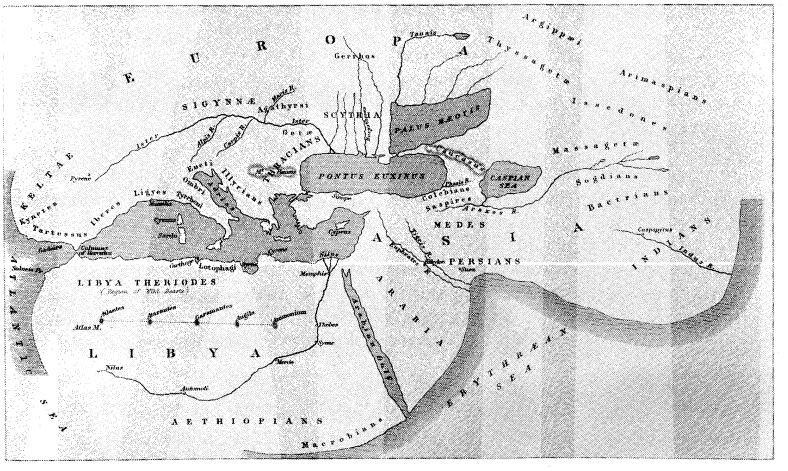


Figure 1: Herodotus' map of the world.

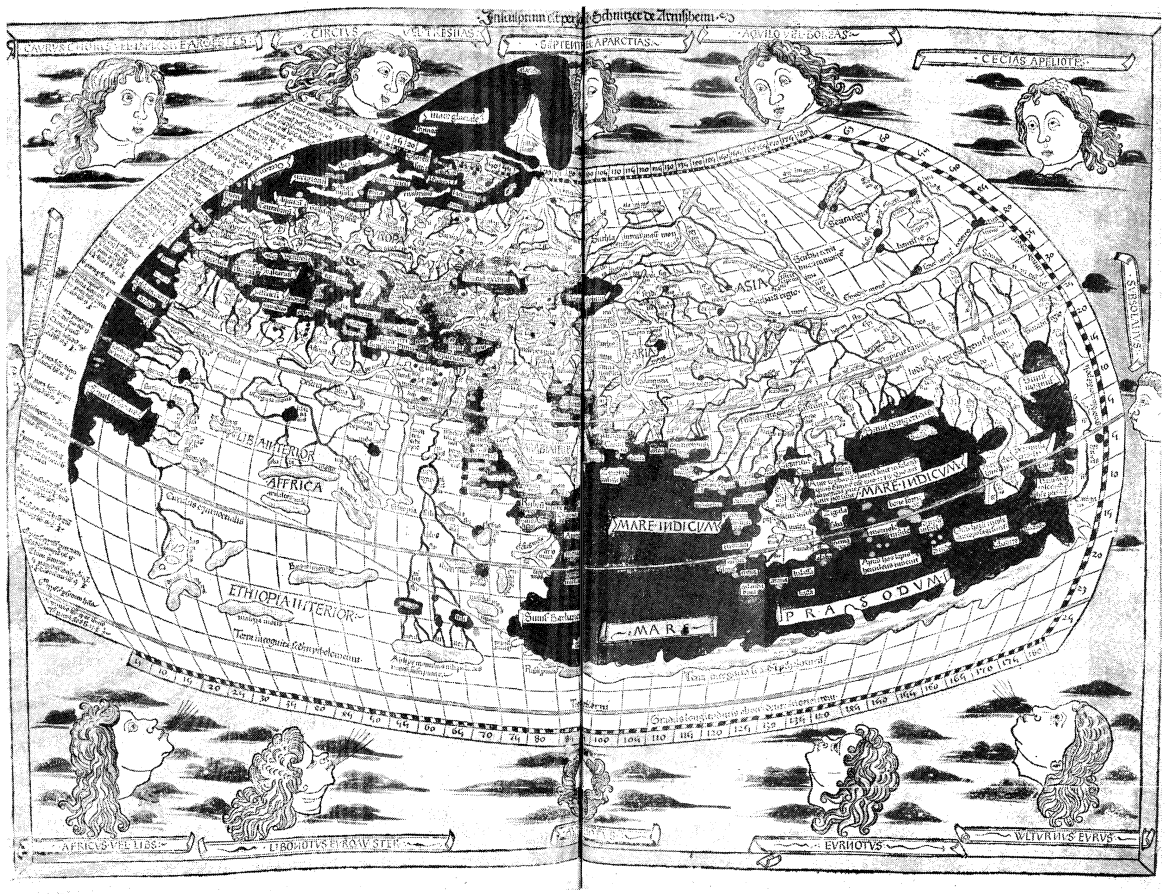


Figure 2: Ptolemy's map of the world, as printed at Ulm, 1482.
By courtesy of the Library of Congress, Washington, D.C.

because they met their needs and were easier to read and understand.

The Roman general Marcus Vipsanius Agrippa, prior to Ptolemy's time, constructed a map of the world based on surveys of the then-extensive system of Roman military roads. References to many other Roman maps have been found, but very few actual specimens survived the Dark Ages. It is quite probable that the Peutinger Table, a parchment scroll showing the roads of the Roman world, was originally based on Agrippa's map and subjected to several revisions through medieval times.

The tragic turn of world events during the first few centuries of the Christian Era wrought havoc to the accumulated knowledge and progress of mankind. As with other fields of science and technology, progress in geography and cartography was abruptly curtailed. After Ptolemy's day there even appears to have been a retrogression, as exemplified by the Roman trend away from the mathematical approach to mapping.

Great accumulations of documents and maps were destroyed or lost, and the survival of a large part of Ptolemy's work was probably due to its great prestige and popularity. The only other major work on mapping to survive was Strabo's earlier treatise, albeit with some changes from recopying. Few of the maps and related works of the ancient world have come down to us in their original forms. The tendencies to revise and even recapitulate, when copying manuscripts, are readily understood. Doubtless, the factual content was improved more often than not, but a residual confusion remains when the specimen at hand may be either a true copy of an ancient document or a medieval scholar's version of the subject matter.

THE MIDDLE AGES

Progress in cartography during the early Middle Ages was slight. The medieval mapmaker seems to have been dominated by the church, reflecting in his work the ecclesiastical dogmas and interpretations of Scripture. In fact, during the 6th century Constantine of Antioch created a "Christian topography" depicting the Earth as a flat disk. Thus the Roman map of the world, along with other concepts, continued as authoritative for many centuries. A contemporary Chinese map shows that country occupying most of the world, while the Roman Empire dominates most other maps produced during early Christian times.

Later medieval mapmakers were clearly aware of the Earth's sphericity, but for the most part, maps remained small and schematic, as exemplified by the T and O renderings, so named from the stylized T-form of the major water bodies separating the continents and the O as the circumfluent ocean surrounding the world. The orientation with east at the top of the map was often used, as the word (orientation) suggests.

The earliest navigators coasted from headland to headland; they did not require charts until adoption of the magnetic compass made it possible to proceed directly from one port to another. The earliest record of the magnetic compass in Europe (1187) is followed within a century by the earliest record of a sea chart. This was shown to Louis IX, king of France, on the occasion of his participation in the Eighth Crusade in 1270. The earliest surviving chart dates from within a few years of this event. Found in Pisa and known as the Carta Pisana, it is now in the Bibliothèque Nationale, Paris. Thought to have been made about 1275, it is hand drawn on a sheepskin and depicts the entire Mediterranean Sea. Such charts, often known as portolans named for the portolano or pilot book, listing sailing courses, ports, and anchorages, were much in demand for the increasing trade and shipping. Genoa, Pisa, Venice, Majorca, and Barcelona, among others, cooperated in providing information garnered from their pilots and captains. From repeated revisions, and new surveys by compass, the portolan charts eventually surpassed all preceding maps in accuracy and reliability. The first portolans were hand drawn and very expensive. They were based entirely on magnetic directions and map

Portolans

projections that assumed a degree of longitude equal to a degree of latitude. The assumption did little harm in the Mediterranean but caused serious distortions in maps of higher latitudes. Development of line engraving and the availability, in the 16th century, of large sheets of smooth-surfaced paper facilitated mass production of charts, which soon replaced the manuscript portolans.

Many specimens of portolan charts have survived. Though primarily of areas of the Mediterranean and Black Sea, some covered the Atlantic as far as Ireland, and others the western coast of Africa. Their most striking feature is the system of compass roses (Figure 3), showing directions from various points, and lines showing shortest navigational routes.

Another phenomenon of the late Middle Ages was the great enthusiasm generated by the travels of Marco Polo in the 1270s and 1280s. New information about faraway places, and the stimulation of interest in world maps, promoted their sale and circulation. Marco Polo's experiences also kindled the desire for travel and exploration in others and were, perhaps, a harbinger of the great age of discovery and exploration.

During Europe's Dark Ages Islāmic and Chinese cartography made progress. The Arabs translated Ptolemy's treatises and carried on his tradition. Two Islāmic scholars deserve special note. Ibn Haukal wrote a *Book of Ways and Provinces* illustrated with maps, and al-Idrīsī constructed a world map in 1154 for the Christian king Roger of Sicily, showing better information on Asian areas than had been available theretofore. In Baghdad astronomers used the compass long before Europeans, studied the obliquity of the ecliptic, and measured a part of the Earth's meridian. Their sexagesimal (based on 60) system has dominated cartography since, in the concept of a 360-degree circle.

Mapmaking, like so many other aspects of art and science, developed independently in China. The oldest known Chinese map is dated about 1137. Most of the area that is now included in China had been mapped in crude form before the arrival of the Europeans. The Jesuit missionaries of the 16th century found enough information to prepare an atlas, and Chinese maps thereafter were influenced by the West.

THE AGE OF DISCOVERY AND EXPLORATION

Revival of Ptolemy. The fall of Byzantium sent many refugees to Italy, among them scholars who had preserved some of the old Greek manuscripts, including Ptolemy's *Geography,* from destruction. The rediscovery of this great work came at a fortunate time because the recent development of a printing industry capable of handling map reproduction made possible its circulation far beyond the few scholars who otherwise would have enjoyed access to it. This, together with a general reawakening of scholarship and interest in exploration, created a golden era of cartography.

The *Geography* was translated into Latin about 1405. Although it had not been completely lost (the Arabs had preserved portions of it), recovery of the complete work, with maps, greatly stimulated general interest in cartography. About 500 copies of the *Geography* were printed at Bologna in 1477, followed by other editions printed in Germany and Italy. The printing process, in addition to permitting the wide diffusion of geographic knowledge, retained the fidelity of the original works. By 1600, 31 Latin or Italian editions had been printed.

Printing of Ptolemy's *Geography*

Maps of the discoveries. Progress in other technologies such as navigation, ship design and construction, instruments for observation and astronomy, and general use of the compass tended continuously to improve existing map information, as well as to encourage further exploration and discovery. Accordingly, geographic knowledge was profoundly increased during the 15th and 16th centuries. The great discoveries of Columbus, da Gama, Vespucci, Cabot, Magellan, and others gradually transformed the world maps of those days. "Modern" maps were added to later editions of Ptolemy. The earliest was a map of northern Europe drawn at Rome in 1427 by Claudius Claussön Swart, a Danish geographer. Cardinal Nicholas Krebs drew the first modern map of Germany, engraved in 1491. Martin Waldseemüller of St. Dié prepared an edition with more than 20 modern maps in 1513. Maps showing new discoveries and information were at last transcending the classical treatises of Ptolemy.

The most important aspect of postmedieval maps was their increasing accuracy, made possible by continuing exploration. Another significant characteristic was a trend toward artistic and colourful rendition, for the maps still had many open areas in which the artist could indulge his imagination. The cartouche, or title block, became more and more elaborate, amounting to a small work of art. Many of the map editions of this age have become collector's items. The first map printings were made from woodcuts. Later they were engraved on copper, a process that made it possible to reproduce much finer lines.

Archivio di Stato di Firenze, Carte nautiche n. 1. Autorizzazione del Ministero dell'Interno, Direzione Generale degli Archivi di Stato (Italia), parere n. 526 del 1971

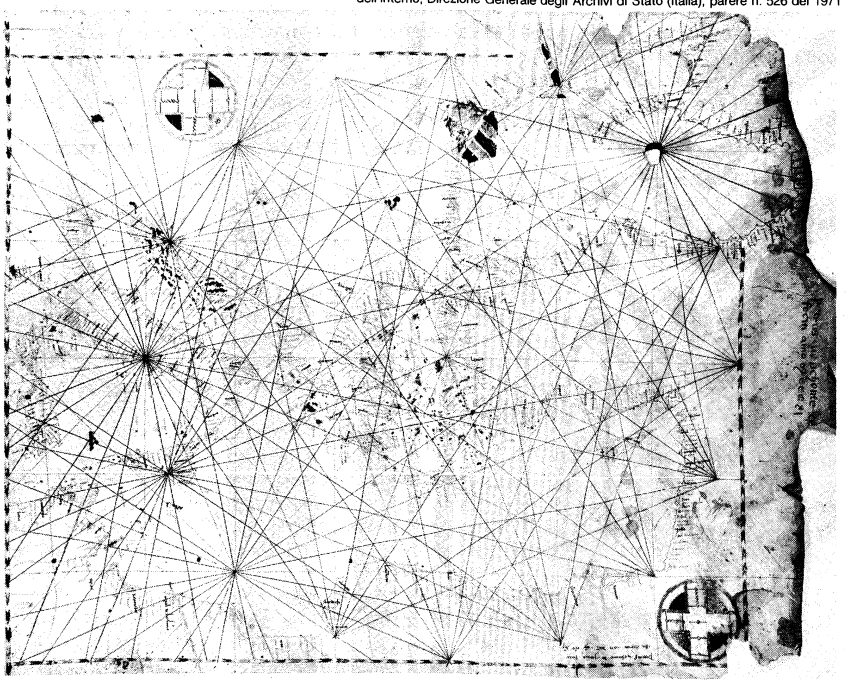


Figure 3: Portolan chart of the eastern Mediterranean area by Petrus Vesconte, 1311.

The finished plates were inked and wiped, leaving ink in the cut lines. Dampened paper was then pressed on the plate and into the engraved line work, resulting in very fine impressions. The process remained the basis of fine map reproduction until the comparatively recent advent of photolithography.

The *Cosmographiae,* textbooks of geography, astronomy, history, and natural sciences, all illustrated with maps and figures, first appeared in the 16th century. One of the earliest and best known was that of Petrus Apianus in 1524, the popularity of which extended to 15 more editions. That of Sebastian Münster, published in 1544, was larger and remained authoritative and in demand until the end of the century, reflecting the general eagerness of the times for learning, especially geography.

Mercator

The foremost cartographer of the age of discovery was Gerhard Kremer, known as Gerardus Mercator, of Flanders. Well educated and a student of Gemma Frisius of Louvain, a noted cosmographer, he became a maker of globes and maps. His map of Europe, published in 1554, and his development of the projection that bears his name made him famous. The Mercator projection solved an age-old problem of navigators, enabling them to plot bearings as straight lines.

Other well-known and productive cartographers of the Dutch-Flemish school are Abraham Ortelius of Antwerp, who prepared the first modern world atlas in 1570, and Jadocus Hondius. Early Dutch maps were among the best for artistic expression, composition, and rendering. Juan de la Cosa, the owner of Columbus' flagship, *Santa María,* in 1500 produced a map recording Columbus' discoveries, the landfall of Cabral in Brazil, Cabot's voyage to Canada, and da Gama's route to India. The first map showing North and South America clearly separated from Asia was produced in 1507 by Martin Waldseemüller. An immense map, 4½ by 8 feet (1.4 by 2.4 metres), printed in 12 sheets, it is probably the first map on which the name America appeared, indicating that Waldseemüller was impressed by the account written by the Florentine navigator Amerigo Vespucci.

In 1529 Diego Ribero, cosmographer to the king of Spain, made a new chart of the world on which the vast extent of the Pacific was first shown. Survivors of Magellan's circumnavigation of the world had arrived in Seville in 1522, giving Ribero much new information.

The first known terrestrial globe that has survived was made by Martin Behaim at Nürnberg in 1492. Many others were made throughout the 16th century. The principal centres of cartographic activity were Spain, Portugal, Italy, the Rhineland, the Netherlands, and Switzerland. England and France, with their growing maritime and colonial power, were soon to become primary map and chart centres. Capt. John Smith's maps of Virginia and New England, the first to come from the English colonies, were published in London in 1612.

18TH CENTURY TO THE PRESENT

A reformation of cartography that evolved during the 18th century was characterized by scientific trends and more accurate detail. Monsters, lions, and swash lines disappeared and were replaced by more factual content. Soon the only decorative features were in the cartouche and around the borders. The map interiors contained all the increasing information available, often with explanatory notes and attempts to show the respective reliabilities of some portions.

Application of science to mapmaking

Where mapmakers formerly had sought quick, profitable output based on information obtained from other maps and reports of travelers and explorers, the new French cartographers were scientists, often men of rank and independent means. For expensive ventures, such as the triangulation of two degrees of a meridian to determine the Earth's size more accurately, they were subsidized by the king or the French Academy. Similar trends were developing across Europe.

The new cartography was also based on better instruments, the telescope playing an important part in raising the quality of astronomical observations. Surveys of much higher accuracy were now feasible. The development of the chronometer (an accurate timepiece) made the computation of longitude much less laborious than before; much more information on islands and coastal features came to the map and chart makers.

The rise of national surveys. The development in Europe of power-conscious national states, with standing armies, professional officers, and engineers, stimulated an outburst of topographic activity in the 18th century, reinforced to some extent by increasing civil needs for basic data. Many countries of Europe began to undertake the systematic topographic mapping of their territories. Such surveys required facilities and capabilities far beyond the means of private cartographers who had theretofore provided for most map needs. Originally exclusively military, national survey organizations gradually became civilian in character. The Ordnance Survey of Britain, the Institut Géographique National of France, and the Landestopographie of Switzerland are examples.

In other countries, such as the United States, where defense considerations were not paramount, civilian organizations—*e.g.,* the U.S. Geological Survey and the National Ocean Service (originally Survey)—were assigned responsibility for domestic mapping tasks. Only when World War II brought requirements for the mapping of many foreign areas did the U.S. military become involved on a large scale, with the expansion of the Oceanographic Office (Navy), Aeronautical Chart Service (Air Force), and the U.S. Army Topographic command.

Elaborate national surveys were undertaken only in certain countries. The rest of the world remained largely unmapped until World War II. In some instances colonial areas were mapped by military forces, but except for the British Survey of India, such efforts usually provided piecemeal coverage or generalized and sketchy data. Some important national surveys will be outlined briefly.

Surveys of colonies

The work in France was organized by the French Academy, and in 1748 the *Carte géométrique de la France,* comprising 182 sheets, was authorized. Most of the field observations were accomplished by military personnel. The new map of France as a whole, drawn after the new positions had been computed, caused Louis XV to remark that the more accurate data lost more territory than his wars of conquest had gained. Napoleon, an ardent map enthusiast, planned a great survey of Europe on a 1:100,000 scale, which was well under way when he was overthrown.

During the 18th century Great Britain became the foremost maritime power of Europe, and the Admiralty sponsored many developments in charting as well as improvements in navigation facilities. Because of the Admiralty's prestige, other maritime nations accepted its proposal that the prime meridian for longitude reference should pass through Greenwich. Other achievements in early oceanography were Edmond Halley's magnetic chart, which has been continuously revised from new data. Later similar charts for currents, tides, and prevailing winds were developed.

French progress in mapping stimulated the English to start a national survey, which was completed in 1787. The Ordnance Survey was organized in 1791, and the first sheet, on the scale of one inch to the mile, was published in 1801. When the original coverage was completed, a new series, at six inches to the mile, was begun; England was soon the best mapped country in the world. An earlier "first" was John Ogilby's *Britannia,* published in 1675, an atlas of road strip maps plotted by odometer and compass, presaging the modern road map.

A survey of Spain was started in the 18th century. Surveys of several German principalities were combined after unification into the Reichskarte at 1:100,000 scale. A topographic survey of Switzerland was begun in 1832. An Austrian series was started in 1806, from which the Specialkarte, later considered the most detailed maps of Europe, were derived. In China, under the Communist regime, survey and cartography groups have provided coverage of much of the country with a new 1:50,000-scale map series. Japan established an Imperial Land Survey in 1888, and by 1925, topographic coverage of the home islands, at a scale of 1:50,000, was complete.

International Map of the World (IMW). The International Geographical Congress in 1891 proposed that the participating countries collaborate in the production of a 1:1,000,000-scale map of the world. Specifications and format were soon established, but production was slow in the earlier years since it was first necessary to complete basic surveys for the required data, and during and after World War II there was little interest in pursuing the project. The intention to complete the series was reestablished, however, and many countries have returned to the task. By the mid-1980s the project was nearing completion.

World War II and after. World War I, and to a much greater extent World War II, brought great progress in mapping, particularly of the unmapped parts of the Earth; an appraisal by the U.S. Air Force indicated that in 1940 less than 10 percent of the world was mapped in sufficient detail for even the meagre requirements of pilot charts. A major program of aerial photography and reconnaissance mapping, employing what became known as the trimetrogon method, was developed. Vast areas of the unmapped parts of the world were covered during the war years, and the resulting World Aeronautical Charts have provided generalized information for other purposes since that time. Many countries have used the basic data to publish temporary map coverage until their more detailed surveys can be completed.

World Aeronautical Charts

The Cold War atmosphere of the 1940s and 1950s promoted a continuation of militarily oriented mapping. Both NATO and Warsaw Pact countries continued to improve their maps; NATO developed common symbols, scales, and formats so that maps could be readily exchangeable between the forces of member countries. Postwar economic development programs, in which maps were needed for planning road, railroad, and reservoir constructions, also stimulated much work. The United Nations provides advisory assistance in mapping to countries wishing it.

Among other collaborations, the Inter-American Geodetic Survey, in which the U.S. Army provides instruction and logistic support for mapping, was organized. Although this cooperation primarily involved Latin-American countries, similar arrangements were made with individual countries in other parts of the world. Cooperation and exchange of data in hydrographic surveys, aeronautical charting, and other fields has continued.

Although some terrain data are available for practically all of the world, the data for many sectors remain sketchy. Surveys of Antarctica by the several countries active there are in progress, but the continent will not be completely mapped for some years. The goal of most countries is to achieve adequate coverage for general development needs. Much remains to be done. Even in countries like the United States that have not yet completed the initial coverage, many of the maps prepared in earlier years are already in need of revision. Thus, even when mapping is completed, requirements for greater detail and revision will continue to make demands upon the funds available.

Aerial photography, which permits accurate and detailed work within feasible cost ranges, has dominated basic mapping in recent years. During World War I aerial photography was used for reconnaissance mapping, and after the war rapid progress was made in optics, cameras, plotting devices, and related equipment. By World War II much of the highly sophisticated equipment now in use had been designed. Electronic distance-measuring devices have made field surveys easier and more accurate, while much improved circle graduation has made theodolites (transits) lighter as well as more precise. Computers and automation, which together have transformed the mapping procedures of yesterday, are described below in the section *Modern mapmaking techniques.*

Mapmaking

ELEMENTS

Map design is a twofold process: (1) the determination of user requirements, with attendant decisions as to map content and detail, and (2) the arrangement of content, involving publication scale, standards of treatment, symbolizations, colours, style, and other factors. To some extent user requirements obviously affect standards of treatment, such as publication scale. Otherwise, the latter elements are largely determined on the basis of efficiency, legibility, aesthetic considerations, and traditional practices.

In earlier productions by individual cartographers or small groups, personal judgments determined the nature of the end product, usually with due respect for conventional standards. Map design for large programs, such as the various national map series of today, is quite formal by comparison. In most countries, the requirements of official as well as private users are carefully studied, in conjunction with costs and related factors, when considering possible changes or additions to the current standards.

Study of user requirements

Requirements of military agencies often have a decisive influence on map design, since it is desirable to avoid the expense of maintaining both civil and military editions of maps. International organizations and committees are additional factors in determining map design. The fact that development of changes in design and content of national map series may become rather involved induces some reluctance to change, as does the fact that map stocks are usually printed in quantities intended to last for 10 or more years. Also, frequent changes in treatments result in extensive overhauls at reprint time, with consequent inconsistencies among the standing editions.

Planning for the production of a national series involves both technical and program considerations. Technical planning involves the choice of a contour interval (the elevation separating adjacent contour lines, or lines of constant elevation), which in turn determines the height of aerial photography and other technical specifications for each project. The sequence of mapping steps, or operational phases, is determined by the overall technical procedures that have been established to achieve the most efficiency.

The program aspects of planning involve fiscal allotments, priorities, schedules, and related matters.

Production controls also play important roles in large programs, where schedules must be balanced with capacities available in the respective phases to avoid backlogs or dormant periods between the mapping steps. Considering that from authorization to final printing, topographic maps may require two years or more to complete, the importance of careful planning is evident. Many factors, including the weather, can converge to cause delays.

Map scales and classifications. Map scale refers to the size of the representation on the map as compared to the size of the object on the ground. The scale generally used in architectural drawings, for example, is 1/4 inch to one foot, which means that 1/4 of an inch on the drawing equals one foot on the building being drawn. The scales of models of buildings, railroads, and other objects may be one inch to several feet. Maps cover more extensive areas, and it is usually convenient to express the scale by a representative fraction or proportion, as 1/63,360, 1:63,360, or "one-inch-to-one-mile." The scale of a map is smaller than that of another map when its scale denominator is larger: thus, 1:1,000,000 is a smaller scale than 1:100,000. Most maps carry linear, or bar, scales in one or more margins or in the title blocks.

Nautical charts are constructed on widely different scales and can be generally classified as follows: ocean sailing charts are small-scale charts, 1:5,000,000 or smaller, used for planning long voyages or marking the daily progress of a ship. Sailing charts, used for offshore navigation, show a generalized shoreline, only offshore soundings, and are at a scale between 1:600,000 and 1:5,000,000. As an illustration of chart use, a 10-knot ship covers about 29 inches (74 centimetres) at 1:600,000 scale in a day.

General charts, as exemplified by Figure 4, are used for coastwise navigation outside outlying reefs and shoals and are at a scale between 1:100,000 and 1:600,000. Coast charts are intended for use in leaving and entering port or navigating inside outlying reefs or shoals and are at a scale between 1:50,000 and 1:100,000. Harbour charts are for use in harbours and small waterways, with a scale usually larger than 1:50,000.

In rare instances reference may be made to the areal scale of a map, as opposed to the more common linear scale.

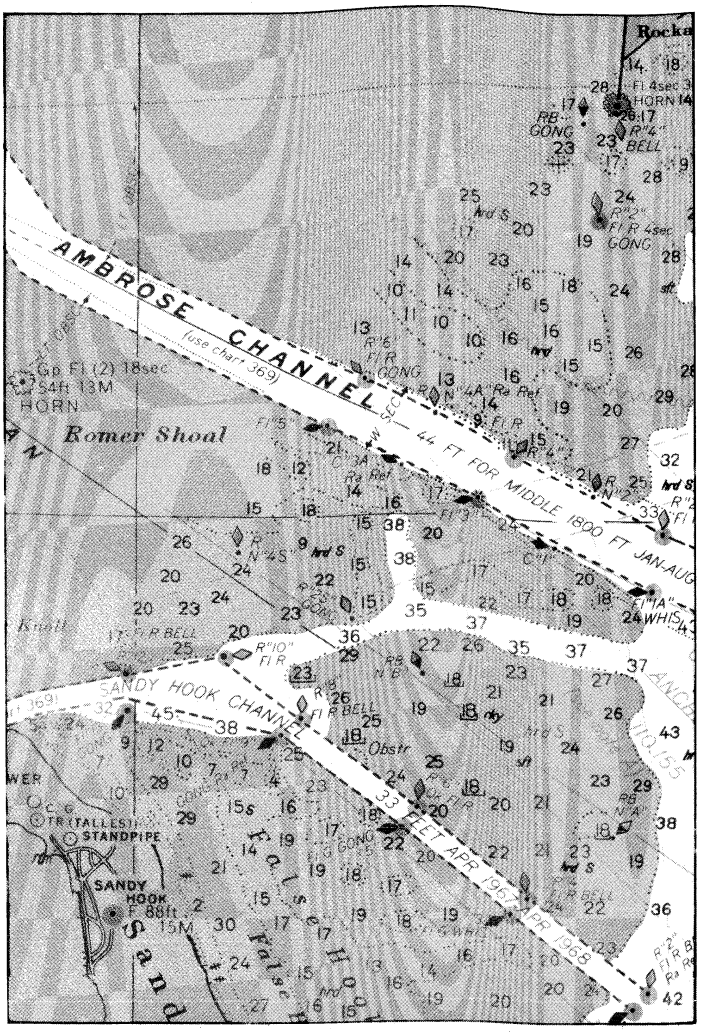


Figure 4: Small section of a large hydrographic chart (U.S. Coast and Geodetic Survey No. 1215) of the approaches to New York harbour. The scale is 1:80,000.

By courtesy of the U.S. Department of Commerce, Environmental Science Services Administration

In such cases the denominator of the fractional reference would be the square of the denominator of the linear scale.

The linear scale may vary within a single map, particularly if the scale is small. Variations in the scale of a map because of the sphericity of the surface it represents may, for practical purposes, be considered as nil. On maps of very large scale, such as 1:24,000, such distortions are negligible (considerably less than variations in the paper from fluctuations of humidity). Precise measurements for engineering purposes are usually restricted to maps of that scale or larger. As maps descend in scale, and distortions inherent to their projection of the spherical surface increase, less accurate measurements of distances may be expected.

Types of classification

Maps may be classified according to scale, content, or derivation. The latter refers to whether a map represents an original survey or has been derived from other maps or source data. Some contain both original and derived elements, usually explained in their footnotes. Producing agencies, technical committees, and international organizations have variously classed maps as large, medium, or small scale. In general, large scale means inch-to-mile and larger, small scale, 1:1,000,000 and smaller, leaving the intermediate field as medium scale. As with most relative terms, these can occasionally lead to confusions but are useful as one practical way to classify maps.

The nature of a map's content, as well as its purpose, provides a primary basis of classification. The terms aeronautical chart, geologic, soil, forest, road, and weather map make obvious their respective contents and purposes. Maps are therefore often classified by the primary purposes they serve. Topographic maps usually form the background for geologic, soil, and similar thematic maps and provide primary elements of the bases upon which many other kinds of maps are compiled.

Map projections. A great variety of map projections has been devised to provide for the various properties that may be desired in maps. In effect, a projection is a systematic method of drawing the Earth's meridians and parallels on a flat surface. Some projections have equal-area properties, while others provide for conformal delineations in which, for small areas, the shape is practically the same as it would be on a globe. Only on a globe can areas and shapes be represented with true fidelity. On flat maps of very large areas, distortions are inevitable. These effects may be minimized by selecting the projection best suited to the purpose of the map to be produced.

Most types of projection can be grouped according to their geometric derivations as cylindrical, conic, or azimuthal (Figure 5). A few cannot be so related or are combinations of these. Terms such as network, graticule, or grid might have been preferable to describe the transposition of meridians and parallels from globe to flat surface, since few systems are actually derived by projection, and most in fact have been formulated by analytic and mathematical processes. The term projection, however, is well established and has some merit in helping the layman to understand the problems and solutions. The theory of trigonometric surveying was disclosed in 1533 by Gemma Frisius, a Flemish mathematician. In 1569 Gerardus Mercator solved the projection problem by producing his famous world map with the meridians vertical and parallels having increased spacing in proportion to the secant (a trigonometric function) of the latitude. Edward Wright published mathematical tables (1599) giving the basis of Mercator's projection. Tables for the construction of other commonly used projections have been developed by mapping agencies.

Mercator projection

Cylindrical projections treat the Earth as a cylinder on which parallels are horizontal lines and meridians appear as vertical lines. The familiar Mercator projection is of this class and has many advantages in spite of the great distortions that it causes in the higher latitudes. Compass bearings may be plotted as straight segments on these projections, which have been traditionally used for nautical charts. On cylindrical projections places of similar latitude appear at the same height. Parallels and meridians may, if desired, be omitted from the body of the map and instead simply indexed at the margins, while lettering can be placed horizontally rather than in a curve. Among the variations of cylindrical projections is the Transverse Mercator, in which the cylinder is tangent to the Earth not along the Equator but along a chosen meridian, a treatment that has advantages in drawing maps that are long in the north–south direction.

Virtually all navigational charts are constructed on the ordinary Mercator projection; the only navigational charts not on ordinary Mercator projections are great-circle charts and charts of the polar regions. Great-circle charts, which are maps of large areas, such as the entire Pacific

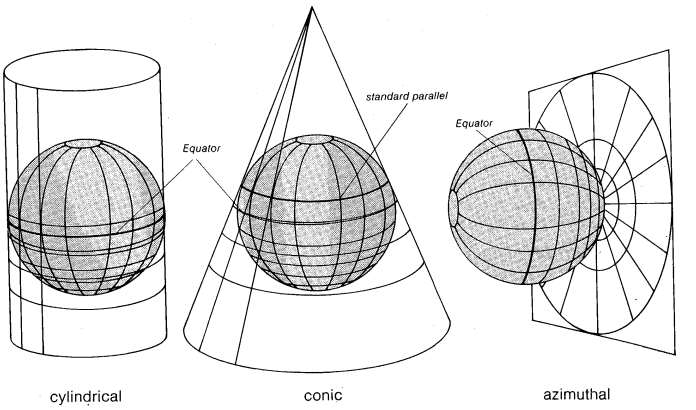


Figure 5: Derivations of basic projections.

Ocean, are ordinarily on very small scales with gnomonic projection. The navigator uses them to lay out a track between ports perhaps thousands of miles apart and then transfers the latitudes corresponding, for example, to each 5° of longitude, to his ocean sailing chart. He thus arrives at a series of short rhumb-line courses, each of which makes the same angle with all meridians, that closely approximate the shortest distance between the two ports.

Conic projections are derived from a projection of the globe on a cone drawn with the point above either the North or South Pole and tangent to the Earth at some standard or selected parallel. Occasionally the cone is arranged to intersect the Earth at two closely spaced standard parallels. A polyconic projection, used in large-scale map series, treats each band of maps as part of a cone tangent to the globe at the particular latitude.

Azimuthal, or zenithal, projections picture a portion of the Earth as a flattened disk, tangent to the Earth at a specified point, as viewed from a point at the centre of the Earth, on the opposite side of the Earth's surface, or from a point far out in space. If the perspective is from the centre of the Earth, the projection is called gnomonic; if from the far side of the Earth's surface, it is stereographic; if from space, it is called orthographic (Figure 6).

Modified from I. Fisher and O.M. Miller, *World Maps and Globes*

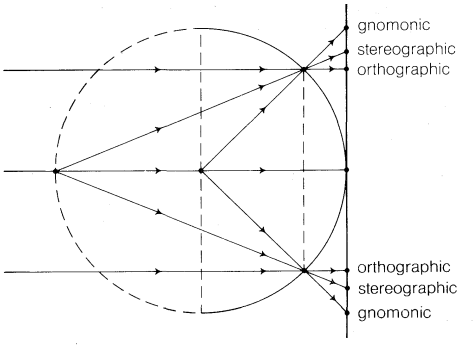


Figure 6: Basis of gnomonic, stereographic, and orthographic projections.

A type of projection often used to show distances and directions from a particular city is the Azimuthal Equidistant. Such measurements are accurate or true only from the selected central point to any other point of interest.

The polar projection is an azimuthal projection drawn to show Arctic and Antarctic areas. It is based on a plane perpendicular to the Earth's axis in contact with the North or South Pole. It is limited to 10 or 15 degrees from the poles. Parallels of latitude are concentric circles, while meridians are radiating straight lines.

Development of reference spheroids. Tables from which map projections of the more familiar kinds may be plotted have been available for some years and have been based on the best determinations of the size and shape of the Earth available at the time of their compilation. The dimensions of Clarke's Spheroid (introduced by the British geodesist Alexander Ross Clarke) of 1866 have been much used in polyconic and other tables. A later determination by Clarke in 1880 reflected the several geodetic surveys that had been conducted during the interim. An International Ellipsoid of Reference was adopted by the Geodetic and Geophysical Union in 1924 for application throughout the world.

Clarke's Spheroid

The development of electronic distance-measuring systems has facilitated geodetic surveys. During the late 20th century, satellite observation and international collaborations have led to an accurate determination of the size and shape of the Earth and to the possibility of adjusting all existing primary geodetic surveys and astronomical observations to a single world datum.

Geographic and plane coordinate systems. The standard geographic coordinate system of the world involves latitudes north or south of the Equator and longitudes east or west of the Prime Reference Meridian of Greenwich. Map and control point references are stated in degrees, minutes, and seconds carried to the number of decimal places commensurate with the accuracy to which locations have been established.

Geodetic surveys, being of extensive areas, must be adjusted for the Earth's curvature, and reductions must be made to mean sea level for scale. The computations are therefore somewhat involved. As a convenience for engineers and surveyors, many countries have established official plane coordinate systems for each province, state, or sector thereof. By this means, all surveys can be "tied" to control points in the system without transposition to geographic coordinates.

In large countries such as the United States, two basic projections are commonly selected to provide systems with minimum distortions for each state or region. For those long in north–south dimension, the Transverse Mercator is generally used, while for those long in east–west direction, the Lambert conformal (intersecting cone) projection is usually employed. In the case of large regions, two or more zones may be established to limit distortions. Positions of geodetic control points have been computed on the plane coordinate systems and have been made available in published lists.

Basic data for compilation. Maps may be compiled from other maps, usually of larger scale, or may be produced from original surveys and photogrammetric compilations. The former are sometimes referred to as derived maps and may include information from various sources, in addition to the maps from which they are principally drawn. Most small-scale series, such as the International Map of the World and World Aeronautical Charts, are compiled from existing information, though new data are occasionally produced to strengthen areas for which little or doubtful information exists. Thus compiled maps may contain fragments of original information while those representing original surveys may include some existing data of higher order, such as details from a city plat.

Road and city map compilations

Road maps, produced by the millions, are compiled from road surveys, topographic maps, and aerial photography. City maps often represent original surveys, made principally to control engineering plans and construction. Some are, however, compiled from enlargements of topographic maps of the area.

Notations regarding the sources from which they were drawn are usually carried on compiled maps. This sometimes includes a reliability diagram showing the areas for which good information was available and those that may be less dependable. Comments regarding certain features or areas, which the editor may deem helpful to the user, may be made in the map itself.

Maps reflecting original surveys, such as a national topographic map series, carry standard marginal information. Date of aerial photography, process and instrumentation employed, notes regarding control and projection, date of field edit, and other information may be included. References to the availability of adjoining maps and those of other scales or series may also be included. Marginal ticks for intervals of plane coordinate systems, military grids, and other reference features are also shown and appropriately labeled.

Symbolization. Symbols are the graphic language of maps and charts that has evolved through generations of cartographers. The symbols doubtless had their origins as simple pictograms that gradually developed into the conventions now generally used.

Early cartographers recognized that common usages and conventions would minimize confusion and to some extent simplify compilation and engraving. Efforts in this direction were made over the years, but cartographers, being artists of a sort, preferred to vary their styles, and effective standardization was not achieved until comparatively recent times. National agencies in most countries established conventions with due regard to practices in other countries. International Map of the World agreements, NATO conventions, and the efforts of the United Nations and of international technical societies aid standardization.

Symbols may be broadly classed as planimetric or hypsographic or may be grouped according to the colours in which they are conventionally printed. Black is used for names and culture, or works of man; blue for water

features, or hydrography; brown for relief, or hypsography; green for vegetation classifications; and red for road classes and special information. There are variations, however, particularly in special-purpose series, such as soil and geologic maps. Symbols will also vary, perforce, because of limitations of space in the smaller scales and the feasibility of drawing some features to true scale on large maps. Legends explain the less obvious symbols on many maps, while explanatory sheets or booklets are available for most standard series, providing general data as well as symbol information. When less familiar symbols are used on maps they are often labeled to prevent misunderstanding. The general located-object symbol, with label, is often used in preference to specific symbols for such objects as windmills and lookout towers for similar reasons.

Planimetric features (those shown in "plan," such as streams, shorelines, and roads) are easier to portray than shapes of land and heights above sea level. Mountains were shown on early maps by sketchy lines simulating profile or perspective appearance as envisioned by the cartographer. Little effort was made at true depiction as this was beyond the scope of available information and existing capabilities. Form lines and hachures, among other devices, were also used in attempting to show the land's shape. Hachures are short lines laid down in a pattern to indicate direction of slope. When it became feasible to map rough terrain in more detail, hachuring developed into an artistic speciality. Some hachured maps are remarkable for their detail and fidelity, but much of their quality depends on the skill of draftsman or engraver. They are little used now, except where relief is incidental.

Hachuring

Contours are by far the most common and satisfactory means of showing relief (see Figure 7). Contours are lines that connect points of equal elevation. The shorelines of lakes and of the sea are contours. Such lines were little used until the mid-19th century, mainly because surveys had not generally been made in sufficient detail for them to be employed successfully. Mean sea level is the datum to which elevations and contour intervals are generally referred. If mean sea level were to rise 20 feet (six metres) the new shoreline would be where the 20-foot contour line is now shown (assuming that all maps on which it is delineated are reasonably accurate).

The quality of contour maps, until recent times, depended largely on the sketching skill of the topographer. In earlier days funds available for topographic mapping were limited, and not much time could be spent in accurate placement. Later, the accurate location of more control points became feasible. An approximate scale of reliability is therefore indicated by the date of a topographic survey, taking into account the respective situations that existed in various countries. Modern surveys, being based on aerial photos and accurate plotting instruments, are generally better in detail and accuracy than earlier surveys. The personal skill of individual topographers, long a factor in map evaluations, has therefore been substantially eliminated.

Hill shading, or shaded relief, layer or altitude tinting, and special manipulations of contouring are other methods of indicating relief. Hill shading requires considerable artistry, as well as the ability to visualize shapes and interpret contours. For a satisfactory result, background contours are a necessary guide to the artist. Hypsographic tinting is relatively easy, particularly since photomechanical etching and other steps can be used to provide negatives for the respective elevation layers. Difficulty in the reproduction process is sometimes a deterrent to the use of treatments involving the manipulation of contours.

In the past, three-dimensional maps were laboriously constructed for studies in military tactics and for many other purposes. They were costly to produce, as contour layers had to be cut and assembled, filled with plaster and painted, after which streams, roads, etc., had to be drawn on the surface. Lettering then was applied, and models of large structures, such as buildings and bridges, were added. In view of the time and cost involved in such productions, they were sparingly used until recent years when better production methods and materials became available. During and after World War I a process was developed and improved whereby an aluminum sheet was "raised" by tapping along the contours copied on its surface. When the contours selected for tapping were completed, the sheet became, in effect, a mold for shaping plastic sheets to its convolutions. The map was printed on plastic sheets prior to the thermal process of shaping them to the mold. Sets of relief maps were soon produced in this manner for use in schools, military briefings, and many other activities.

During and after World War II the production of plastic relief maps was greatly expanded, while the processes and equipment were further improved and refined. Most significant among these developments was a pantograph-router, which cuts a model from plaster or other suitable material as the selected contours are followed by the operator on a topographic map. This eliminated the distortions inherent in shaping metal sheets by the tapping process. Selected topographic maps are now published in limited relief editions for military instruction, special displays, and general classroom instruction.

Pantograph-router

Most relief maps are exaggerated severalfold in the vertical scale. The Earth is remarkably smooth, when viewed in actual scale, and many significant features would hardly be distinguishable on a map without some vertical exaggeration. Mt. Everest, for example, is actually only one-seventh of 1 percent of the Earth's radius in height, or only one-third of an inch (about eight millimetres) at a scale of 1:1,000,000. For this reason relief is usually shown at five, or even 10, times actual scale, depending upon the nature of the area represented. This exaggerated relief scale is always explained in the map legend.

Nomenclature. All possible places and features are identified and labeled to maximize the usefulness of the map. Some names must be omitted, particularly from maps of smaller scales, to avoid overcrowding and poor legibility. The editor must decide which names may be eliminated, while arranging placements so that a maximum number may be accommodated.

Geographic names are the most important, and sometimes the most troublesome, part of the map nomenclature as a whole. Research on existing maps and related documents for a given area may reveal different names for the same features, variations in spelling, or ambiguous applications of names. The field engineer often finds that local usage is confused and sometimes controversial. Various types of official organizations have been established to study the problems submitted and decide the forms and applications that are to be used in government maps and documents. This function is exercised in the United States by the Board on Geographic Names and in the United Kingdom by the Permanent Committee on Geographical Names; worldwide these activities are coordinated by the United Nations Conference on the Standardization of Geographical Names.

By courtesy of U.S. Geological Survey

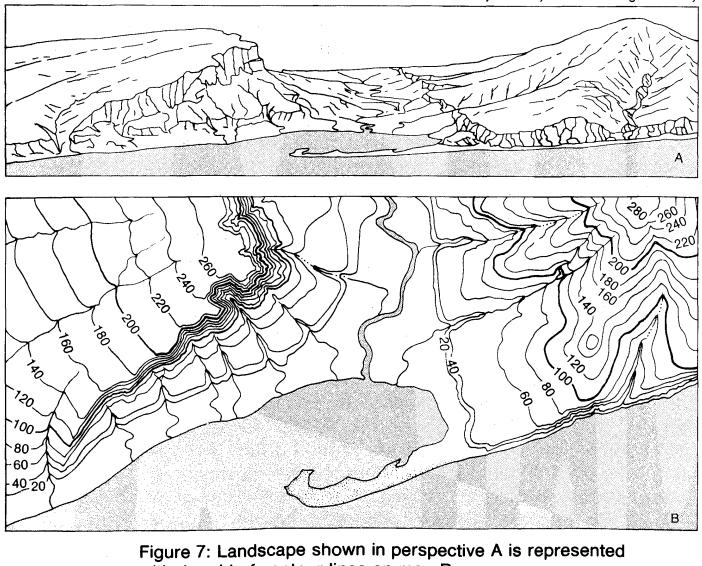


Figure 7: Landscape shown in perspective A is represented with the aid of contour lines on map B.

Toponymics

The science of place-names, or toponymics, has become a significant specialty since World War II, and efforts have been made to establish uniform usages and standards of transliteration throughout the world. Renewed interest in completing the remaining sheets of the International Map of the World, collaborations resulting from military alliances, and efforts of committees of international scientific societies and the United Nations have contributed to these efforts.

At the local levels, however, there are different kinds of problems. The larger scales of most basic topographic map series permit the naming of quite minor hilltops, ridges, streams, and branches, for which designations can be obtained locally. In sparsely settled country few names in actual use may be obtained for minor features, while in other areas inquiries may reveal inconsistencies and confusions in both spelling and application of local names. In some areas, for example, local residents may tend to refer to small streams by the name of the present occupant of the headwater area. The occupants of opposite sides of a mountain sometimes refer to it by different names. In coastal areas the waterman and landsman may use different references for the same features.

A prime opportunity for resolving these problems is presented when a topographic map of an area is prepared for publication. By extensive inquiry and documentation and research of local records and deeds, the appropriate form and application of nearly all names can be determined. Publication and distribution of the map as an official document may then tend to solidify local usage and eliminate the confusions that previously existed.

Lettering

Lettering is selected by the map editor in styles and sizes appropriate to the respective features and the relative importance of each. For topographic maps and most others that follow conventional practice, four basic styles of lettering are used in the Western world. The Roman style is generally used for place-names, political divisions, titles, and related nomenclature. Italic is used for lakes, streams, and other water features. Gothic styles are usually applied to land features such as mountains, ridges, and valleys. Man-made works such as highways, railroads, and canals are usually labeled in slope Gothic capitals, but other distinctive styles are often used for these, together with descriptive notes.

The relative importance of map features is reflected in the different sizes of lettering selected to label them. The most prominent places and features are usually shown in capitals, while lesser ones are labeled with lowercase lettering. In the labeling of cities, however, uppercase lettering is often reserved for state or province capitals. County seats are also labeled in this manner on topographic maps of the United States. For other towns, where lowercase lettering is in order, the sizes selected reflect their relative importance. The use of hand lettering has been abandoned in favour of words and figures printed by type or by a photographic process onto transparent material that is "floated" onto the compilation and anchored by an adhesive wax backing in the proper place. Compass roses and graphic scales are added in the same manner.

TYPES AND USES OF MAPS AND CHARTS

World status of mapping and basic data. Before World War I only a few countries, such as Great Britain, France, and Germany, had detailed maps covering their whole national areas. Now many countries have completed coverage of their territories, while others have carried out small-scale coverage and are beginning engineering surveys in selected areas.

It has been demonstrated that the full potential of map usage in a country, state, or province is not realized until some time after complete coverage has become available. When a modern, detailed map replaces an earlier issue, annual distribution can increase dramatically.

Topographic maps provide the basic data for many other kinds as well as working bases for thematic maps showing geology, soils, and vegetation types. The progress of such mapping in the various parts of the world is therefore a primary indicator of the status of cartography in general. A United Nations survey of the status of world mapping is taken periodically. Inquiries are made to the mapping organizations of all member countries regarding the extent of their respective map coverage, publication scales, and related data.

About a third of the world's land area is now covered by maps at scales of 1:75,000 and larger. Some of such coverage is culturally obsolescent or of low structural quality. An additional third is covered by medium-scale topographic maps; *i.e.*, up to 1:125,000 (about two miles to the inch). Some of this is inferior coverage at medium scales, lacking in geodetic control and topographic detail. This is the case with much of China, but most of the mapping is quite adequate for purposes of reconnaissance and as source information for smaller scale maps.

This provides a general indication of the relative reliability of data contained in such world series maps as the 1:1,000,000-scale aeronautical charts and International Maps of the World. Areas of doubtful information are left blank or are drawn with broken lines. In spite of this dearth of reliable data, most of the IMW sheets have been compiled, and most of the aeronautical pilotage charts have been published, to provide navigation continuity across water areas as well as over unmapped parts of the world.

In some areas, however, large-scale topographic maps are not required. Australia, for example, has large-scale coverage only of its populated coastal areas in the east; in the Outback areas 1:250,000-scale maps are considered adequate for most needs, and a program for their production is well under way. Likewise, large areas of tundra, as in Siberia, deserts in many parts of the world, and other sparsely populated areas may be adequately served with medium- or small-scale coverage until specific development sites require engineering maps.

Nautical chart coverage of the world leaves much to be desired. Good progress has been made, however, on areas bordering the continents and islands. The Arctic, Antarctic, South Pacific, and South Atlantic oceans are the most deficient in good coverage. The Defense Mapping Agency, through agreement with the British Admiralty and other chart-producing countries, maintains worldwide coverage that is constantly updated. The National Ocean Service (originally Survey) maintains charts of U.S. coastal waters. The International Hydrographic Organization (until 1967 Bureau), based at Monaco, attempts to stimulate cooperation in improvement of hydrographic data in general. This organization's General Bathymetric Chart of the Oceans shows existing knowledge and is revised from time to time as new data are accumulated.

International Hydrographic Organization

Coverage of reliable aeronautical charts parallels the availability of topographic maps that provide the essential terrain and cultural data. For this purpose, good 1:250,000-scale maps contain sufficient information for clearance safety and position identification.

Until recently the progress of geodetic triangulation, the basic survey method, was more or less limited to areas either covered by good topographic maps or scheduled for mapping. Preparations for cadastral surveys, where land partition problems abound, have occasionally led to early geodetic programs. Coastal and other surveys also require good basic control to be fully effective; however, it is again the developed and heavily populated areas that are encompassed with the best geodetic surveys. Electronic distance-measuring systems accelerated the progress of geodetic surveys during the 1960s and extended continental schemes over many ocean areas. International cooperation on satellite triangulation is now in progress, with the prospect that existing triangulation of the continents may soon be tied together and adjusted into a single world datum. The Inter-American Geodetic Survey has made progress in the Americas.

In addition to other applications, aerial photographs provide a useful supplement to topographic maps. Indeed, where maps are not available, aerial photographs invariably serve as map substitutes in spite of inherent distortions and lack of elevation data. Most of the world is covered by aerial photography.

During World War II the U.S. Air Force photographed vast areas of the world, providing reconnaissance maps

that were used as bases for aeronautical charts. Much of this information now forms the basis for small-scale map coverage in still remote areas. The system of photography and mapping became known as the trimetrogon process. In it, three wide-angle cameras are used to photograph the terrain from horizon to horizon across the line of flight from an elevation of 20,000 feet (6,100 metres). Detail is usually discernible and plottable for several miles on each side of the line of flight, and occasional points, required for photo-triangulation, can be identified farther out. With higher flight capabilities, wider angle cameras, and lenses of fine resolution the progress of aerial photography has been accelerated. Films have been much improved for fineness of emulsion grain and scale stability. Satellite photography and high-altitude flights with super-wide-angle cameras are now under way in the remaining areas of the world. Infrared and colour film developments have greatly improved photo-interpretation capabilities, providing much better delineations for coastal charts, geologic maps, timber and soil classifications, and other thematic mapping.

Trimetrogon process

Types of maps and charts available. Although the range of maps and charts now available in many countries is so extensive that a complete listing is impractical, any list of the principal types would have to include aeronautical (worldwide and national), congressional or political districts, population distribution, geologic (various scales), highways (national and secondary political units), historical, hydrographic (coastal areas, inland waters, foreign waters), national forests, forest types, public land survey plats, soil, and topographic and related (national and foreign areas).

The National Atlas of the United States of America, published by the Geological Survey in 1970, contains contributions from all of that country's mapping agencies. Summaries are provided of all thematic and economic data of interest. The atlas also indicates where more detailed information or large-scale specialized maps may be obtained. Many countries have centres where detailed information on existing map series and related data may be obtained. In the United States this service is performed by the Map Information Office of the U.S. Geological Survey, which publishes and distributes indexes of each state showing map coverage and ordering information. Summary data on geodetic control and aerial photography are also maintained.

The situation is less complex in other countries where mapping activities are concentrated in one or two organizations; *e.g.,* Ordnance Survey in Great Britain and Institut Géographique National in France. The main agencies can advise where maps produced by others may be obtained. Technical societies maintain large map reference libraries and are prime sources of information, as are the map sections of national libraries and museums.

Government and other mapping agencies. The following are the primary agencies of selected countries having advanced mapping programs.

Military agencies play large roles in the mapping activities of many countries. Frequently, a small cadre of officers administers the mapping facilities, while most of the production personnel are civilian. Many countries, such as Iran and the United States, have both civilian and military organizations that collaborate in developing their respective programs and in performing the actual mapping.

Most countries have private and commercial organizations that produce maps. The widely distributed road maps noted earlier are printed by a few large producers who, in cooperation with others, compile the maps. Very large-scale maps, for road construction and other engineering works, are produced under contract by a number of mapping companies. Some local highway departments have their own photogrammetric units to provide or supplement such productions. City surveys and maps for real-estate developments, tax records, power lines, and so on are largely produced by commercial organizations.

Principal Mapping Agencies of Selected Countries

Australia	Division of National Mapping, Department of National Development
Brazil	Serviço Geográfico do Exercito
Canada	Surveys and Mapping Branch, Department of Energy, Mines and Resources
Chile	Instituto Geográfico Militar
France	Institut Géographique National
Germany, West	Institut für Landeskunde Geodäsie
Iran	Army Geographic Department and National Geographic Organization
U.K.	Ordnance Survey and Directorate of Overseas Surveys
U.S.S.R.	Glavnoye Upravlenie Geodezii i Kartografii

Large societies, such as the American Geographical Society, the National Geographic Society, and the Royal Geographical Society, play important roles in addition to being centres of reference as noted above. The National Geographic Society produces popular small-scale maps of the various regions of the world. The American Geographical Society has compiled many maps, most notably a 1:1,000,000 coverage of Hispanic America on standards similar to those of the International Map of the World. Technical societies, such as the American Congress on Surveying and Mapping, the American Society of Photogrammetry, the American Society of Civil Engineers, and others, lend their support to mapping programs and activities. They issue technical papers and hold frequent meetings where new processes and instrumentation are discussed and displayed. *The Manual of Photogrammetry* and *Journal,* produced by the American Society of Photogrammetry, *Photogrammetria,* published by the International Society for Photogrammetry and Remote Sensing, and *Surveying and Mapping,* published by the American Congress on Surveying and Mapping, are prime examples of important contributions that societies make to the overall progress of mapping.

The work of geographic societies

International organizations. Many societies and other types of organizations are now engaged in activities associated with maps and mapping. In general, they encourage cooperation through meetings and articles in their journals; some are more directly concerned with the dissemination of information on the progress of particular kinds of mapping and charting. Standardizations of map treatments and conventional signs as well as the promotion of progress in technical processes are further objectives of such groups.

The United Nations Office of Cartography plays an important role in all of the activities noted above. It maintains records of progress on the International Map of the World and performs related services formerly handled by the Central Bureau of the IMW. Technical assistance in the development of mapping facilities and programs is provided on request. Occasional regional meetings are arranged for groups of countries having similar problems, while the journal *World Cartography* publishes related papers.

The Inter-American Geodetic Survey is a special unit of the U.S. Corps of Engineers organized to forward the completion of geodetic surveys and mapping in the Americas. Through technical training and assistance with programs, geodetic surveys in Central and South America have been greatly advanced in recent years. Training in photogrammetry is offered and has promoted the establishment of mapping facilities and programs in many of the collaborating countries.

The Pan American Institute of Geography and History has sponsored regular meetings and consultations on cartography, much in the manner of scientific societies. The consultations are held in different countries each year.

The International Hydrographic Bureau was founded in 1921 in Monaco, where it has been headquartered through the years. It serves as a clearinghouse for information related to hydrography and charting and maintains a General Bathymetric Chart of the World, which is revised periodically to include data furnished by the maritime nations participating in their programs and conferences. Other organizations that promote progress in the various aspects of mapping and charting are the International Association of Geodesy, the International Cartographic Association, the International Civil Aviation Organization, the International Geographical Union, the International Federation of Surveyors, the International Society for Photogramme-

General Bathymetric Chart of the World

try and Remote Sensing, and the International Union of Geodesy and Geophysics.

MODERN MAPMAKING TECHNIQUES

Compilation from existing materials. The preparation of derived maps—*i.e.*, maps that are compiled from other maps or existing data—involves the search for, and evaluation of, all extant data pertaining to the subject area. Depending on the nature of the map to be compiled, thoroughgoing research includes boundary references, historical records, name derivations, and other materials. Selection of the most authentic items, on the frequent occasions when some ambiguities are detected, requires careful study and references to related materials. The sources finally selected may require some adjustment or compromise in order to fit properly with adjacent data. When it becomes evident that some sources are of questionable reliability, the cartographer explains this in the margin of his compilation. Sometimes this is placed in the body of the map where the doubtful features or delineations are located.

When selected materials have been assembled they are reduced to a common scale and copied on the compilation base, often in differentiating colours for the respective features. Reductions to a common scale are usually made by photography but may be made by projection and traced directly on the drawing. Minor adjustments may have to be made during compilation even though the source materials are of good quality. In particular, the need to make appropriate generalizations, omitting some details in smaller scale maps, requires much study and judgment.

Except for the new methods of preparing final colour separation plates by scribing (described below), rather than by drafting or copperplate engraving, compilation processes have changed little over the years. Automatic-focusing projectors and better illumination have made the tracing of selected data at compilation scale easier. Better and more extensive facilities for photoreduction and copying, improved light tables, and a wider choice of drafting materials and instruments have served to facilitate compilation. The basic chores of research, selection of best data, and adjustment of these into the compilation however, remain essentially the same.

The preparation of small-scale maps from large ones is sometimes simpler than the process just described, which pertains to compilation from a miscellany of differing sources. The relatively straightforward preparation of 1:62,500-scale maps from those of the 1:24,000-scale series, for example, may require little more than photoreduction and colour-separation drafting, or scribing. Even in this case some generalizations, as well as omission of a few of the least important details, are in order. To avoid the considerable expense involved in such scale conversions, straight photoreduction of colour separation plates appears to be a promising procedure.

Straight photo-reduction

Larger reductions from one map series to another—1:62,500 to 1:250,000 for instance—are more of a problem, since the need for generalization is greater and the omission of many details is involved. The considerable differences in road and other symbol sizes also create displacement problems.

The component maps are reduced, and the negatives are cut and assembled into a mosaic on a clear sheet of plastic, the master negative of which provides guide copy for the several colour-separation plates required, which are then completed for reproduction. More often, however, it is necessary to make an intermediate compilation rather than burden the draftsman with too many adjustments to be made while following copy on the colour-separation plates. The intermediate scale for initial reduction of the component maps provides better legibility than direct photography to reproduction scale. This negative mosaic is copied on a metal-mounted drafting board. A compiler then inks the whole map, usually in three or more contrasting colours. He also draws roads and other symbols at the intermediate size, so that they will reduce to proper dimensions at reproduction scale, and makes the necessary displacement adjustments. Minor features and terrain details to be omitted on the new map are not inked in. The drawing is now ready for photoreduction to the final colour-separation plates, providing much better copy for the draftsman or engraver than direct reduction in one step would have produced.

Most smaller scale map series are prepared from large-scale maps as described above. In earlier days original reconnaissance surveys were made at small scales such as 1:192,000 for publication at 1:250,000. Ideally, the small-scale series of maps should be compiled progressively from those of larger scale and greater detail. Most countries, however, started their mapping programs with relatively small-scale reconnaissance surveys because of economic considerations. Later, affluence and technical competence permitted mapping at larger scales with better accuracy.

Geologic, soil, and other thematic maps usually have a topographic base from which woodland tints and road classification printings have been omitted. Such a map, therefore, has a topographic background printed in subdued colours on which the geologic or soil patterns are overprinted in prominent colours. Small-scale thematic maps showing weather patterns, vegetation types, and a large amount of economic and other information are of similar origin. Backgrounds are drawn from appropriate outline maps of provinces, countries, or regions of the world, while overlaying subject matters are compiled from specialized sources of information.

Use of topographic backgrounds

The generalization of detail is a problem that frequently confronts the cartographer in original mapping and in reducing the scale of existing maps. There are two principal reasons for taking such liberties (or topographic license in the case of the original mapping). The primary purpose is to avoid overcrowding and the resulting poor legibility. In addition, the degree of generalization or detail should be as consistent as possible throughout the map. Generalizations in some parts and excessive detail in others confuse the user and make the map's reliability suspect. Effective generalization requires good judgment based on seasoned knowledge and experience.

In approaching such problems as the thousands of islets in the Stockholm archipelago or the thousands of small lakes in the Alaskan tundra areas, when the map scale will accommodate only a small number, the cartographer may decide to draw the features in groupings that reflect the patterns shown in the large-scale source maps or aerial photos. This is difficult and at best presents the nature of the respective areas rather than a literal portrayal. There is also the possibility that the source maps may already have been generalized by some omissions to accommodate to their own scales. Another device is to note, in appropriate text or marginal references, that many minor lakes or islets are omitted because of scale. Such areas may also be symbolized and explained. The "pattern" representation noted above is actually a form of symbolization.

Intricate coastlines are also extremely difficult to generalize consistently. Here again, the purpose is to omit minor details while retaining the main features and their distinguishing characteristics. These and many equally perplexing questions arise in preparing maps of very small scale from any source. The problems of equalization of detail are also present in such cases. The topographer of earlier days had the equalization problem between areas close at hand and those viewed distantly. In addition, the topographer had to deal with terrain on the far sides of obscuring features.

Photogrammetrists—that is, persons who compile original maps from aerial photos—have similar problems when, for example, one side of a ridge is seen in more detail than the opposite side. Indeed, in steep terrain, parts of the far sides of some mountains are not seen at all. Appropriate steps must be taken in such cases to avoid differing renditions on opposite sides of the mountain. This may be accomplished by adding, in field completion of the manuscript map, the segments not seen by the photogrammetrist; or additional aerial photography, patterned to cover the obscured sectors, may be requested.

Map production from original surveys. The instrumentation, procedures, and standards involved in making original surveys have improved remarkably in recent years. Geodetic, topographic, hydrographic, and cadastral sur-

veys have been facilitated by the application of electronics and computer sciences. At the same time, superior optics and more refined instruments, in general, have enhanced the precision of observations and accuracies of the end products.

The improved quality of surveys has increased the reliability of maps and charts based on them. In turn, the greater output of basic data has accelerated the production of maps and charts, while parallel improvements in processing steps have increased the volume and improved the final product. In a sense the production of maps from original surveys parallels the process steps after a compilation is made from derived sources. This phase is sometimes referred to as map finishing and involves editing, colour separation, and printing. In original surveys for topographic maps and nautical charts, however, the end products are provided for in all the process steps leading to the completed basic manuscript. The manuscript scale is, for example, selected to accommodate the plotting instruments involved as well as the final rendering for printing. In early years it was usual to choose a manuscript scale somewhat larger than that prescribed for publication. This was to allow for some generalization and line refinement in the final reduction. Thus, maps to be published at 1:62,500 scale were plotted in the field at 1:48,000 or thereabouts. With modern photogrammetric instruments, plotting is usually at reproduction scale.

Maps are not directly derived from geodetic surveys, and only land-line plats are produced from cadastral surveys. Accordingly, the primary original map and chart productions are those from topographic and hydrographic surveys. The surveys are somewhat similar as the nautical chart is, in effect, a topographic map of the coast with generalized offshore topography interpolated from depth soundings.

Use of electronic devices in ship surveys

A variety of electronic devices are used to determine a survey ship's precise location while taking soundings, which are also made with electronic equipment. Both hydrographic and topographic surveys now employ aerial photography and precise plotting instruments to develop the base map. In order to simplify the description of modern mapmaking techniques, the process developed for topographic mapping will be described below, with comments where procedures for nautical charts differ significantly. Both processes start by expanding upon the basic control previously established from geodetic surveys.

Surveying, in which the facts are discovered and recorded, must precede mapping, in which the facts are presented in graphic form. Surveying involves (1) global positioning, in which the area to be mapped is located on the Earth's surface, usually by fixing a number of points in the area by astronomical observations or, after the techniques became available, by satellite or radar procedures; (2) establishing the framework, in which these points, and commonly many others connected by some combination of distance and angle measurements, are integrated into an accurately defined structure—like the steel framework of a modern building—on which the detail survey is based; and (3) making the detail survey, which establishes by less accurate (and therefore cheaper) methods the relative positions and shapes of the features being mapped. Constant reference to the framework prevents the errors in the detail survey from accumulating and growing unacceptably large.

Mapping also consists of three steps: (1) fair drawing, in which the accurate but not publishable surveyor's plot is redrawn by a skilled cartographer with uniform lines and lettering and, if a multicoloured map is being produced, is separated into several drawings, one for each colour; (2) reproduction, in which a negative is prepared from each of the fair-drawn originals and special colouring (to represent areas of vegetation, for example) is added; and (3) printing, in which a printing plate is made from each negative, the plates are mounted on a press, proofs (a few trial copies) are made to facilitate correction of errors and blemishes, and the final maps are produced.

Final steps in map preparation. After all the features visible in the aerial photographs have been mapped, the manuscripts are contact printed on coated plastic sheets for review by the field engineer. He examines the whole map, adding such details as houses, trails, and fences that were not visible or were overlooked by the photogrammetrist. Political lines such as state, county, and township limits are located, as are geographic and other names in local use. Roads are classified, and woodland outlines are checked.

Contour accuracy is tested if the operator has noted areas that may be weak. The determination of names involves extensive local inquiry, as do political lines, and both may require research of records.

In remote areas it is more efficient to combine the above activities with supplemental control survey to avoid the extra field phase. Then the photos must be carefully examined and annotated for the compiler, while buildings must be encircled or pricked. Roads are classified and political lines located in the usual manner and noted on the photos or overlays.

Handling corrections

Field corrections are applied to the original manuscripts. They must be scribed (engraved) on the originals so that guide copies can be prepared by contact printing for final colour-separation scribing. At this time all factual detail is carefully checked. Editing may proceed, to conserve time, while the colour-separation scribing is in progress. The editor reviews all names, boundaries, and related data, comparing them to information thereon that may be available from other sources. The editor's function is to see that the map conforms to standard conventions and is clear, legible, and free of errors.

Controversial names, or those found to be in confused or ambiguous spelling or usage, are documented and referred to an appropriate official body. The designation of type styles and sizes as well as placement of lettering is another function of the map editor.

Because modern topographic maps are printed in several colours, separate plates must be prepared for each. Some of the earliest maps were printed from woodcuts, usually in a single colour. Various hand processes were developed through the years, culminating in the fine rendering of copperplate engraving, which dominated the map production industry for many years. The process became obsolete, however, with increased production demands and the development of efficient printing presses. After World War II engraving on glass, and later on coated plastic sheets, was developed to a point that recovered the fineness of copper engraving. These methods of engraving have become firmly established in map production throughout the world.

Scribing process

In the negative engraving or scribing process, guide copy is printed on several sheets of plastic coated with an opaque paint, usually yellow. The scriber follows copy on the respective plates by engraving through the coating. Because arc light can pass only through the engraving scratches, the completed engravings are, in effect, negatives from which the press plates are made. The finest lines (0.002 inch, or 0.05 millimetre, wide), such as intermediate contours, are engraved freehand. Heavier lines, such as index contours, engraved at 0.007 inch, may require a small tripod to assure that the scriber is perfectly vertical. Gravers for double-lined roads, others for buildings, and templates, or patterns, for a variety of symbols are used. Woodland and similar boundaries and shorelines are contact printed and etched on their respective coated sheets, and the areas of the woodland or water are then peeled off, leaving open windows for their respective features. If portions of scrub, orchard, or vineyard are contained in the "woodland" plate, negative sections for these are stripped into their respective locations. Press plates are then processed from the negatives.

A combined-colour proof is then made by successively printing the several completed negatives on a sensitized white plastic sheet that serves for the final checking and review of all aspects of the map. After all corrections have been made, the negatives are ready for the reproduction process.

Nearly all maps are now printed by rotary offset presses, using flexible aluminum-alloy printing plates. The system uses surface plates (very slightly raised or recessed) as opposed to the letterpress and intaglio processes, which involve greater image heights and depths respectively. In

the printing sequence, ink goes from the plate to a rubber blanket to the paper. Thus, the printing plate is positive, or right-reading, as is the printed map. The negatives from which the printing plates are prepared are accordingly wrong-reading. This is the process for so-called surface plates. To retain fineness of line on very long runs (10,000 or more impressions), some map printers prefer "etched" plates, prepared from film positives. Both may be considered essentially surface plates, however, since the respective raise or recess is quite small.

Presses are of many varieties and makes. Huge multicolour types are used in large plants, printing several colours at a time. In effect, a multicolour press is several presses built into one. Each unit has three cylinders for plate, rubber blanket, and paper as well as rollers for water and ink. Presses with automatic feed may produce as many as 6,000 impressions per hour, while hand-fed types are limited to about 2,500 per hour.

Nautical charts are commonly large, 28 by 40 inches (70 centimetres by 1 metre) being an internationally accepted maximum size. In order that a navigator may work with them efficiently, charts must be kept with a minimum of folding in drawers in a large chart table in a compartment of the ship having ready access to the navigating bridge, known as the chart room or chart house. Such structures are not possible in small craft, which therefore require charts of a more convenient size. With the recognition that there are many more small boats in the world, particularly recreational craft, than there are ships and that they are navigated primarily by piloting rather than by celestial or electronic means, many hydrographic offices have given attention to the production of special chart series in a small format for yacht navigators.

SC charts

A typical series is that produced by the U.S. Coast and Geodetic Survey with the designation SC (for small craft). Such charts are only 15 inches (38 centimetres) in the vertical dimension and thus need to be folded only in the vertical direction. Printed on both sides of the sheet, they are oriented along the most probable route rather than by parallels and meridians. Several are stapled together into a stiff cardboard folder for protection. Along with the ordinary chart information, they contain a year of tide tables and information on small-craft facilities in the area. New editions are produced annually.

Practical uses of charts impose some constraints on the selection of colour. Red, for example, would logically be chosen as the color in which to print warnings of navigational hazards. But navigators, who must work at night, prefer to retain the darkness adaptation of their eyes by viewing their charts under red light. Under such illumination, red, orange, and buff are invisible. Hence these colours have been superseded by magenta, purple, and gray.

Charts are working instruments, and, since ships often voyage far from where replacement charts are readily obtainable, hydrographic offices give attention to the quality of the paper on which charts are printed. A ship's reckoning is kept in pencil and erased after each voyage. Thus printing stock that permits multiple erasures is chosen. In view of the environment where charts are used, another quality commonly sought is high wet strength.

Automation in mapping. During the past few decades there was much interest in the automation of mapping processes, and considerable progress was made in this area. Achievements in the fields of electronics, high-speed digital computers, and related technologies provided a favourable period for such progress. In Great Britain, development of a set of procedures utilizing automatic elements, known as the Oxford System, was begun.

Automatic plotting

Some success was also achieved in the difficult area of automatic plotting. Instruments now available can automatically scan a stereo model and generate approximate profiles from which contours may be interpolated. Some steps, however, must be closely monitored or else performed completely by the operator. Contouring interpolated from a profile scan is inferior to an operator's delineation. This contouring meets some less exacting requirements for elevation data, and refinements in the system are improving its precision. The need for human intervention when automatic devices get "lost" is not a decisive drawback, as one operator can monitor several machines. The reduction of tedious and repetitive steps for stereo-operators offers a significant advancement.

Coordinatographs with high repeat accuracies facilitate the automatic plotting of control points and projection intersections. Line work can also be drafted or scribed automatically by the same process, but the respective features must first be coded to provide the necessary input tapes. Automatic colour scanning and discrimination is operational but has not become widely used; it is still necessary for an operator to trace the various features on the manuscript to code them. Obviously, little is to be gained by automatic scribing until the input can be provided automatically. Coded line work can be displayed on a cathode-ray tube and corrected with a light pen, but it is much simpler to check and correct the manuscript or finished drawing. Systems of automatic type placement at present offer only marginal advantages over conventional methods. In short, automation has made substantial advances but has not become fully operational in a practical sense.

An aspect of automation that is developing rapidly concerns graphic data acquisition, storage, and retrieval. Data banks are being accumulated by specialized users of topographic information, often to produce thematic maps showing soil types, vegetation classifications, and a variety of other information. Such data banks are usually organized in two parts: one for line work, such as boundaries, and the other for descriptive information or classifications. Assuming that the necessary inputs have been made to the data bank, special plats can be generated speedily. Examples of such graphics include profiles showing elevations along a selected radio propagation path and cross sections for earthwork on roads and other construction.

(C.F.F./Ed.)

Surveying

Surveying is a means of relatively large-scale, accurate measurement of Earth surfaces. It includes the determination of the measurement data, the reduction and interpretation of the data to usable form, and, conversely, the establishment of relative position and size according to given measurement requirements. Thus surveying has two similar but opposite functions: (1) the determination of existing relative horizontal and vertical position, such as that used for the process of mapping, and (2) the establishment of marks to control construction or to indicate land boundaries.

Surveying has been an essential element in the development of the human environment for so many centuries that its importance is often forgotten. It is an imperative requirement in the planning and execution of nearly every form of construction. Surveying was essential at the dawn of history, and some of the most significant scientific discoveries could never have been implemented were it not for the contribution of surveying. Its principal modern uses are in the fields of transportation, building, apportionment of land, and communications.

Except for minor details of technique and the use of one or two minor hand-held instruments, surveying is much the same throughout the world. The methods are a reflection of the instruments, manufactured chiefly in Switzerland, Austria, Great Britain, the United States, Japan, East Germany, and West Germany. Instruments made in Japan are similar to those made in the West.

HISTORY

Egyptian and other ancient devices

It is quite probable that surveying had its origin in ancient Egypt. The Great Pyramid of Khufu at Giza was built *c.* 2700 BC, 755 feet long and 480 feet high. Its nearly perfect squareness and north–south orientation affirm the ancient Egyptians' command of surveying.

Evidence of some form of boundary surveying as early as 1400 BC has been found in the fertile valleys and plains of the Tigris, Euphrates, and Nile rivers. Clay tablets of the Sumerians show records of land measurement and plans of cities and nearby agricultural areas. Boundary stones

marking land plots have been preserved. There is a representation of land measurement on the wall of a tomb at Thebes (1400 BC) showing head and rear chainmen measuring a grainfield with what appears to be a rope with knots or marks at uniform intervals. Other persons are shown. Two are of high estate, according to their clothing, probably a land overseer and an inspector of boundary stones.

There is some evidence that, in addition to a marked cord, wooden rods were used by the Egyptians for distance measurement. They had the groma, which was used to establish right angles. It was made of a horizontal wooden cross pivoted at the middle and supported from above (Figure 8, left). From the end of each of the four arms hung a plumb bob. By sighting along each pair of plumb bob cords in turn, the right angle could be established. The device could be adjusted to a precise right angle by observing the same angle after turning the device approximately 90°. By shifting one of the cords to take up half the error, a perfect right angle would result.

There is no record of any angle-measuring instruments of that time, but there was a level (Figure 8, right) consisting of a vertical wooden A-frame with a plumb bob supported at the peak of the A so that its cord hung past

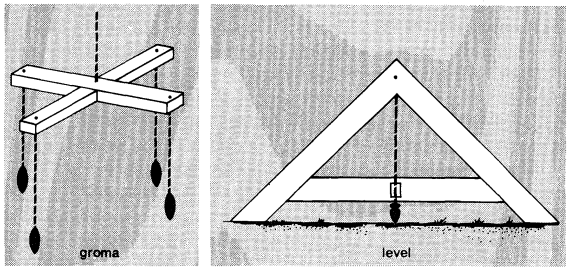


Figure 8: Egyptian surveying instruments.

an indicator, or index, on the horizontal bar. The index could be properly placed by standing the device on two supports at approximately the same elevation, marking the position of the cord, reversing the A, and making a similar mark. Halfway between the two marks would be the correct place for the index. Thus, with their simple devices, the ancient Egyptians were able to measure land areas, replace property corners lost when the Nile covered the markers with silt during floods, and build the huge pyramids to exact dimensions.

Greek log line and Chinese lodestone

The Greeks used a form of log line for recording the distances run from point to point along the coast while making their slow voyages from the Indus to the Persian Gulf about 325 BC. The magnetic compass was brought to the West by Arab traders in the 12th century AD. The astrolabe was introduced by the Greeks in the 2nd century BC. An instrument for measuring the altitudes of stars, or their angle of elevation above the horizon, took the form of a graduated arc suspended from a hand-held cord. A pivoted pointer that moved over the graduations was pointed at the star. The instrument was not used for nautical surveying for several centuries, remaining a scientific aid only.

During their occupation of Egypt, the Romans acquired Egyptian surveying instruments, which they improved slightly and to which they added the water level and the plane table. About 15 BC the Roman architect and engineer Vitruvius mounted a large wheel of known circumference in a small frame, in much the same fashion as the wheel is mounted on a wheelbarrow; when it was pushed along the ground by hand it automatically dropped a pebble into a container at each revolution, giving a measure of the distance traveled. It was, in effect, the first odometer.

The water level consisted of either a trough or a tube turned upward at the ends and filled with water. At each end there was a sight made of crossed horizontal and vertical slits. When these were lined up just above the water level, the sights determined a level line accurate enough to establish the grades of the Roman aqueducts. In laying out their great road system, the Romans are said to have used the plane table. It consists of a drawing board mounted on a tripod or other stable support and of a straightedge—usually with sights for accurate aim (the alidade) to the objects to be mapped—along which lines are drawn. It was the first device capable of recording or establishing angles. Later adaptations of the plane table had magnetic compasses attached.

Plane tables were in use in Europe in the 16th century, and the principle of graphic triangulation and intersection was practiced by surveyors. In 1615 Willebrord Snell, a Dutch mathematician, measured an arc of meridian by instrumental triangulation.

In 1620 English mathematician Edmund Gunter developed a surveying chain, which was only superseded by the steel tape in the beginning of the 20th century.

The study of astronomy resulted in the development of angle-reading devices that were based on arcs of large radii, making such instruments too large for field use. With the publication of logarithmic tables in 1620, portable angle-measuring instruments came into use. They were called topographic instruments, or theodolites. They included pivoted arms for sighting and could be used for measuring both horizontal and vertical angles. Magnetic compasses may have been included on some.

17th-century portable instruments

The vernier, an auxiliary scale permitting more accurate readings (1631), the micrometer microscope (1638), telescopic sights (1669), and spirit levels (about 1700) were all incorporated in theodolites by about 1720. Stadia hairs were first applied by James Watt in 1771. The development of the circle-dividing engine about 1775, a device for dividing a circle into degrees with great accuracy, brought one of the greatest advances in surveying methods, as it enabled angle measurements to be made with portable instruments far more accurately than had previously been possible.

By the late 18th century modern surveying can be said to have begun. One of the most notable early feats of surveyors was the measurement in the 1790s of the meridian from Barcelona, Spain, to Dunkirk, Fr., by two French engineers, Jean Delambre and Pierre Méchain, to establish the basic unit for the metric system of measurement.

The start of modern surveying

Many improvements and refinements have been incorporated in all the basic surveying instruments. These have resulted in increased accuracy and speed of operations and have opened up possibilities for improved methods in the field. In addition to modification of existing instruments, two revolutionary mapping and surveying changes have been introduced: photogrammetry, or mapping from aerial photographs (about 1920), and electronic distance measurement, including the adoption of the laser for this purpose as well as for alignment (in the 1960s). Important technological developments benefiting surveying in the 1970s included the use of satellites as reference points for geodetic surveys and electronic computers to speed the processing and recording of survey data. (J.Ly./Ed.)

MODERN SURVEYING

Basic control surveys. Geodetic surveys involve such extensive areas that allowance must be made for the Earth's curvature. Baseline measurements for classical triangulation (the basic survey method that consists of accurately measuring a base line and computing other locations by angle measurement) are therefore reduced to sea-level length to start computations, and corrections are made for spherical excess in the angular determinations. Geodetic operations are classified into four "orders," according to accuracy, the first-order surveys having the smallest permissible error. Primary triangulation is performed under rigid specifications to assure first-order accuracy.

Efforts are now under way to extend and tie together existing continental networks by satellite triangulation so as to facilitate the adjustment of all major geodetic surveys into a single world datum and determine the size and shape of the Earth spheroid with much greater accuracy than heretofore obtained. At the same time, current national networks will be strengthened, while the remaining amount of work to be done may be somewhat reduced. Satellite triangulation became operational in the United States in 1963 with observations by Rebound A-13, launched that year, and some prior work using the Echo

1 and Echo 2 passive reflecting satellites. The first satellite specifically designed for geodetic work, Pageos 1, was launched in 1966.

A first requirement for topographic mapping of a given area is an adequate pattern of horizontal and vertical control points, and an initial step is the assembly of all such existing information. This consists of descriptions of points for which positions (in terms of latitude and longitude) and elevations above mean sea level have been determined. They are occasionally located at some distance from the immediate project, in which case it is necessary to expand from the existing work. This is usually done on second- or third-order standards, depending upon the length of circuits involved.

The accuracy of survey measurements can be improved almost indefinitely but only at increased cost. Accordingly, control surveys are used; these consist of a comparatively few accurate measurements that cover the area of the project and from which short, less accurate measurements are made to the objects to be located. The simplest form of horizontal control is the traverse, which consists of a series of marked stations connected by measured courses and the measured angles between them. When such a series of distances and angles returns to its point of beginning or begins and ends at stations of superior (more accurate) control, it can be checked and the small errors of measurement adjusted for mathematical consistency. By assuming or measuring a direction of one of the courses and rectangular coordinates of one of the stations, the rectangular coordinates of all the stations can be computed.

A system of triangles usually affords superior horizontal control. All of the angles and at least one side (the base) of the triangulation system are measured. Though several arrangements can be used, one of the best is the quadrangle or a chain of quadrangles. Each quadrangle, with its four sides and two diagonals, provides eight angles that are measured. To be geometrically consistent, the angles must satisfy three so-called angle equations and one side equation. That is to say the three angles of each triangle, which add to 180°, must be of such sizes that computation through any set of adjacent triangles within the quadrangles will give the same values for any side. Ideally, the quadrangles should be parallelograms. If the system is connected with previously determined stations, the new system must fit the established measurements.

When the survey encompasses an area large enough for the Earth's curvature to be a factor, an imaginary mathematical representation of the Earth must be employed as a reference surface. A level surface at mean sea level is considered to represent the Earth's size and shape, and this is called the geoid. Because of gravity anomalies, the geoid is irregular; however, it is very nearly the surface generated by an ellipse rotating on its minor axis—*i.e.*, an ellipsoid slightly flattened at the ends, or oblate. Such a figure is called a spheroid. Several have been computed by various authorities; the one usually used as a reference surface by English-speaking nations is (Alexander Ross) Clarke's Spheroid of 1866. This oblate spheroid has a polar diameter about 27 miles (43 kilometres) less than its diameter at the Equator.

The imaginary Earth, or geoid

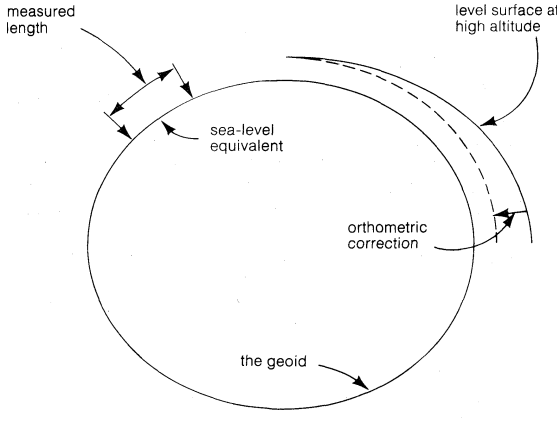


Figure 9: Measured length corrected to its sea-level equivalent (on the geoid) and the orthometric correction for high altitude measurements (see text).

Because the directions of gravity converge toward the geoid, a length of the Earth's surface measured above the geoid must be reduced to its sea-level equivalent—*i.e.*, to that of the geoid (see Figure 9). These lengths are assumed to be the distances, measured on the spheroid, between the extended lines of gravity down to the spheroid from the ends of the measured lengths on the actual surface of the Earth. The positions of the survey stations on the Earth's surface are given in spherical coordinates.

Bench marks, or marked points on the Earth's surface, connected by precise leveling constitute the vertical controls of surveying. The elevations of bench marks are given in terms of their heights above a selected level surface called a datum. In large-level surveys the usual datum is the geoid. The elevation taken as zero for the reference datum is the height of mean sea level determined by a series of observations at various points along the seashore taken continuously for a period of 19 years or more. Because mean sea level is not quite the same as the geoid, probably because of ocean currents, in adjusting the level grid for the United States and Canada all heights determined for mean sea level have been held at zero elevation.

Because the level surfaces, determined by leveling, are distorted slightly in the area toward the Earth's poles (because of the reduction in centrifugal force and the increase in the force of gravity at higher latitudes), the distances between the surfaces and the geoid do not exactly represent the surfaces' heights from the geoid. To correct these distortions, orthometric corrections (see Figure 9) must be applied to long lines of levels at high altitudes that have a north–south trend.

Trigonometric leveling often is necessary where accurate elevations are not available or when the elevations of inaccessible points must be determined. From two points of known position and elevation, the horizontal position of the unknown point is found by triangulation, and the vertical angles from the known points are measured. The differences in elevation from each of the known points to the unknown point can be computed trigonometrically.

The National Ocean Service in recent years has hoped to increase the density of horizontal control to the extent that no location in the United States will be farther than 50 miles (80 kilometres) from a primary point, and advances anticipated in analytic phototriangulation suggest that the envisioned density of control may soon suffice insofar as topographic mapping is concerned. Existing densities of control in Britain and much of western Europe are already adequate for mapping and cadastral surveys.

Global positioning. The techniques used to establish the positions of reference points within an area to be mapped are similar to those used in navigation. In surveying, however, greater accuracy is required, and this is attainable because the observer and the instrument are stationary on the ground instead of in a ship or aircraft that is not only moving but also subject to accelerations, which make it impossible to use a spirit level for accurate measurements of star elevations.

The technique of locating oneself by observations of celestial objects is rapidly going out of date. In practicing it, the surveyor uses a theodolite with a spirit level to measure accurately the elevations of the Sun at different times of the day or of several known stars in different directions. Each observation defines a line on the Earth's surface on which the observer must be located; several such lines give a fix, the accuracy of which is indicated by how closely these lines meet in a point. For longitude it is necessary also to record the Greenwich Mean Time of each observation. This has been obtained since 1884 by using an accurate chronometer that is checked at least once a day against time signals transmitted telegraphically over land lines and submarine cables or broadcast by radio.

A more recent procedure for global positioning relies on satellites, whose locations at any instant are known precisely because they are being continuously observed from a series of stations in all parts of the world. The coordinates of these stations were established by very large scale triangulation based on a combination of radar observations

of distances and measurements of the directions of special balloons or flashing satellites, obtained by photographing them at known instants of time against the background of the fixed stars.

The principal method of using satellites for accurate positioning is based on an application of the Doppler effect. A radio signal is transmitted at a steady frequency by the satellite, but a stationary observer detects a higher frequency as the satellite approaches and a lower one as it recedes. The speed of the frequency drop depends on the distance of the observer from the satellite's track, so a determination of this speed provides a measure of that distance. At the instant of the satellite's closest approach, the observed frequency is the same as that transmitted, so at that time the observer must be located somewhere along the line at right angles to the satellite's track. Since this track over the Earth's surface is accurately known at all times, these data define the observer's position.

Establishing the framework. Most surveying frameworks are erected by measuring the angles and the lengths of the sides of a chain of triangles connecting the points fixed by global positioning. The locations of ground features are then determined in relation to these triangles by less accurate and therefore cheaper methods. Establishing the framework ensures that detail surveys conducted at different times or by different surveyors fit together without overlaps or gaps.

For centuries the corners of these triangles have been located on hilltops, each visible from at least two others, at which the angles between the lines joining them are measured; this process is called triangulation. The lengths of one or two of these lines, called bases, are measured with great care; all the other lengths are derived by trigonometric calculations from them and the angles. Rapid checks on the accuracy are provided by measuring all three angles of each triangle, which must add up to 180 degrees.

In small flat areas, working at large scales, it may be easier to measure the lengths of all the sides, using a tape or a chain, rather than the angles between them; this procedure, called trilateration, was impractical over large or hilly areas until the invention of electromagnetic distance measurement (EDM) in the mid-20th century. This procedure has made it possible to measure distances as accurately and easily as angles, by electronically timing the passage of radiation over the distance to be measured; microwaves, which penetrate atmospheric haze, are used for long distances and light or infrared radiation for short ones. In the devices used for EDM, the radiation is either light (generated by a laser or an electric lamp) or an ultrahigh-frequency radio beam. The light beam requires a clear line of sight; the radio beam can penetrate fog, haze, heavy rain, dust, sandstorms, and some foliage. Both types have a transmitter-receiver at one survey station. At the remote station the light type contains a set of corner mirrors; the high-frequency type incorporates a retransmitter (requiring an operator) identical to the transmitter-receiver at the original station. A corner mirror has the shape of the inside of a corner of a cube; it returns light toward the source from whatever angle it is received, within reasonable limits. A retransmitter must be aimed at the transmitter-receiver.

EDM

In both types of instrument, the distance is determined by the length of time it takes the radio or light beam to travel to the target and back. The elapsed time is determined by the shift in phase of a modulating signal superimposed on the carrier beam. Electronic circuitry detects this phase shift and converts it to units of time; the use of more than one modulating frequency eliminates ambiguities that could arise if only a single frequency had been employed.

EDM has greatly simplified an alternative technique, called traversing, for establishing a framework. In traversing, the surveyor measures a succession of distances and the angles between them, usually along a traveled route or a stream. Before EDM was available, traversing was used only in flat or forested areas where triangulation was impossible. Measuring all the distances by tape or chain was tedious and slow, particularly if great accuracy was required, and no check was obtainable until the traverse closed, either on itself or between two points already fixed by triangulation or by astronomical observations.

In both triangulation and traversing, the slope of each measured line must be allowed for so that the map can be reduced to the horizontal and referred to sea level. A measuring tape may be stretched along the ground or suspended between two tripods; in precise work corrections must be applied for the sag, for tension, and for temperature if these differ from the values at which the tape was standardized. In work of the highest order, known as geodetic, the errors must be kept to one millimetre in a kilometre, that is, one part in 1,000,000.

Though for sketch maps the compass or graphic techniques are acceptable for measuring angles, only the theodolite can assure the accuracy required in the framework needed for precise mapping. The theodolite consists of a telescope pivoted around horizontal and vertical axes (Figure 10) so that it can measure both horizontal and

Theodolite

From *Ground and Air Surveys for Field Scientists* by John Wright, published by the Clarendon Press, Oxford; © John Wright 1982

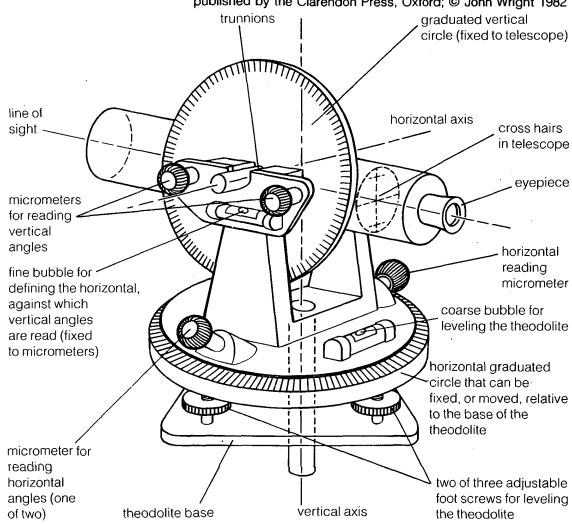


Figure 10: Principal parts of the theodolite. Clamping and slow-motion screws are not shown.

vertical angles. These angles are read from circles graduated in degrees and smaller intervals of 10 or 20 minutes. The exact position of the index mark (showing the direction of the line of sight) between two of these graduations is measured on both sides of the circle with the aid of a vernier or a micrometer. The accuracy in modern first-order or geodetic instruments, with five-inch glass circles, is approximately one second of arc, or $^1/_{3,600}$ of a degree. With such an instrument a sideways movement of the target of one centimetre can be detected at a distance of two kilometres. By repeating the measurement as many as 16 times and averaging the results, horizontal angles can be measured more closely; in geodetic surveying, measurements of all three angles of a triangle are expected to give a sum of 180 degrees within one second of arc.

In the most precise long-distance work, signaling lamps or heliographs reflecting the Sun are used as targets for the theodolite. For less demanding work and work over shorter distances, smaller theodolites with simpler reading systems can be used; targets are commonly striped poles or ranging rods held vertical by an assistant.

An extensive set of these measurements establishes a network of points both on the map, where their positions are plotted by their coordinates, and on the ground, where they are marked by pillars, concrete ground marks, bolts let into the pavement, or wooden pegs of varying degrees of cost and permanence, depending on the importance and accuracy of the framework and the maps to be based on it. Once this framework has been established, the surveyor proceeds to the detail mapping, starting from these ground marks and knowing that their accuracy ensures that the data obtained will fit precisely with similar details obtained elsewhere in the framework.

Detail surveying. The actual depiction of the features to be shown on the map can be performed either on the

ground or, since the invention of photography, aviation, and rocketry, by interpretation of aerial photographs and satellite images. On the ground the framework is dissected into even smaller areas as the surveyor moves from one point to another, fixing further points on the features from each position by combinations of angle and distance measurement and finally sketching the features between them freehand. In complicated terrain this operation can be slow and inaccurate, as can be seen by comparing maps made on the ground with those made subsequently from aerial photographs.

Ground survey still has to be used, however, for some purposes; for example, in areas where aerial photographs are hard to get; under the canopy of a forest, where the shape of the ground—not that of the treetops—is required; in very large scale work or close contouring; or if the features to be mapped are not easily identifiable on the aerial photographs, as is the case with property boundaries or zones of transition between different types of soil or vegetation. One of two fundamental differences between ground and air survey is that, as already mentioned, the ground survey interpolates, or sketches, between fixed points, while air survey, using semiautomatic instruments, can trace the features continuously, once the positions of the photographs are known. One effect of this is to show features in uniform detail rather than along short stretches between the points fixed in a ground survey.

The second difference is that in ground survey different techniques and accuracies may be adopted for the horizontal and vertical measurements, the latter usually being more precise. Accurate determinations of heights are required for engineering and planning maps, for example, for railway gradients or particularly for irrigation or drainage networks, since water in open channels does not run uphill.

The methods used for fixing locations within the horizontal detail framework are similar to, but less accurate than, those used for the primary framework. Angles may be measured with a hand-held prismatic compass or graphically with a plane table, or they may be estimated as right angles in the case of points that are offset by short distances from straight lines between points already fixed. Detail points may be located by their distances from two fixed points or by distance and bearing from only one.

The surveyor may record measurements made in the field and plot them there on a sketch board or in the office afterward, but if the country is open and hilly, or even mountainous, the plane table offers the best way of recording the data. A disadvantage of plane-table work is that it cannot be checked in the office, and so it requires greater intelligence and integrity of the surveyor. The plane table reached its most efficient form of use in the Survey of India, begun in 1800, in which large areas were mapped with it by dedicated Indian surveyors. It consists of a flat board that is mounted on a tripod so that it can be fixed or rotated around a vertical axis. It is set up over a framework point or one end of a measured baseline with its surface (which is covered with paper or other drawing medium) horizontal. It is turned until the line joining its location with another framework point or the other end of the baseline is parallel to the same line as drawn on the paper. This alignment is performed with the aid of an alidade, or sight rule, a straightedge fitted with simple sights. The alidade is then directed toward points on features that are to be fixed, and pencil rays are drawn along the sight rule toward them. The procedure is repeated at the other framework point or the other end of the baseline; the points where the rays intersect on the table will be the map positions of the features (see Figure 11).

Use of the plane table

In surveying for engineering projects, more sophisticated instruments are employed to maximize accuracy. For example, distances may be measured by EDM or by tachymetry, a geometric technique in which the vertical distance on a graduated vertical staff, seen between two stadia hairs in the theodolite eyepiece, is a measure of the horizontal distance between the theodolite and the staff—usually 100 times the difference between the two readings (see Figure 12). This method requires at least one assistant to move the staff from place to place. Modern surveying instruments combine a theodolite, EDM equipment, and a computer that records all the observations and calculates the height differences obtained by measuring vertical angles.

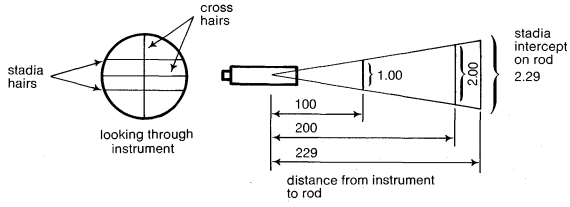


Figure 12: Principles involved in the use of telescope stadia hairs and a staff to determine distance. In the example illustrated, the distance to the rod is equal to 100 times the stadia intercept on the rod (see text).

From *Ground and Air Surveys for Field Scientists* by John Wright, published by the Clarendon Press, Oxford, © John Wright 1982

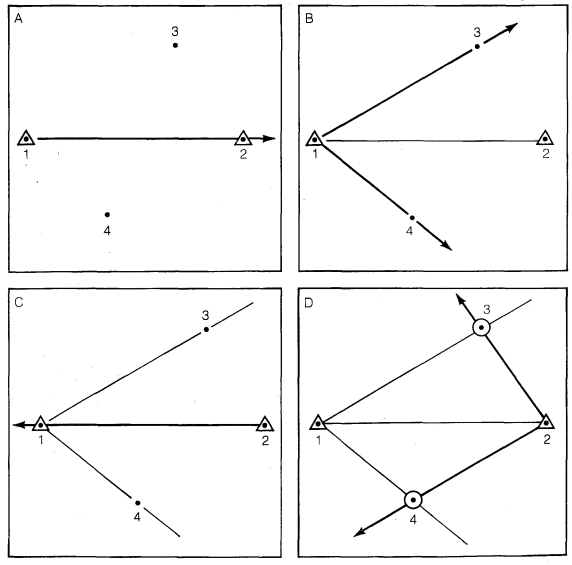


Figure 11: *Use of the plane table to determine positions of previously unmapped points (3 and 4) in relation to positions of points of known locations (1 and 2).*
(A) Table at point 1, rotated to make mapped line 1–2 coincide with actual direction of 2 from 1 and clamped; (B) table at 1, rays drawn in the directions of 3 and 4; (C) table at 2, rotated to make mapped line 2–1 coincide with actual direction of 1 from 2 and clamped; (D) rays drawn in the directions of 3 and 4. The intersections of the rays drawn in steps B and D establish the positions of points 3 and 4.

Aerial surveying. Aviation and photography have revolutionized detailed mapping of features visible from the air. An aerial photograph, however, is not a map, as can be seen in Figure 13, a view of the Houses of Parliament and Westminster Bridge, London, from directly overhead. On a map the tops of the towers at the corners of the building should coincide with the corners of the foundations, but in the picture they do not, being displaced radially from the centre. The photograph demonstrates an important property of vertical aerial photographs: angles are correctly represented at their centres, but only there. Similar distortions are present in photographs of hilly ground. This problem may be dealt with in two principal ways, depending on the relative scales of the map and the photographs and on whether contours are required on the map. The older method, adequate for planimetric maps at scales smaller than the photographs, was used extensively during and after World War II to map large areas of desert and thinly populated country; mountainous areas could be sketched in, but the relief was not accurately shown.

As in ground survey, a framework of identified points is necessary before detailed mapping can be carried out from the air. The photographs are ordinarily taken by a vertically aligned camera in a series of strips (see Figure 14) in which each picture overlaps about 60 percent of the preceding one; adjacent strips overlap only slightly. The overlaps make it possible to assemble a low-order



Figure 13: Aerial photograph of (top left) the Houses of Parliament and (right of centre) Westminster Bridge, London. Because they are viewed from a finite altitude, the tops of buildings appear farther from the centre of the picture than the foundations.
By courtesy of J.A. Story & Partners

framework or control system based on small, recognizable features that appear in more than one photograph. In the simplest form of this procedure each photograph is replaced by a transparent template on which rays are drawn (or slots are cut) from the centre of the picture to the selected features. The angles between these rays or slots are correct, and slotted templates can be fitted together by inserting studs, which represent the features, into the appropriate slots and sliding the templates so that each stud engages the slots in all the pictures showing the corresponding feature. This operation ensures that the centres of the pictures and the selected features are in the correct relationship. The array of overlapping photographs can be expanded or contracted by sliding them about on the work surface as long as the studs remain engaged in the slots, so the assemblage can be positioned, oriented, and scaled by fitting it to at least two—preferably several—ground-control points identified on different photographs.

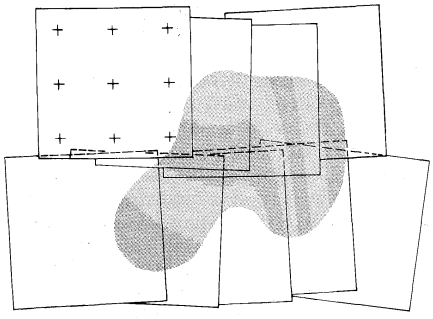


Figure 14: Photogrammetric photographs from two short, overlapping flight strips arranged for supplying mapping details. Photo-control points are shown on only one photograph; shading indicates a typical terrain feature such as a lake (see text).

This technique may be extended by using two additional cameras, one on each side, aimed at right angles to the line of flight and 30 degrees below the horizontal. The photographs taken by the side cameras overlap those taken by the vertical one and also include the horizon; the effect is to widen the strip of ground covered and thus to reduce the amount of flying required. Points in the backgrounds of the oblique photographs can be incorporated in the overlapping array as before to tie the adjacent flight paths together. Photography from high-flying jet aircraft and satellites has rendered this technique obsolete, but before those advances took place it greatly facilitated the mapping of underdeveloped areas.

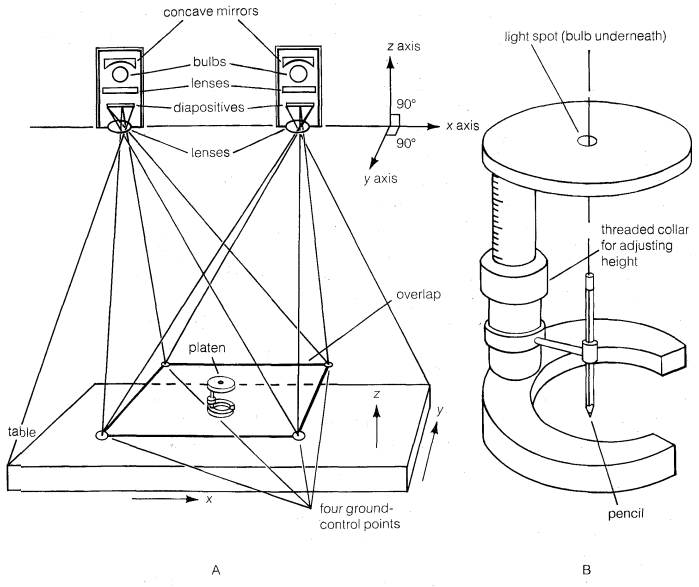


Figure 15: *Multiplex projector system of mapping contours and other details from aerial photographs* (see text).
(A) Arrangement of equipment, showing only two projectors.
(B) Details of the platen.

From *Ground and Air Surveys for Field Scientists* by John Wright, published by the Clarendon Press, Oxford; © John Wright 1982

For the production of maps with accurate contours at scales five or six times that of the photographs, a more sophisticated approach is necessary. The ground-survey effort must be expanded to provide the heights as well as the positions of all the features employed to establish the framework.

In this technique the details within each segment of the map are based not on individual photographs but on the overlap between two successive ones in the same strip, proceeding from the positions and heights of features in the corners of each area. A three-dimensional model can be created by viewing each pair of consecutive photographs in a stereoscope; by manipulation of a specially designed plotting instrument, the overlapping area can be correctly positioned, scaled, and oriented, and elevations of points within it can be derived from those of the four corner points. These photogrammetric plotting instruments can take several forms. In projection instruments (Figure 15) the photographs are projected onto a table in different colours so that, through spectacles with lenses of complementary colours, each eye sees only one image, and the operator visualizes a three-dimensional model of the ground. A table or platen, with a lighted spot in the middle, can be moved around the model and raised or lowered so that the spot appears to touch the ground while

From *Ground and Air Surveys for Field Scientists* by John Wright, published by the Clarendon Press, Oxford; © John Wright 1982

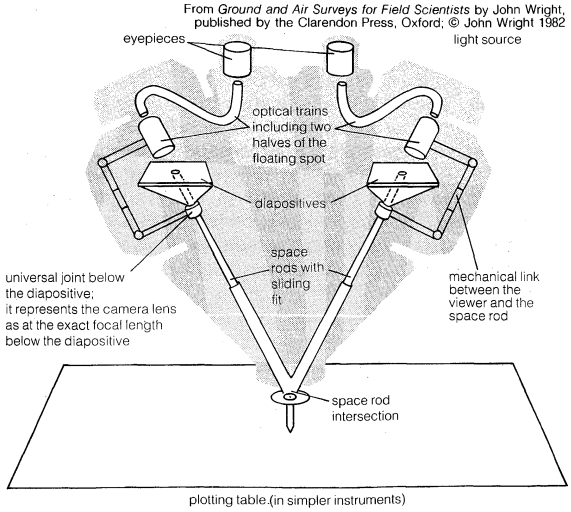


Figure 16: Optical-mechanical plotting instrument (see text).

the operator scans any feature, even if it is located on a steep hillside. A pencil directly beneath the spot then plots the exact shape and position of the feature on the map. For contouring the platen is fixed at the selected height (at a scale adjusted to that of the model), and the spot is permitted to touch the model surface wherever it will; the pencil then draws the contour.

With more complex mechanical devices, rays of light reaching the aircraft taking the two photographs are represented by rods meeting at a point that represents the position of the feature of the model being viewed (Figure 16). With a complicated system of prisms and lenses the operator, as with projection instruments, sees a spot that can be moved anywhere in the overlap and up or down to touch the model surface, as described above. A mechanical or electronic system moves a pencil into the corresponding position on a plotting table to which the map manuscript is fixed.

With computerized analytic instruments the mechanical operation is limited to measuring coordinates on the two photographs, and the conversion to a three-dimensional model is performed entirely by the computer (Figure 17). It is possible with the most precise plotting instruments of either type to draw a map at four to six times the scale of the photographs and to plot contours accurately at a vertical interval of about one one-thousandth of the height from which the photographs were taken. With such analytic instruments the record can be stored in digital as well as graphic form to be plotted later at any convenient scale.

From *Ground and Air Surveys for Field Scientists* by John Wright, published by the Clarendon Press, Oxford; © John Wright 1982

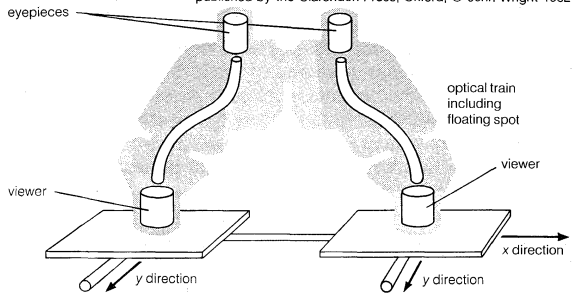


Figure 17: Geometry of analytical and approximate photogrammetric instruments. Both diapositives (or paper prints) are in the same plane and can move only in the *x* and *y* directions together and relative to each other.

All these methods produce a line or drawn map; some of them also create a data file on disk or tape, containing the coordinates of all the lines and other features on the map. On the other hand, aerial photographs can be combined and printed directly to form a photomap. For flat areas this operation requires simply cutting and pasting the photographs together into a mosaic. For greater accuracy the centres of the photographs may be aligned by the use of slotted templates as described above to produce a photomap called a controlled mosaic.

A much more precise technique is based on the use of an orthophotoscope. With this device, overlapping photographs are employed just as in the stereoscopic plotter already described, but the instrument, rather than the manual tracing of the features and contours, scans the overlap and produces an orthophotograph by dividing the area into small sections, each of which is correctly scaled. This procedure is best applied to areas of low relief without tall buildings; the resulting maps can then be substituted for line maps in rural areas where they are particularly useful in planning resettlement in agricultural projects. Because no fair drawing is required, the final printed map can be produced much more quickly and cheaply than would otherwise be possible.

Hydrography. Surveying of underwater features, or hydrographic surveying, formerly required techniques very different from ground surveying, for two reasons: the surveyor ordinarily was moving instead of stationary, and the surface being mapped could not be seen. The first problem, making it difficult to establish a framework except near land or in shoal areas, was dealt with by dead reckoning

between points established by astronomical fixes. In effect a traverse would be run with the ship's bearing measured by compass and distances obtained either by measuring speed and time or by a modern log that directly records distances. These have to be checked frequently, because however accurate the log or airspeed indicator and compass, the track of a ship or aircraft is not the same as its course. Crosscurrents or winds continually drive the craft off course, and those along the course affect the speed and the distance run over the ground beneath.

The only way a hydrographer could chart the seabed before underwater echo sounding and television became available was to cast overboard at intervals a sounding line with a lead weight at the end and measure the length of the line paid out when the weight hit the bottom. The line was marked in fathoms, that is, units of one one-thousandth of a nautical mile, or approximately six feet (1.8 metres).

Echo sounding

Sounding by lead line is obviously very slow, especially in deep waters, and the introduction of echo sounding in the early 20th century marked a great improvement. It was made possible by the invention of electronic devices for the measurement of short intervals of time. Echo sounding depends on timing the lapse between the transmission of a short loud noise or pulse and its return from the target—in this case the bottom of the sea or lake. Sound travels about 5,000 feet (1,500 metres) per second in water, so that an accuracy of a few milliseconds in measurements of the time intervals gives depths within a few feet.

The temperature and density of water affect the speed at which sound waves travel through it, and allowances have to be made for variations in these properties. The reflected signals are recorded several times a second on a moving strip of paper, showing to scale the depth beneath the ship's track. The echoes may also show other objects, such as schools of fish, or they may reveal the dual nature of the bottom, where a layer of soft mud may overlie rock. Originally only the depth directly beneath the ship was measured, leaving gaps between the ship's tracks. Later inventions, which include sideways-directed sonar and television cameras, have made it possible to fill these gaps. Measurements of depths away from the ship's track are not so accurate, but the pictures reveal any dangerous objects such as rock pinnacles or wrecks; the survey vessel can then be diverted to survey them in detail.

Modern position-fixing techniques using radar have made the whole process much simpler, for the ship's location is now known continuously with reference to fixed stations on shore or to satellite tracks. Another modern technique is the use of pictures taken from aircraft or satellites to indicate the presence and shape of shoal areas and to aid the planning of their detailed survey.

An alternative to the use of radar or satellite signals for continuous and automatic recording of a ship's position is the employment of inertial guidance systems. These devices, developed to satisfy military requirements, detect every acceleration involved in the motion of a craft from its known starting point and convert them and the elapsed time into a continuous record of the distance and direction traveled.

For studying the seabed in detail, the bottom of the sounding lead was hollowed to hold a charge of grease to pick up a sample from the sea floor. Today television cameras can be lowered to transmit pictures back to the survey ship, though their range is limited by the extent to which light can penetrate the water, which often is murky. Ordinary cameras also are used in pairs for making stereoscopic pictures of underwater structures such as drilling rigs or the wreckage of ancient ships.

Height determination. Heights of surface features above sea level are determined in four main ways: by spirit leveling, by measuring vertical angles and distances, by measuring differences in atmospheric pressure, and, since the late 20th century, by using three-dimensional satellite or inertial systems. Of these the first is the most accurate; the second is next in accuracy but faster; the third is least accurate but can be fastest if heights are to be measured at well-separated points. The last two techniques require sophisticated equipment that is still very expensive.

Spirit leveling

In spirit leveling the surveyor has for centuries used a surveying level, which consists of a horizontal telescope fitted with cross hairs, rotating around a vertical axis on a tripod, with a very sensitive spirit level fixed to it; the instrument is adjusted until the bubble is exactly centred. The reading on a graduated vertical staff is observed through the telescope. If such staffs are placed on successive ground points, and the telescope is truly level, the difference between the readings at the cross hairs will equal that between the heights of the points. By moving the level and the staffs alternately along a path or road and repeating this procedure, differences in height can be accurately measured over long horizontal distances.

In the most precise work, over a distance of 100 kilometres the error may be kept to less than a centimetre. To achieve this accuracy great care has to be taken. The instrument must have a high-magnification telescope and a very sensitive bubble, and the graduated scale on the staff must be made of a strip of Invar (an alloy with a very small coefficient of thermal expansion). Moreover, the staffs must be placed on pegs or special heavy steel plates, and the distance between them and the level must always be the same to cancel the effects of aerial refraction of the light.

In less precise work a single wooden staff can be used; for detailed leveling of a small area, the staff is moved from one point to another without moving the level so that heights can be measured within a radius of about 100 metres (Figure 18). The distances of these points from the

From *Ground and Air Surveys for Field Scientists* by John Wright, published by the Clarendon Press, Oxford; © John Wright 1982

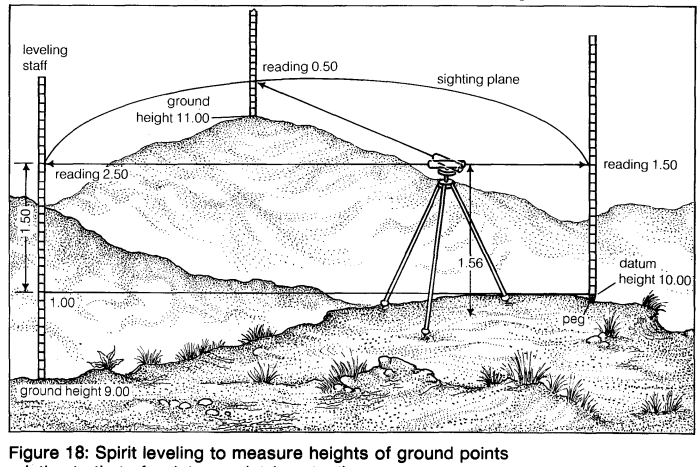


Figure 18: Spirit leveling to measure heights of ground points relative to that of a datum point (see text).

instrument can be measured by tape or, more commonly, by recording not only the reading at the central cross hair in the field of view of the telescope but also those at the stadia hairs, that is by tachymetry, as described above. The bearing of each point is observed by compass or on the horizontal circle of the level so that it can be plotted or drawn on the map.

Since the 1950s levels have been introduced in which the line of sight is automatically leveled by passage through a system of prisms in a pendulum, thus removing the need to check the bubble. The disadvantage of spirit leveling is the large number of times the instrument has to be moved and realigned, particularly on steep hills; it is used primarily along practically flat stretches of ground.

For faster work in hilly areas, where lower accuracies usually are acceptable, trigonometric height determination is employed using a theodolite to measure vertical angles and measuring or calculating the distances by triangulation. This procedure is particularly useful in obtaining heights throughout a major framework of triangulation or traverse where most of the points are on hilltops. To increase precision, the observations are made simultaneously in both directions so that aerial refraction is eliminated; this is done preferably around noon, when the air is well mixed.

The third method of height determination depends on measurements of atmospheric pressure differences with a sensitive aneroid barometer, which can respond to pres-

sure differences small enough to correspond to a foot or two (0.3 to 0.6 metre) in height. The air pressure changes constantly, however, and to obtain reliable results it is necessary to use at least two barometers; one at a reference point of known height is read at regular intervals while the surveyor proceeds throughout the area, recording locations, times, and barometer readings. Comparison of readings made at the same time then gives the height differences.

An alternative to the barometer for pressure measurement is an apparatus for measuring the boiling point of a liquid, because this temperature depends on the atmospheric pressure. Early explorers determined heights in this way, but the results were very rough; this technique was not accurate enough for surveyors until sensitive methods for temperature measurement were developed. The airborne profile recorder is a combination of this refined apparatus with a radar altimeter to measure the distance to the ground below an aircraft.

Analysis of the signals received simultaneously from several satellites gives heights as accurately as positions. Heights determined in this way are useful in previously unmapped areas as a check on results obtained by faster relative methods, but they are not accurate enough for mapping developed areas or for engineering projects. All-terrain vehicles or helicopters can carry inertial systems accurate enough to provide approximate heights suitable for aerial surveys of large areas within a framework of points established more accurately by spirit leveling.

(J.W.Wr./Ed.)

BIBLIOGRAPHY

History of mapping and surveying: A list of sources on the topic is presented in WALTER W. RISTOW, *Guide to the History of Cartography: An Annotated List of References on the History of Maps and Mapmaking* (1973). ROBERT C. DURU, *Maps and Map Making* (1977); and G.R. CRONE, *Maps and Their Makers: An Introduction to the History of Cartography,* 5th ed. (1978), are brief overviews. Essential aspects are explored in LEO BAGROW, *History of Cartography,* 2nd ed., rev. by R.A. SKELTON (1985, originally published in German, 1951); CHARLES BRICKER, *Landmarks of Mapmaking: An Illustrated Survey of Maps and Mapmakers* (1968; U.K. title, *A History of Cartography: 2500 Years of Maps and Mapmakers,* 1969; reissued 1977); LLOYD A. BROWN, *The Story of Maps* (1949, reprinted 1979), and *Map Making: The Art That Became a Science* (1960); EDWARD LYNAM, *The Mapmaker's Art: Essays on the History of Maps* (1953); and JOHN NOBLE WILFORD, *The Mapmakers* (1981). Maps of specific periods are studied in CHARLES H. HAPGOOD, *Maps of the Ancient Sea Kings: Evidence of Advanced Civilization in the Ice Age,* rev. ed. (1979); R.A. SKELTON, *Decorative Printed Maps of the 15th to 18th Centuries* (1952, reprinted 1966); HUGH CORTAZZI, *Isles of Gold: Antique Maps of Japan* (1983); and RAYMOND LISTER, *How to Identify Old Maps and Globes: With a List of Cartographers, Engravers, Publishers, and Printers Concerned with Printed Maps and Globes from c. 1500 to c. 1850* (1965). The development of cartographic institutions can be traced in MARY BLEWITT, *Surveys of the Seas: A Brief History of British Hydrography* (1957); SIR ARCHIBALD DAY, *The Admiralty Hydrographic Service, 1795–1919* (1967); ADRIAN H.W. ROBINSON, *Marine Cartography in Britain: A History of the Sea Chart to 1855* (1962); and G.S. RITCHIE, *The Admiralty Chart: British Naval Hydrography in the Nineteenth Century* (1967).

Mapmaking: The procedures involved in creating a map are described in many books, including WELLMAN CHAMBERLIN, *The Round Earth on Flat Paper: Map Projections Used by Cartographers* (1947); ARTHUR D. MERRIMAN, *An Introduction to Map Projections* (1947); ERWIN J. RAISZ, *General Cartography,* 2nd ed. (1948), and *Principles of Cartography* (1962); EDUARD IMHOF, *Cartographic Relief Presentation* (1982; originally published in German, 1965); J.S. KEATES, *Cartographic Design and Production* (1973); ARTHUR H. ROBINSON *et al., Elements of Cartography,* 5th ed. (1984); JOHN LOXTON, *Practical Map Production* (1980); and MORRIS M. THOMPSON, *Maps for America: Cartographic Products of the U.S. Geological Survey and Others,* 2nd ed. (1981). A range of special projects, of which cartography is an essential part, are described in A.H.A. HOGG, *Surveying for Archaeologists and Other Fieldworkers* (1980); TEODOR J. BLACHUT, ADAM CHRZANOWSKI, and JOUKO H. SAASTAMOINEN, *Urban Surveying and Mapping* (1979); P.F. DALE, *Cadastral Surveys Within the Commonwealth* (1976); S. ROWTON SIMPSON, *Land Law and Registration* (1976); R.A. SKELTON, *The Legal Elements of Boundaries and Adjacent Properties* (1930); CURTIS M. BROWN, WALTER G. ROBILLARD, and DONALD A. WILSON, *Evidence and Procedures for Boundary Location,* 2nd ed. (1981); T.W. BIRCH, *Maps: Topographical and Statistical,* 2nd ed. (1964, reprinted with corrections, 1976); B.W. LUCKE, *A Course on the Chart* (1966); M. CHRISS and G.R. HAYES, *An Introduction to Charts and Their Use,* 4th ed. (1977); and D.A. MOORE, *Marine Chartwork,* 2nd ed. (1981).

Technical developments: The many special studies in the field include CHESTER C. SLAMA (ed.), *Manual of Photogrammetry,* 4th ed. (1980); JOHN WRIGHT, *Ground and Air Survey for Field Scientists* (1982); C.D. BURNSIDE, *Electromagnetic Distance Measurement,* 2nd ed. (1982); and WILLIAM RITCHIE *et al., Mapping for Field Scientists: A Problem-Solving Approach* (1977). Current developments are covered in special journals: *Surveying and Mapping* (quarterly); *Photogrammetria* (bimonthly); *Military Engineer* (bimonthly); *The American Cartographer* (semiannual); *Cartography* (semiannual); and *Cartographica* (quarterly).

Modern surveying: UNITED STATES. GOVERNMENT PRINTING OFFICE, *Surveying and Mapping* (1982), is a subject bibliography listing current sources. Informative introductory texts include WILLIAM C. WATTLES, *Land Survey Descriptions,* rev. ed., edited by GURDON H. WATTLES (1974); CHARLES A. HERUBIN, *Principles of Surveying,* 3rd ed. (1982); J.G. OLLIVER and J. CLENDINNING, *Principles of Surveying: Plane Surveying,* 4th ed. (1978); PHILIP KISSAM, *Surveying Practice: The Fundamentals of Surveying,* 3rd ed. (1978); R.H. DUGDALE, *Surveying,* 3rd ed. (1980); and JACK C. MCCORMAC, *Surveying Fundamentals* (1983). Comprehensive treatments include CHARLES B. BREED and GEORGE L. HOSMER, *The Principles and Practices of Surveying,* 11th ed., rev. by W. FAIG, 2 vol. (1977); RAYMOND E. DAVIS *et al., Surveying, Theory and Practice,* 6th ed. (1981); RUSSELL C. BRINKER and PAUL R. WOLF, *Elementary Surveying,* 7th ed. (1984); FRANCIS H. MOFFITT and HARRY BOUCHARD, *Surveying,* 7th ed. (1982); and G. BOMFORD, *Geodesy,* 4th ed. (1980). Instruments used in the practice are discussed in J. CLENDINNING and J.G. OLLIVER, *Principles and Use of Surveying Instruments,* 3rd ed. (1969, reissued 1972); and M.A.R. COOPER, *Modern Theodolites and Levels,* 2nd ed. (1982). Hydrography is the subject of A.E. INGHAM (ed.), *Sea Surveying,* 2 vol. (1975); and MELVIN J. UMBACH, *Hydrographic Manual,* 4th ed. (1976). INTERNATIONAL HYDROGRAPHIC BUREAU, *Hydrographic Dictionary,* 3rd ed. (1970), contains useful information; and *International Hydrographic Bulletin* (monthly), and *International Hydrographic Review* (semiannual), published by the same organization, provide data on current research and developments.

(C.F.F./J.W.Wr.)

Marketing and Merchandising

Marketing in its most general definition is the directing of the flow of goods and services from producers to consumers or users. It is not confined to any particular type of economy, since goods must be marketed in all economies and societies except perhaps the most primitive. Nor is marketing a function only of profit-oriented business; even such service institutions as hospitals, schools, and museums engage in some forms of marketing. Within the broad scope of marketing, merchandising is concerned more specifically with promoting the sale of goods and services to consumers (*i.e.*, retailing) and hence is more characteristic of free-market economies.

This article treats the functions, varieties, and techniques of marketing. Advertising, perhaps the most conspicuous of the elements of marketing, is treated at length and in greater detail than some others.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 532 and 533.

The article is divided into the following sections:

The functions of marketing 538
  Types of goods and channels of distribution 538
    Fast-moving consumer goods
    Consumer durables
    Industrial goods
    Services
    Channels of distribution
  Marketing as a business function 539
    Product management
    Pricing
    Distribution
    Advertising and promotion
    Selling
The elements of marketing 539
  Market research 539
    Kinds of market research
    Basic procedures
  Product development 540
    Product design
    Product line and product mix
    Packaging and branding
  Advertising 540
    Establishment of modern advertising
    Development of the advertising agency
    Advertising media
    Advertising strategy
    Advertising design
  Distribution 547
    Retailing
    Wholesaling
    Marketing goods to industry
    Marketing farm products
Economic and social aspects of marketing 550
Bibliography 551

The functions of marketing

The increasing scope of marketing

In Western business organizations the marketing function was once seen as encompassing only sales and advertising and perhaps, also, market research. The purpose of the staff engaged in marketing was primarily to maximize sales volume. More recently the marketing manager's responsibilities have broadened to include all activities relating to the firm's customers, comprising product development, customer communications, and customer services; in some firms, the principal marketing executives have even been given authority over credit, physical distribution, public relations, and research and development. In a few firms, the integrated marketing concept has been pushed to the point where the head of marketing becomes the chief executive officer. The emphasis on marketing has been most conspicuous in industries serving the general consumer, such as retail store chains and manufacturers of foods, toiletries, and household appliances. Service industries, such as hotels, banks, and insurance companies, have been slower in taking it up. These tendencies have been visible in most of the leading industrial countries, although they have been pushed farthest in the United States.

The activities involved in marketing account for a large proportion of consumer expenditures. Studies made in the United States estimate the proportion at between 40 and 60 percent of total final sales value of all tangible goods. This includes the buying and selling expenses of producers, of wholesalers' and retailers' margins, and of all other promotional, storage, and transportation expenses. The proportion has increased through the years: in 1870 in the United States, 88 persons were employed in distribution for every 1,000 employed in production; in 1950, 407 persons were employed in distribution for every 1,000 in production. By 1980, according to slightly different criteria, about 675 persons were so employed. But if marketing increases in significance with the growth of an industrial economy, it also plays a large part in those less-developed economies where distribution is in the hands of myriad small tradesmen.

TYPES OF GOODS AND CHANNELS OF DISTRIBUTION

Goods and services can be classified in a variety of ways. One common scheme uses the following categories:

Fast-moving consumer goods. These are the items that make up the contents of the everyday shopping basket, such as groceries, toiletries, and general household consumables. Individual items are typically low- or moderately priced; they are purchased frequently, and households do not, as a rule, stockpile them. They are bought from supermarkets and other common retail outlets. Patterns of purchasing of these kinds of goods change in response to social developments, such as the increase in the proportion of women working outside the home. Goods such as these are often referred to as "low-involvement" purchases. This does not mean that they are unimportant but that purchasers are familiar with them and the variety of brands available and, given the small investment involved, can make purchase decisions quickly and without elaborate searches for information.

Consumer durables. These are items that, as their name implies, are used over substantial periods of time, such as home appliances and motor vehicles. They are bought less frequently and represent sizable investments; consequently consumers take more trouble over the purchase, seeking information from friends, salespeople, consumer guidebooks, and other sources before making a decision. Goods of this type are therefore also known as "high-involvement" purchases.

Industrial goods. These are goods purchased for use in the production of other goods and services. Some industrial goods, such as office equipment and stationery, are distributed through retail outlets, but most are distributed through separate channels. Industrial goods can be subdivided into raw or semifinished materials, which go into finished goods; consumables, such as stationery, fuel, and packaging materials; and capital goods, such as machinery, ships, and major projects like factories, petroleum refineries, and transport systems.

Services. All tangible goods contain an element of service and the reverse may also be true, but a separate cat-

egory of intangibles can readily be identified. In addition to everyday services such as theatres and other entertainments, restaurant meals, dry cleaning, and various forms of transportation are professional services, such as those of physicians, accountants, or bankers.

Channels of distribution. Goods move to the final consumer through various channels of distribution. Raw materials pass from the extractive industries (forestry, fishing, mining, and agriculture) to manufacturing establishments. Finished products may be sent directly by the manufacturer to the retailer or, less often, to the consumer, but as a rule they flow from the manufacturer to one or more wholesalers before they reach the retailer and, finally, the consumer. Although legal possession of goods generally accompanies their physical transfer from one party to another, this is not always the case; in consignment selling the producer retains ownership while the goods are in the hands of the wholesaler or retailer, that is, until the merchandise is sold to the final user or consumer.

In less-developed countries, channels of distribution tend to be shorter and simpler than in the more highly industrialized nations. There are notable exceptions, however. Cacao beans in Ghana are collected by the Ghana Cocoa Marketing Board, which licenses trading firms to process the commodity. Similar marketing organizations exist in several other West African nations. Because of the large numbers of small producers, these agents work through middlemen who, in turn, employ sub-buyers who may have "runners" to bring in the produce from remote areas. Japan's marketing organization was until recently characterized by lengthy channels of distribution and a multiplicity of wholesalers. It was not unusual for a product to pass through at least five separate wholesalers before it reached a retailer: a general wholesaler, a basic product specialty wholesaler, a special wholesaler, a regional wholesaler, and a local wholesaler. A move toward direct selling by manufacturers is compressing the length of these channels in Japan, as it is in Switzerland, Great Britain, West Germany, and most other highly industrialized nations.

Distribution in Africa and Japan

MARKETING AS A BUSINESS FUNCTION

Business people whose primary concern is the marketing of goods and services usually distinguish five functions basic to their work: (1) product or service management, (2) pricing, (3) distribution, (4) advertising and promotion, and (5) selling.

Product management. Product management involves the progress of an idea from conception through research and development, design, and engineering to the production of a marketable product and its launch into the marketplace. While many of these processes will be carried out by others, the marketing manager's job is to ensure that the end product is one that will appeal to and suit the needs of customers. The marketer is also responsible for the management of the existing product line, ensuring that it remains responsive to the demands of the marketplace and that obsolete or weak products are withdrawn.

Pricing. In determining prices, the marketer has to take account of costs, demand, and competition. Pricing for a new product presents quite different problems than does price management for existing products. And environmental factors—such as inflation, technological change, or government regulation—also must be considered.

Distribution. As the previous section indicated, companies have a wide choice of distribution channels open to them. They may sell direct to the final customer, as in the cases of some mail-order consumer goods or most industrial capital goods, or they may use one or more intermediaries. The choice is usually an economic one: Can the company perform the task itself more cheaply than through intermediaries? Frequently, however, there is also a question of access. In overseas markets it may be very difficult for an exporter to make contact with end users, and various kinds of agents must therefore be used.

Advertising and promotion. There is a saying that "if you make a better mousetrap, the world will beat a path to your door," but unfortunately the world often has other things to do. Customers have to be told, in the first place, that the product exists, and they have to be reminded frequently thereafter. In many cases much effort has to be expended to create and then sustain demand for a new and unfamiliar product. In developed countries some 2 percent of gross national product is spent on all forms of advertising and promotion.

Selling. Finally, there is, in most cases, a need for human contact between buyer and seller somewhere in the distribution chain. Consumer-goods manufacturers, even though they do not sell directly to the final consumer, maintain large sales forces to call on their immediate customers, whether wholesalers, retailers, or some other intermediaries. In the sale of complex capital goods the role of the salesperson is crucial.

The marketer's job has two main dimensions: diagnosis and action. He must first understand potential customers' needs, wants, and motivations, how they make up their minds, and how they can be made aware of the marketer's own product. This information can be gathered by observation of what succeeds for oneself or one's competitors, by trial and error (although this can be costly), or by carrying out market research. Having identified a target group, or segment, of customers, the marketer then puts together and carries out the package of actions outlined above. This package is known as the marketing mix.

The elements of marketing

MARKET RESEARCH

Market research has been defined as the application of the scientific method to the solution of marketing problems. It studies people as buyers, sellers, and consumers, examining their attitudes, preferences, habits, and purchasing power. Market research is also concerned with the channels of distribution, with promotion and pricing, and with the design of the products and services to be marketed. Although there is hardly a form of commercial activity that does not use market research with some frequency, it finds its widest application in manufacturing—probably because manufacturing is further removed from the ultimate consumer than are most public utilities and wholesale and retail enterprises.

Analysis of marketing problems

Kinds of market research. Advertising research attempts to evaluate the effectiveness of advertising expenditures and to guide future advertising efforts. Market analysis and forecasting try to identify and measure the markets for specific products and to estimate sales possibilities. Performance analysis helps a company evaluate how well it is meeting its objectives in such areas as sales, its share of the market, sales-force performance, costs, and profits. Research in physical distribution aims to increase a manufacturer's efficiency in moving products along the channels of distribution. Product research runs a gamut from identifying the needs and characteristics of product users to testing package designs.

Although the basic practices of market research originated in the United States, the methods have been adopted in most parts of the world in varying degree. Western European and Japanese market research expanded after 1950. The Soviet Union and other eastern European countries have established agencies to furnish market research to their nationalized industries.

Market research is not an exact science; it has not eliminated important areas of uncertainty in managerial decision making and probably never will. It indicates only the conditions that exist, not what to do about them. It may also be expensive, and its value must be measured against its cost. Finally, the information provided soon becomes out of date.

Basic procedures. There are certain general procedures in market research, although they are not applicable to all cases. A six-step scheme for conducting research on a marketing problem is the following: (1) preliminary analysis, in which potentially influential factors and variables are identified and examined, often based upon an extensive search for relevant data; (2) definition of the problem; (3) planning of the research study, which consists essentially of determining what data are sought, how the data will be obtained, and how they will be organized in order to permit sound conclusions and recommendations; (4)

collection and tabulation of data, which may be primary data gathered by observation, experimentation, or survey, or secondary data obtained from government or private sources; (5) interpretation of the results; and (6) presentation of the results to management, including assistance to management in following up the study's conclusions and recommendations.

Each of the six steps listed above involves serious problems of methodology and technique for the market researcher. In step four, for example, if the researcher wants to gather data by survey, he must usually decide among mail, telephone, and personal interviews. He is also faced with the questions of the survey's validity (whether it measures what it is intended to measure) and of its reliability as a sample of the entire market. The researcher must further decide how much stress to put on quantitative "facts" in his survey as against the qualitative issues of motivation—that is, why respondents answer in various ways.

PRODUCT DEVELOPMENT

In traditional economies, the goods produced and consumed often remain the same from generation to generation—including food, clothing, and housing. As economies develop, the range of products available tends to expand, and changes occur in the products themselves. In contemporary industrialized societies, products, like people, go through life cycles: birth, growth, maturity, and decline. This constant replacement of existing products with new or altered products is of great concern to professional marketers. The development of new products involves all the departments of a business—production, finance, research and development, even personnel administration and public relations.

Product life cycle

Product design. When a customer purchases a product he receives not only the item itself but, along with it, labeling, branding, packaging, sometimes a warranty, and perhaps the assurance of the availability of such services as repairs. In designing the product, these other things must be considered with it. Participants in the designing of a product may include consumers, middlemen, and industrial buyers, as well as the manufacturer. In a buyer's market—that is, one in which sellers must compete actively for the consumer's patronage—the greatest influence on product design is likely to be the consumer's reaction, actual or anticipated. Because of the closeness of middlemen (particularly retailers) to the consumer, they are often in an excellent position to anticipate consumer reaction to a planned product. Some giant retailers (*e.g.*, Marks and Spencer, Ltd. in Great Britain and Europe, or Sears, Roebuck and Company in the United States) play major roles in designing the products they sell. Industrial users may also take an active hand in designing the items they buy because the items often fit into an existing or planned production system. The degree of patent protection offered by the government is another factor affecting product design. Manufacturers of pharmaceuticals in Italy, for instance, lacking adequate patent protection, engage in very little original product research; they tend to copy innovations in other countries. France, on the other hand, grants an unusual degree of protection to designers of fashion products, such as women's clothing, despite the fact that manufacturers in most other countries are more or less free to copy those designs.

Three main types of product development

Product development takes basically three forms: innovation, alteration of an existing product, and the development of new applications for a product that remains the same. Examples of innovation include many antibiotic drugs, a wide variety of plastics, and innumerable electrical appliances, ranging from electric toothbrushes to compact disc players. Most invention today is done by the product research and development departments of manufacturing firms, rather than by independent inventors. Examples of alteration of an existing product range from the addition of stereo sound capability to television receivers to such minor changes as the addition of a butter drawer to a refrigerator. Other changes merely vary the style of a product—an office typewriter, a coat—to make it more attractive to the consumer. An example of a new application for an existing product may be seen in the electric blender, which was originally used to reduce soft fruits and vegetables into pulp or juices; eventually its kitchen uses were expanded so that it was promoted as a device for chopping and mixing a wide variety of foods.

Product line and product mix. A product (or service) may be thought of merely as giving the owner basic utilitarian satisfactions, as an automobile, for example, provides transportation. On the other hand, many goods go beyond merely practical requirements: the automobile may offer aesthetic satisfaction because of its design, or enhance its owners' status, or appeal to unconscious desires for virility, power, youthfulness, or dignity. The way a product is defined may also depend on the orientation of the manufacturer. Thus, a maker of electrical refrigerators may limit his market interests to the selling of his product as narrowly defined or, on the other hand, may think of his business more broadly as the arresting of food deterioration. In defining his product, however, the manufacturer must consider the marketing, technological, financial, and even legal problems that may arise when he tries to enlarge his scope.

Marketers are continually altering their "product lines" by adding new products, changing existing ones, and dropping others. A product line is an assortment of items that have some affinity to one another, whether from the way in which the customer makes use of them or the way in which he mentally relates the products to each other. Within each product there are additional classifications by style, colour, size, or other characteristics.

In setting up a product line, the marketer must consider the whole process of selling and distribution, including the prices at which the goods are to be sold, the approach to middlemen within the channels of distribution, the personal selling efforts required, and so forth. These elements must, in turn, be integrated with production, finance, research and development, and other nonmarketing functions within the firm.

Packaging and branding. Customers in France still carry their long loaves of bread unwrapped; small produce dealers in Italy wrap their vegetables in newspapers or place them in customers' string bags; and industrial wholesalers in many countries still sell nuts, bolts, nails, and other materials in bulk. But in most industrial countries the packaging of merchandise has become a part of the selling effort; marketers now specify exactly the types of packaging that will be most appealing to prospective customers. The importance of packaging in distributing the product has increased with the spread of self-service—in wholesaling as well as in retailing. Packaging is sometimes designed to facilitate the use of the product, as with aerosol containers for room deodorants; in Europe, such condiments as mustard, mayonnaise, and catsup are often packaged in tubes. Some packages are reusable, making them attractive to customers in some of the poorer countries where metal containers, for instance, are often highly prized.

The importance of externals

Branding is used to identify the product of a specific seller. Like packaging, a brand may become part of the product in the mind of the customer. Manufacturers' brands are found in most parts of the world, including the Soviet Union. Wholesalers' or retailers' brands (also called dealers' brands, private brands, or private labels) are employed by large department stores, large wholesale establishments, chain stores, and retail cooperatives.

Part of the marketer's product policy is his product guarantee or warranty, in which the manufacturer or middleman promises the purchaser certain services and performance standards. By explicitly stating his guarantees, the manufacturer may limit his liability, while also acquiring an additional promotion message. In most countries, large manufacturers and retailers are more likely to have specific policies on guarantees than are their smaller competitors, probably because they have more to gain or lose by the way in which their guarantees are interpreted.

(D.Ca./P.J.S.L.)

ADVERTISING

Advertising is a form of mass communication intended, in its usual meaning, to promote the sale of a product or service. The term is also used for mass communication



"War Gardens Victorious," a World War I poster intended to influence public opinion in favour of victory gardens.

intended to influence public opinion, to gain political support, to advance a particular cause, or to elicit some other response desired by the advertiser, although these practices may involve different techniques, measures, and legal requirements. The advertising message, or advertisement, is delivered to its intended audience through the various media, including newspapers, magazines, television, radio, billboards, and direct mail. Advertising is distinguished from other forms of communication in that the advertiser pays the medium to deliver the message. For this payment the advertiser receives the opportunity to control the message. Within legal constraints designed to prevent deception and assure fair competition, and within standards of practice enforced by the media, the advertiser is free to say what he wants to say the way he wants to say it. Within limits set by availability, he can select the particular issue of a newspaper or magazine in which the message will run; in the case of television and radio, he can select the hour and even the minute when the message will be broadcast. The style of presentation usually identifies the message as an advertisement. In some countries if the message is not readily distinguishable from a medium's editorial or program content, the notation "This is an advertisement" and the name of the advertiser or sponsoring organization are included.

Users of advertising

Different kinds of businesses use advertising to motivate different kinds of markets toward different kinds of responses. A retailer such as a department store advertises to build store traffic and patronage and to keep his regular customers informed about available merchandise. The manufacturer of a branded item such as soap advertises to build a brand preference—that is, to add value to his brand in the consumer's mind. The manufacturer of industrial goods, such as steel, machinery, lubricants, and office equipment, advertises to other business firms to get them to use his product in their operations. The producer of ethical drugs, building materials, or textbooks advertises to doctors, architects, or educators to induce these professional people to prescribe, specify, or recommend his product to patients, clients, or students. An association of firms in a particular industry, such as sugar refiners, orange growers, or dairies, advertise to stimulate consumption of a generic product such as sugar, orange juice, or milk. All business firms, whether they be retailers, manufacturers, or service institutions such as banks, insurance companies, and airlines, advertise to build a respected corporate identity, to make their corporate name well known and highly regarded.

Advertising is also used to awaken, enlighten, and activate the public at large concerning matters that affect society generally. Private, nonprofit organizations in some countries create and distribute various public service advertisements including those for forest fire prevention, rehabilitation of the handicapped, traffic safety, drug abuse information, crime prevention, and continuing education.

Generally speaking, advertising is more prevalent in highly industrialized countries where consumer purchasing power is broadly distributed, where the economic system is competitive, and where mass communications media are more generally available. Although consumer advertising exists in many Communist countries, its volume is relatively insignificant.

Establishment of modern advertising. The first advertising was by public criers in ancient times who circulated through the streets calling attention to the sale of such items as slaves, cattle, and imports. A written advertisement, perhaps 3,000 years old, was discovered by an archaeologist delving in the ruins of Thebes. It offered a "whole gold coin" as the reward for the return of a runaway slave named Shem. During the Middle Ages there was little of the activity that is now called advertising. The invention of movable type about 1450 by Johannes Gutenberg made it possible to produce many copies of books and periodicals cheaply and quickly and ushered in the modern era of advertising.

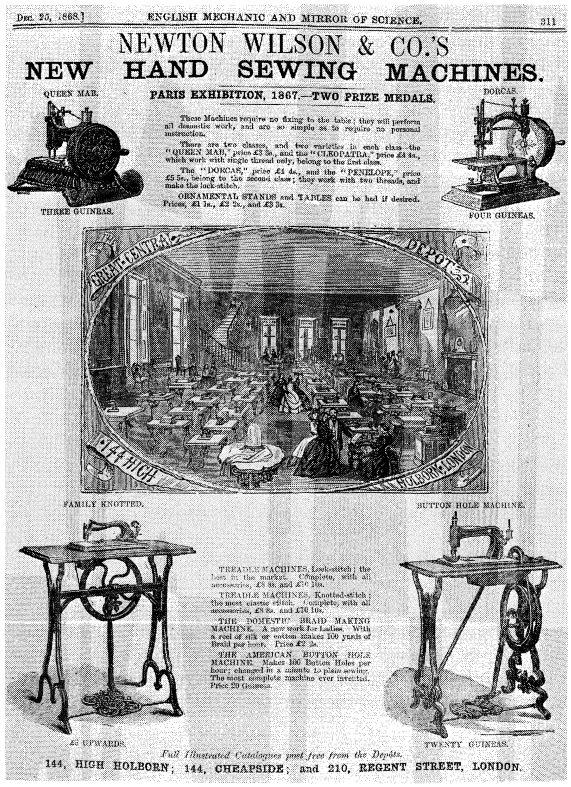
Printed advertisements

Printing made possible the transition from simple announcement to the system of argument and suggestion that constitutes modern advertising, and the medium of this development was the newspaper. The weekly papers sometimes carried a few advertisements, including the first offerings of coffee (1652), chocolate (1657), and tea (1658) in England. In June 1666 the *London Gazette,* No. 62, announced the first advertisement supplement as follows:

> An Advertisement—Being daily prest to the publication of Books, Medicines, and other things not properly the business of a Paper of Intelligence. This is to notifie once for all, that we will not charge the *Gazette* with Advertisements, unless they be matter of state; but that a Paper of Advertisements will be forthwith printed apart, and recommended to the Publick by another hand.

Advertising was prospering by the following century, and in 1758 Samuel Johnson could write in "The Idler":

Mary Evans Picture Library, London



An advertisement for sewing machines in a 19th-century English magazine.

Advertisements are now so numerous that they are very negligently perused, and it is therefore become necessary to gain attention by magnificence of promise and by eloquence sometimes sublime and sometimes pathetick.

The 19th century was a period of expansion in advertising as well as in business generally. The Industrial Revolution expanded the output of factories, and advertising helped market this output. The growth of the penny press and of magazines, combined with an increase in the educational level in the Western nations, brought forth an increase in audiences. Improvements in transportation made possible vast increases in circulation and reductions in price. Slogans and jingles became favourite advertising techniques. Some of the trademarks famous today were popularized during this period.

By courtesy of Levi Strauss & Co.



Label bearing the registered trademark of Levi Strauss & Co., which dates from the 1870s.

During the first two decades of the 20th century, advertising was reexamined. The common use of excessive, false, or deceptive claims caused widespread resentment. Regulation came both from within industries and from governments. After World War I came the era of salesmanship; advertising became "salesmanship in print." Advertising was accepted as an essential tool in selling the booming output of all the nations' factories. "It pays to advertise" became a standard slogan. It was discovered that testimonials by movie stars were effective in selling such items as perfumes, soaps, and cigarettes.

A major new advertising medium, radio, was added during the 1920s. At first, most radio commercials were indirect and there was doubt as to whether advertising would be allowed on the air. In such countries as Great Britain and France, it was banned from public-owned radio; in the United States, however, commercial radio thrived.

Another new medium, television, was added to the advertising scene immediately after World War II. It had a spectacular growth in the United States and became the second largest advertising medium (after newspapers) in terms of total dollars spent and the largest in the top national advertisers' budgets. In the postwar era, too, American advertising agencies began popularizing the concept of the "brand image." According to one agency president, "Every advertisement should be thought of as a contribution to the complex symbol which is the brand image. . . . Every advertisement is . . . a long-term investment in the total personality of the brand." The concept of a brand image was extended to building favourable images for corporations, retail stores, service institutions, and political candidates.

Development of the advertising agency. An advertising agency, Reynell and Son, was founded in London as early as 1812. In 1841 Volney Palmer, who had been an editorial writer and later a member of the advertising staff of the *Pottsville* (Pa.) *Miner's Journal,* opened in Philadelphia what is generally considered to have been the first advertising agency in the United States. Unlike modern agents, he was primarily an agent for newspaper publishers, who paid him a commission of 25 percent for selling space, although he did not assume liability for the credit extended to advertisers. Shortly after he had begun operations, he had several competitors.

Advertising agencies

George P. Rowell, who opened an agency in 1865, added two features—wholesaling and the "list system." He contracted with a large number of weekly newspapers to sell him a column of space for a certain period of time; he then offered this space to advertisers in smaller units at retail rates that were below rates charged by competitors. He was, like Palmer, primarily a seller of space, but he offered exclusive lists of newspapers and magazines, and any advertiser who bought space in these had to buy through the exclusive agent.

In 1875 N.W. Ayer & Son of Philadelphia developed marketing policies that gave advertisers access to the true rates charged by the publishers. Ayer thus put the emphasis on the agency's service to the advertiser rather than on selling space for the media.

The next major step was the development of a full line of services for the advertiser. In the late 19th and early 20th century the emphasis was on the creation of advertising copy. After copy came art services, advertising production, and the selection of media. By 1920 most agencies could plan complete campaigns for their clients, including research and budgeting, and could also execute these campaigns by preparing the advertisements and arranging for their insertion in the various media, which now included outdoor advertising and direct mail.

The typical advertising agency today is staffed to perform the following services for its clients: (1) planning marketing and advertising strategies; (2) writing copy; (3) preparing artwork, layouts, and television storyboards; (4) buying artwork; (5) arranging for mechanical reproduction of printed advertisements; (6) producing television commercials; (7) budgeting expenditures; (8) selecting media; (9) contracting for space and time; (10) checking, billing, and paying media; (11) researching markets, media, copy, and results; (12) merchandising promotional programs to the trade; (13) counseling and engaging in public relations activities. Agencies performing all of these functions are referred to as full-service agencies and are likely to be the larger firms. Many smaller agencies concentrate on planning, preparing, and placing advertising.

Advertising media. Advertising media can be classified as follows: newspapers, magazines (consumer magazines, business publications, farm publications, professional journals), television, radio, direct mail, outdoor media (signs, posters, painted bulletins, electric displays), transit media (car cards, outside displays, station posters), and miscellaneous (dealer displays, theatre-screen advertising, specialties, directories, point of sale).

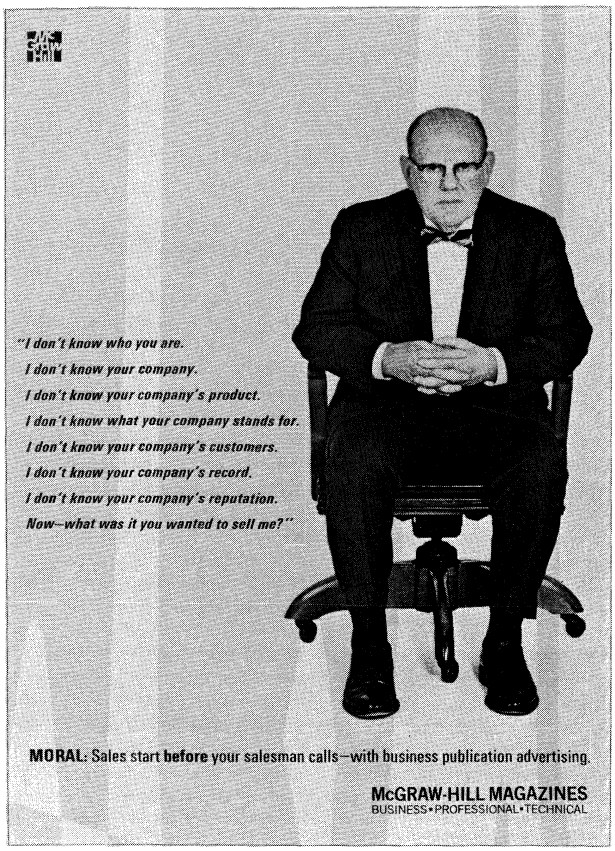
There is no single best medium for all advertising situations. Each has its own character, and each advertising situation presents a unique set of circumstances. Each medium must therefore be considered in terms of how well it meets the individual advertiser's requirements.

Newspapers. Newspapers, because of their large and general circulations, are the most widely used advertising medium. Through advertising, newspapers keep their readers informed of the myriad goods and services available in the marketplace, including the apartment for rent, the house for sale, and the employment opportunity appearing in the classified section.

The three main kinds of newspapers suit different advertisers. National newspapers are ideal for national brands and, given that different newspapers appeal to different readerships, the advertiser can be broadly selective about the income and social characteristics of his readers. International newspapers like the *International Herald Tribune* and the *Financial Times* reach across national boundaries and reach cosmopolitan consumers of internationally distributed products. Local newspapers, with their limited geographic distribution, are better suited to local distributors and retailers, whose marketing area they may match quite closely. There are a growing number of "free sheets"—newspapers that carry comparatively little news and are paid for entirely by advertising. These can be distributed on a tightly defined geographic or demographic basis and can thus be used by the advertiser to reach very specific target audiences.

Local newspapers

Magazines. Magazines offer a wide range of advertising opportunities. General magazines such as *Reader's Digest*



A magazine advertisement for McGraw-Hill magazines dating from 1958 to the present.

© McGraw-Hill Publications Company; reprinted with permission

(U.S.), *Paris-Match* (France), *Der Spiegel* (West Germany), and *Asahi* (Japan) are edited to serve the interests of many different population segments, thereby offering advertisers the opportunity to reach a broad cross section of people in all parts of a nation. Women's magazines such as *Elle* (France), *Ms.* (U.S.), and *Vogue* (both U.S. and France) appeal to various special interests of women, thereby offering advertisers the opportunity to reach these important consumers in a variety of editorial environments. There are magazines that concentrate on sports, business, science, or virtually any other subject, serving virtually every interest group imaginable. Each type of magazine offers the advertiser an opportunity to reach his best prospects when they are reading subject matter relevant to the advertiser's product. A manufacturer of milking machines, for example, can reach dairymen in a receptive frame of mind when he advertises in a dairymen's journal. The many professional journals read by doctors, dentists, engineers, architects, and educators appeal to advertisers who wish to influence those who are in a position to prescribe, recommend, or specify what others will buy.

The geographic selectivity of consumer magazines has been increased through regional editions that enable advertisers to place an advertisement in a part of the circulation that goes into a particular market area. The advertiser must buy a certain minimum circulation, but this is likely to be sufficiently low to permit him to advertise in only one of the major cities. Regional editions let local and regional advertisers benefit from the prestige of a national publication.

The ideal measure of a newspaper's or magazine's value to a given advertiser would be the number of his prospective customers who read it. It would be extremely difficult, however, to isolate and count prospects for each and every advertiser. Therefore, research organizations that measure advertising for advertisers report the composition of magazine audiences in rather standard demographic terms such as age, sex, occupation, income, and family size. Using these breakdowns the advertiser selects newspapers and magazines whose audiences best match the demographic characteristics of his market.

Television. In the affluent nations today, more people spend more time with television than any other medium. In the United States, for example, during the peak viewing hours of an average winter evening, more than 50 percent of the households are tuned in. Viewing per household averages six hours daily. Viewing declines about 20 percent during the summer months. Changes in the hour-by-hour activities of various members of the family influence the composition and size of the television audience. Men do most of their viewing at night and on weekends. Preschool-age children watch more daytime television than school-age children. Women usually do more viewing than men in all time categories. The older age groups, 50 and over, do more viewing than younger groups. The audience composition also varies by program types, whether drama, comedy, feature film, documentary, or other.

Units of advertising time

Television advertisers buy units of time: a minute, 30 seconds, 20 seconds, 10 seconds, a quarter hour, a half hour, an hour or longer. They may buy time from a single station or from a network of stations. The cost of a unit varies by time of day and from program to program, depending on the size of the audience. In such countries as West Germany all commercials are placed between programs; there are no program interruptions. In the United States they are placed both between programs and within programs. The advertiser may specify the time his commercial is to be broadcast, or for a lower rate he may leave its placement to the station's discretion. Program time units may be bought for exclusive sponsorship or multiple sponsorship.

There are many methods for measuring television audiences. The "Audimeter," for instance, developed in the United States by the A.C. Nielsen Company, is an electronic device that is connected to all television receiving sets in a representative sample of households. The Audimeter makes a minute-by-minute record on photographic film of the time periods when each set is turned on, the specific stations to which the set is tuned, the dialing from one station to another, and the length of time spent with each station. The film cartridges are sent in every two weeks to the Nielsen organization where the data are tabulated and matched against the program schedules of the television stations involved. Although the Audimeter provides a precise record of set tuning, it does not show how many (or if any) people were viewing nor which members of the family were doing so. The "coincidental telephone" method consists of telephoning a sample of households and asking if a television receiver is turned on and, if so, to what station or program. This method is relatively inexpensive and therefore is used extensively in local market areas. The "roster-recall" method involves personal interviews in which a roster of all stations and programs readily available in the area is presented to a sample of people. The respondents are asked to check those programs or stations they recall having watched. The "diary" method requires that individual members of the family use a prepared diary form and keep a record of their viewing as it occurs. This method makes it possible to estimate the composition as well as the size of the audience.

Commonly used measurements of television audiences are the following: "program rating," which reveals the percentage of all households in the sample tuned to the particular program; "share-of-audience," which gives the percentage of households using television tuned to each program; "projected-audience," which gives the estimated number of households reached by a program (a projection of the program rating to total population); and "audience-composition," which offers a breakdown by sex, age, family size, income, and so forth.

Radio. Radio is ubiquitous. Battery-operated radios are found in the remotest regions of Africa and Asia, and in the developed countries it is scarcely possible to escape the blare of portable radios. Many governments have considered radio an important medium for education and public information and severely restricted its use for commercial advertising. Some, however, such as Great Britain, have



Car card advertising Wrigley's chewing gum by Otis Shepard, 1949.
By courtesy of the Wm. Wrigley Jr. Company

more recently permitted the establishment of commercial stations. Radio advertising time is sold in much the same way as is television time, and the same methods of audience measurement are used.

Direct mail. By mail the advertiser can send his message directly to individual prospects. The message can take one of many forms: letter, folder, booklet, brochure, postcard, catalog, or coupons. In letter form the message can be personalized by addressing the recipient by name, by appealing to his personal concerns, and by using the language and style of a personal letter. The advertiser also can single out special groups for special appeals. He can be fairly certain that the message will be at least partly read by most recipients and that it will be read on the date it is received.

The mailing list

The mailing list is a critical factor in direct-mail advertising. Frequently used sources of names and addresses are telephone and city directories, trade directories, professional directories, commercial rating books, tax lists, club membership lists, and the records of local governmental bodies. Some firms make a business of compiling lists and selling them to advertisers.

Other media. Organized outdoor media include painted signs, electric displays, and printed posters, the latter being the most important. The standard poster or billboard is located mainly on heavily traveled routes in and around urban markets. The fleeting exposure to motorists and passersby demands a brief message, one that can be seen or read in a glance. A single, simple, visual idea with a caption usually is most appropriate. This brevity of message, however, limits the use of posters to reminder advertising.

Transit advertising, including car cards, outside displays, and station posters, reaches the millions of riders of public transportation in urban areas. Like outdoor advertising, transit advertising offers frequent exposure to large numbers of people at low cost. Its limitations are brevity of message and an environment low in prestige.

Dealer displays reach consumers at the point of purchase and therefore serve as a reminder or as an invitation to immediate buying action. Display space, however, in crowded, self-service supermarkets is at a minimum.

Theatre-screen advertising offers the opportunity to reach a captive audience with the impact of sound and motion pictures on a big screen.

Advertising "specialties" include calendars and other useful items that are imprinted with the advertiser's name. They include ashtrays, lighters, thermometers, key chains, ballpoint pens, memo pads, paperweights, and rulers. Being items that are used frequently, they serve as frequent reminders and engender some goodwill.

Advertising strategy. Advertising strategy is based on a series of key questions: (1) To whom should the advertising be addressed? (2) What communication objectives should be sought? (3) What should the message contain? (4) What media should be employed? (5) How should expenditures be budgeted? Advertising often is characterized as a means of selling to the masses. This leads to the misconception that all advertising should be addressed to everyone. Any population, however, is composed of different individuals with different needs and wants, different life-styles, and different tastes. The most efficient strategy is thus to single out those segments of the population who are the best prospects and to address the advertising to them. Communication objectives include the central idea that is to be learned, the image that is to be perceived, and the action that is to be taken by the target audience. Strategic decisions on message content concern visual, verbal, and aural devices, tone or style, and technique of presentation. Media selection includes the classes of media to be used—for example, daytime television, billboards, or women's magazines. Budgeting involves deciding how much to spend and how expenditures are to be allocated to markets, to media, and to dates. Thus, advertising is essentially a decision-making process. To aid the decision maker, research is used in a variety of ways.

Research

Researching the consumer gives the advertiser a better understanding of the people whom he wants to influence. Motivations that underlie their buying behaviour, their attitudes toward particular product categories and brands, their perceptions of competitive brands, their satisfactions and dissatisfactions with the products they use, their reactions to new product concepts—these are some of the kinds of information that help the advertiser to identify his best prospects and to select the strongest appeal, thereby increasing the effectiveness of his advertising.

Motivational research utilizes depth interviewing and projective techniques such as word associations, sentence completions, and various other tests drawn from clinical psychiatry. Use of these techniques to educe subconscious motives led to the fear in the 1950s that advertisers were gaining the power to manipulate consumer behaviour. The fear was not well founded, for it overlooked the power of the individual to resist persuasion against his own perceived interest. To the degree that techniques of persuasion have become more sophisticated and effective, people have developed sophisticated awareness of and resistance to manipulation.

Traditionally, market segments have been described in demographic terms such as sex, age, income, and occupation. Recently there have been more attempts to use psychological measurements such as personality traits and motivational variables. Geographically, the market for most products is concentrated in urban areas, where most of industrial populations now live. Regional differences in climate, ethnic groups, and life-style produce different sales potentials for different products. Because markets are dynamic, however, continuing research is necessary to keep up with the changes.

A company's advertisement is best viewed as one of a series building a consistent identity for the product or the company. Creative strategy, therefore, should produce a central idea or unifying theme for a continuing campaign. The theme may be expressed verbally in the form of a brief proposition, a headline, or slogan. It may also be expressed visually in pictorial content, layout style, graphic treatment, and such details as trademarks and logos.

Media strategy

Creative requirements have important implications in media selection. If superior full-colour reproduction in print is required, magazines have an advantage over newspapers. New product introductions benefit from the "news" format of newspapers, radio, and television. If the copy approach seeks to inspire confidence or dispel doubt, an authoritative medium such as a professional journal is appropriate. The environment in which the message



Magazine advertisement in *Ladies' Home Journal*, July 1958, by Norman Rockwell, from a three-year Crest toothpaste campaign with a unifying theme; Benton and Bowles, 1957–59.

By courtesy of The Procter & Gamble Company

appears has an appreciable influence on the effectiveness of the message itself.

The basic variables of media strategy are "reach" and "frequency." Reach refers to the number of households or individuals exposed to the advertising message in a given period of time. Frequency is the average number of times that they are exposed. These variables are interdependent. With a given budget it is possible to reach more people less often or fewer people more often. The optimum combination depends on the requirements of the particular advertising situation. When introducing a new product whose market is unknown, for example, one may try to reach as many people as the budget permits. When the market is well defined, however, one may try to "expose" the message as frequently as possible.

The contribution of advertising to sales and profits cannot be readily isolated from all the other variables in the marketing mix. Determining the size of the advertising budget is, therefore, a rather arbitrary procedure. Assuming that a relation exists between advertising and sales (even though it cannot be precisely measured), a common practice is to express the advertising budget as a percentage of sales. The percentage is high for some products—about 20 percent for cosmetics—and low for others—about 1 percent for insurance. The size of the advertising budget tends to be greater when most of the selling burden is placed on advertising instead of personal selling, when a new product is introduced, when there is little or no discernible difference among competing brands, or when competitive practice requires heavy advertising.

(V.R.F./P.J.S.L.)

Advertising design. Advertising requires that its ideas be given a visual form by artistic and graphic design. The means used to this end are typography, drawing, photography, and colour. The secret of good and effective design is to conceive the composition as a whole in which form and content, and representation and expression, are interdependent and matched to each other.

General considerations. The function of advertising design is primarily a practical one directed toward a particular end. It must communicate quickly and clearly and, in so doing, (1) attract attention; (2) lodge the advertising message in the conscious or unconscious mind; (3) induce an active, positive attitude to the message instead of a passive one; and (4) impress a favourable memory of the message on the mind. Because advertising design is aimed at producing an effect desired by an advertiser, it is obviously not free to choose its theme as the fine arts are. Even so, both advertising design and art use the same resources for the realization of their ideas: form, colour, and composition. Advertising design may use these effects successfully without necessarily producing art; it has served its purpose if image and text create the desired effect. A free work of art, on the other hand, becomes true art only when all the components of the design have genuine qualities of form and colour and together make up a vital composition which is complete in itself. Advertising design, however, can also rise to art if the designer has a strong artistic personality and an original imagination and if the commission allows the artist a certain freedom in the way he presents his theme and selects his artistic means—for example, the famous posters of Toulouse-Lautrec.

Both the nature of advertising design and the tasks that it has to perform have been profoundly influenced by advancing technology. New sophisticated techniques of reproducing and representing acoustic and visual material have given advertising new scope. To relief, intaglio, lithography, and silk-screen printing now have been added computer graphics. Hand composition, monotype, and linotype have been supplemented by computerized phototypesetting techniques. Cinema slides and advertising films have been joined by television commercials and holograms. Stationary, two- or three-dimensional signs have been supplemented by moving displays that enable objects to be seen in all their perspectives. Many such displays are programmed by computers.

Techniques used in design

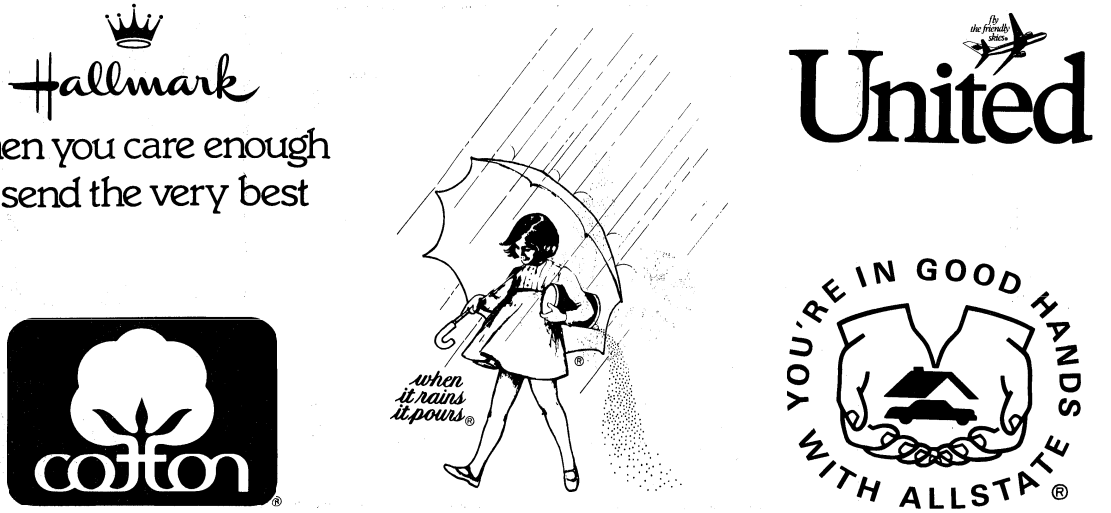
The technical and artistic demands on the advertising designer have become more exacting. The variety of tasks that he has to accomplish calls for detailed knowledge of printing techniques and composition, of cinema and television techniques, of the properties and working qualities of a great diversity of materials used in exhibition design, of the technique of audiovisual presentation and multiprojection, of the potentialities of electronic sign advertising, and of the aids that can be obtained from computers. The advertising designer must familiarize himself with these new techniques and their effects, thereby acquiring more scope for his artistic imagination, a greater field of application for his ideas, and a more comprehensive impact for his advertisements.

The success of advertising design depends on the topicality of what is depicted, on the quality of the design, and, to no less extent, on psychological factors. These largely determine the reactions of the person to whom the advertisement is addressed. Consciously perceived information is capable of inducing moods and feelings that are in harmony with the intention of the advertisement or at variance with it. The static use of the designer's resources in a composition, for instance, may express rest, solidity, conservatism, security, and trust, whereas a dynamic use may evoke love of experiment, flexibility, a forward-looking attitude, aggressiveness, and youthfulness. The correct choice of colours is especially important. Colours with a symbolic significance in politics or religion are taboo and must be used with caution, depending on the country and religion involved.

With the increasing sophistication of advertising after World War II, the importance of the free-lance advertising designer naturally diminished. Advertising campaigns, some of which have become very extensive and complex, can be planned and carried out only by specialists working as a team. A full-service advertising agency handles all the problems arising in connection with an advertising assignment, including budgeting, and thus has a far greater capacity than the designer working on his own. The free-lance advertising designer, generally speaking, can only execute commissions for smaller firms or take on jobs assigned to him by the agencies in connection with advertising campaigns.

The designer as an employee of an agency

The upshot of this development is that the advertising designer has virtually become a staff employee and team colleague of the agency. Depending on his talents, he is employed in a variety of capacities or as a specialist in one particular line—for example, as an illustrator, letterer, draftsman, filmmaker, or display designer. In most cases



U.S. registered trademarks, 1914 to the 1970s.
The Morton Salt girl and slogan (centre) first appeared together on the Morton Salt package in 1914. To keep her fashionable, the girl was given new dresses and hairstyles in 1921, 1933, 1941, 1956, and 1968. The other trademarks date from the 1940s, Hallmark greeting cards; the mid-1960s, United Airlines; 1971, American Cotton Producers; and the 1960s, Allstate Insurance Company.

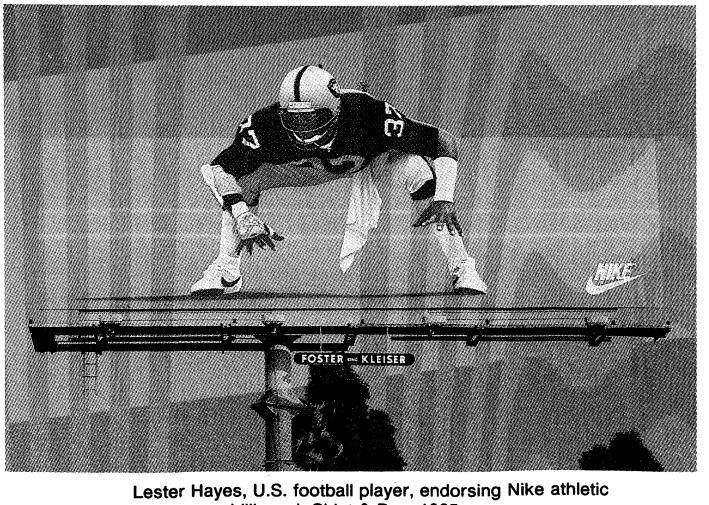
By courtesy of (top left) Hallmark Cards; (centre) Morton Thiokol, Inc. ("The Little Morton Salt Girl" and the slogan "When It Rains It Pours" are registered trademarks of Morton Thiokol, Inc.); (top right) United Airlines; (bottom left) Cotton, Inc., the Fiber Company of American Cotton Producers; (bottom right) Allstate Insurance Company

his work remains outwardly anonymous and is integrated with the overall output of the agency.

Design media. These include advertisements in periodicals, posters, brochures, leaflets, pamphlets, buttons, and the like. Frequently they contain a "device," "trademark," or "trade name" used to denote the maker or supplier of the product or service. The device or trademark represents the most concise graphic form that a particular subject can be given. It should be simple, unmistakable, applicable, and timeless. The trade word has the advantage over the device in that it can be read at once, whereas the meaning of the device must first be made known. Its disadvantages, of course, are that it contains more formal elements (depending on the length of the name) and that the designer's freedom to shape it at will is limited by the demands of legibility; it thus usually has a weaker impact.

Each medium has its own demands. Whereas a poster designed to produce an effect at a distance, for instance, must be drawn and coloured on a generous scale, a newspaper advertisement has to be worked out with greater attention to details. In any case, the idea is to achieve the maximum of clarity and impact with the minimum resources of colour and form. The illustrative poster, for example, presents the theme or object with an atmospheric appeal; the realistic poster supplies objective information through its copy and the truthful depiction of the object; the constructivist poster arranges its pictorial elements in accordance with their thematic importance. The experimental poster surprises the viewer by the novel way in which it presents a problem or by the new techniques it employs.

Photo, courtesy of NIKE, Inc.



Lester Hayes, U.S. football player, endorsing Nike athletic shoes on a billboard; Chiat & Day, 1985.

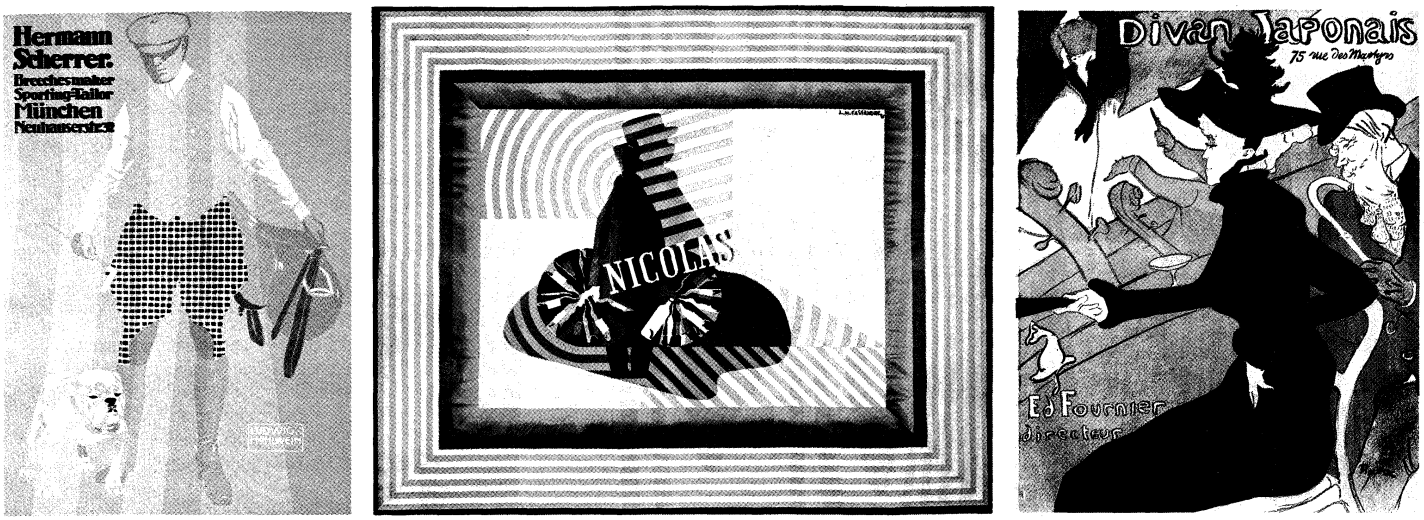
Television commercials and other kinds of films afford the advertising designer a variety of interesting opportunities. If the film or tape consists of live action, then titles, subtitles, possibly functional diagrams, statistics, movie sets, and so on have to be designed. If the film consists of an animated cartoon, then the designer's imagination is given almost unlimited scope; every type of drawing can be used, from fairy-tale illustration to surrealistic montage, and every kind of style can be employed—naturalism, simplified and stylized presentation, abstract symbolism, and so on. But, in any event, it is important that the designer depict his pictorial ideas clearly and simply, that they not change too quickly, and that he not present more than one or two ideas in the picture at the same time. He must consider colours and their sequences with the same care. In all types of films, lettering, image, word, and sound must complement one another.

In parades and demonstrations there are posters, transparencies, banners, and boards, but, unlike most ordinary posters, they are continually in motion and must be visible from various angles. Lettering and images must be recognizable and legible at a distance.

Creating the "image" of the firm

One of the most interesting jobs in advertising design is the creation of a uniform "image" for a business or organization. All advertising aids from visiting cards through commercial forms and newspaper advertisements to outdoor signboards should be given an unmistakable character with regard to graphic form, typography, and colour, so that each element is seen as part of an overall conception. All these different impressions must have a family resemblance, so that their effect is cumulative and produces a clear image in the eye of the beholder. Uniform advertising is a powerful aid in the creation of a firm's image.

In comparing lettered signs on doors, windows, buildings, factories, and so on, the advertising designer is confronted by three specific problems: the lettering must be visible at a certain distance; it must make an impact by reason of its formal originality; and it must be integrated with the architecture in size, form, colour, and material. If the lettering is to be visible at night, the designer can make his choice between neon lettering and spotlighted lettering. If he chooses neon lettering, he must adapt the shape of his letters to technical or engineering requirements.



Poster design from 1893 to 1935.
(Left) "Hermann Scherrer Breechesmaker Sporting-Tailor," realistic poster by Ludwig Hohlwein, 1911; lithograph. (Centre) "Nicolas," experimental poster by A.M. Cassandre for a French wine merchant, 1935; offset lithograph. (Right) "Divan Japonais," illustrative poster by Henri de Toulouse-Lautrec, 1893; lithograph.
By courtesy of the Museum of Modern Art, New York City, (left) gift of Peter Muller-Munk, (right) Abby Aldrich Rockefeller Purchase Fund

In composing lettered signs on delivery vans, trucks, and so on, the designer must take the factor of movement into account: the letters must be somewhat spaced out so that they will not run together when read on the moving vehicle. Similar care is also needed in selecting colours; as a rule dark colours are more suitable because they produce more contrast.

Although not always thought of as part of advertising, directional signs and other communication systems, if well designed, help to promote a firm's good public relations. These are the signs that figure importantly at airports, railway stations, administrative buildings, and other places where large numbers of people have to be directed to their destination by the shortest route in the shortest possible time. The signs and lettering must be uniform in design, clear and factual, and somehow correspond to the organization's advertising design or image. Large airlines, for instance, have recognized that uniform liveries for their employees and uniform lettering for their check-in counters, passageways, and so on are important for advertisement as well as for directional signals, and they have engaged graphic artists to design them.

There is an enormous variety of items that an advertiser can distribute in order to attract the attention of the public it addresses. These include calendars, matchbooks, glassware, ashtrays, notebooks, writing instruments, playthings, and decorative items of all sorts. Such items may be distributed year-round or be used as holiday gifts.

How long an object will continue to attract attention depends on the quality of the material, its workmanship, and the timelessness of its form or colour arrangement. The designer must be guided by aesthetic considerations in affixing the name of the firm, which may be stamped or printed on the object, attached to it on a label, or cast in its interior if the material is transparent. A decorative version of the name may be a factor in determining the form of the object, but often it is placed discretely on the underside. (J.M.-Br./P.J.S.L.)

DISTRIBUTION

From peddlers to chain stores

Retailing. For thousands of years most goods were sold in marketplaces or by peddlers. In tropical Africa, as in many countries in Asia, hawkers still walk the streets of cities and towns crying out their wares or travel from village to village. Marketplaces are still the chief form of retail selling in the villages of many countries. It was once thus also in Europe, where, however, during the Renaissance, market stalls in certain localities became fixed and eventually grew into stores. Many artisans' workshops expanded their retail facilities.

Chains of retail stores are known to have operated in China several centuries before the Christian Era. In 1643 a chain of pharmacies was founded in Japan. Mercantile operations of a chain-store character were carried on by the Fugger family of Germany in the 15th century, and a chain of outposts was developed in Canada by the Hudson's Bay Company following its charter in 1670. The modern chain store, however, is generally believed to stem from the founding of the Great Atlantic & Pacific Tea Company (A&P) in New York City in 1859.

The prototypes of present-day department stores existed in Japan as early as the 17th century, but not until horse-drawn buses encouraged the growth of central business districts did the department store take hold in the Western world. Au Bon Marché in Paris developed from a large specialty store to a department store about 1860. Retailers in the United States soon followed. By World War I de-

By courtesy of The Coca-Cola Company



Advertisement in *The Saturday Evening Post*, December 1931, featuring the first Santa Claus painted by Haddon Sundblom for The Coca-Cola Company. Over the years Sundblom's Santas—or Sundblom-inspired Santas—have become a familiar feature of Coca-Cola's Christmas advertising.

partment stores modeled after the French example existed in large cities throughout the world. Small independent merchants, however, still form the majority of retailers in most countries of the non-Communist world. (Most retail outlets in the Communist countries are nationalized.)

The advantages of chain stores

Chain stores and supermarkets. Centrally managed chains of stores have many advantages over independently managed retail enterprises: they can buy at more favourable terms; employ specialized buying personnel who have a wide knowledge of the supply markets; cut operating costs by integrating wholesaling and retailing functions; maintain lower inventories relative to sales; feature special price reductions on selected items; consolidate their advertising for stores in a single area; and try out new merchandising methods in one store at a time. There are also disadvantages in the chain structure. Personal interest, drive, creativity, and the will to serve the customer are often much weaker in chain operations than among independent retailers. Public sentiment has often been hostile to chains: at one time chain-store tax laws were in effect in 29 states of the United States; a Danish law prohibited retailers from owning more than a single store per municipality; the Poujadist movement in France in the late 1950s and early 1960s brought political pressure to bear in favour of small business and against the retail chain.

The importance of retail chains varies from country to country. In western Europe in the early 1980s, the estimated share of total retail sales made by chains (excluding department stores) ranged from 5 percent in Austria to more than 40 percent in several countries. In Portugal, Spain, and parts of Italy, low purchasing power and the undeveloped structure of commerce and industry inhibited the growth of retail chains. In other countries, including Austria and Denmark, they were impeded by the antichain legislation of the 1930s. The strong position of consumer cooperatives in the Scandinavian countries has made the spread of retail chains difficult.

Retailing in Communist countries is almost entirely in the hands of chains. The retail stores are often tied to wholesale establishments covering a municipality, a region, or even an entire country; in some cases, the integration extends back to producers. The degree of autonomy granted managers of individual stores also varies.

Supermarkets

Supermarkets do not always belong to chains, but in the United States, Great Britain, France, Italy, West Germany, and other highly industrialized countries the chains include about 90 percent of them. Supermarkets are large stores dealing in groceries, meats, produce, dairy products, baked goods, and household sundries; they are organized by departments and operate primarily on a self-service basis. Supermarkets also commonly handle a broad assortment of toilet goods, and they may also deal in wines and other alcoholic beverages (often depending on local licensing laws), clothing, and even cut flowers.

The pattern of supermarket development was set in the United States in 1930 when Michael Cullin opened the first unit of what was to become the King Kullen chain of supermarkets. Other independent food stores followed Cullin's lead. The major food chains gradually began to convert their service stores into supermarkets in the mid-1930s. During the two decades following World War II, when the rate of growth of supermarket chains in the United States was at its peak, tens of thousands of small independent food retailers were driven out of business. Similar developments followed in other countries. In France, for instance, the number of supermarkets and, later, hypermarkets, increased from approximately 50 in 1960 to 4,700 by 1982; during the same period the number of small retail food stores fell from some 130,000 to about 60,000.

Many operators of individual supermarkets, superettes (smaller stores operated on the same pattern), and chains with only a few retail outlets have successfully competed with large chains by banding together into retail cooperatives. The cooperatives service their members by joint buying, warehousing, store delivery, and common advertising and other promotional aids. Retail cooperatives have been strong among foodstores in the Pacific Coast states of the United States and in western Europe in nonfoods, particularly textiles.

Voluntary chains of retailers sponsored by wholesalers are common in many countries. In western Europe in the 1980s there were several huge wholesaler-sponsored chains of retailers, each including 15,000 stores or more. The 25,000 SPAR retail stores were located in 18 countries, each store using the same name and, as a rule, offering the same brands of products. Wholesaler-sponsored chains offer the same types of services for their retailer clients as do the financially integrated retail chains.

How department stores vary

Department stores and discount houses. Department stores are large retail stores that carry a wide variety of merchandise, particularly apparel and accessories, and some home furnishings. In western Europe and Japan, department stores also sell food on a large scale. They are organized in departments, each with its own merchandising, control, promotion, and service. Some departments—such as beauty parlors, restaurants, pet shops, and shoes—that have special merchandising or operational problems may be leased to outside firms.

Although department stores seldom account for more than 5 to 8 percent of total retail sales, they draw large numbers of customers in urban areas. The most influential department stores offer fashion leadership; one example is the promotion of contemporary-style furniture by the Nordiska Kompaniet (NK) department-store group of Sweden. Recreation and entertainment are often used to attract customers, as in the Japanese department stores that offer exhibitions, sideshows, and other diversions, generally without charge.

Giant department stores have also spawned chain organizations, sometimes through mergers with other department stores within the same country or by opening branch units within a region but also by expanding to other countries—for example, Sears, Roebuck and Company's operations throughout Latin America, Canada (in partnership with Simpson of that country), and Spain. Major department stores in several western European nations have branched into variety-store chains, supermarkets, specialty shops, and wholesale establishments. The laRinascente group of Italy, which began with a single department store in Milan, later developed the UPIM chain of variety stores, the SMA supermarket chain, and a buying service for other store groups, including foreign ones.

Moscow's leading department store, GUM, has a staff of well over 4,000; another Moscow department store, Children's World (Detskiy Mir), selling a wide variety of children's items, employs more than 3,000 people. Other large department stores are located in some of Moscow's residential neighbourhoods.

The practice of "discounting," or selling merchandise below the manufacturer's list price, has existed among some segments of retailing for decades. In the years following World War II, large numbers of retail stores in the United States adopted this policy of high-volume, low-margin selling, made possible by reducing services and often locating in low-rent shopping districts. The initial growth of what was often called mass merchandising was in "hard goods" or consumer durables, particularly electrical household appliances. Other goods were added later, until the discount houses (as these establishments came to be known) became essentially department stores featuring reduced prices and minimal services. Specialty discount operations have also grown significantly in foods, pharmaceutical items, forms of hardware, and other lines, usually through a chain-store organization. Some "closed-door" discount houses restrict sales to a special membership, often based upon a group affiliation.

Discount houses appeared also in Canada, France, Mexico, Sweden, Great Britain, West Germany, and elsewhere. They have been most successful in countries that are well supplied with consumer goods and in which retailing is highly competitive. In some countries their growth has been inhibited by pressures from suppliers, governmental agencies, other retail organizations, and labour unions.

A more recent development in retailing is the shopping mall. Malls, often financed by real estate developers, are usually built away from the main shopping streets in a city

or town and provide an indoor shopping area, often on several levels. Branches of large, well-known retail chains are located at key points, and a wide variety of smaller stores are fitted in between. The aim is to provide for a shopper's every need under a single roof and in attractive, climatically controlled conditions. Mall owners often provide entertainments to attract shoppers. City Plaza in Hong Kong, for example, boasts one of the territory's two ice rinks in its main concourse. Given the out-of-town locations of most shopping malls, it is essential that there be good access, and much of the site is used for free vehicle parking.

Consumer cooperatives. Consumer cooperatives are retail stores owned and operated by consumers for their mutual benefit. Most espouse the principles of the first cooperative store established in Rochdale, Eng., in 1844. These include open membership, one vote per member, limited customer services, and a sharing of the profits among members in the form of rebates, generally related to the amounts of their purchases.

Consumer cooperatives have flourished in western and northern Europe, particularly in Denmark, Finland, Iceland, Norway, Sweden, and Great Britain. The Communist countries have their own forms of consumer cooperative, usually integrated into the national distributive systems. In the Soviet Union, for example, consumer cooperatives have a near monopoly of retail sales in the villages. A number of the developing countries have tried to encourage the growth of consumer cooperatives as a means of lowering prices and of counteracting foreign-business influence.

Non-store retailing. Some retailers have no stores. The three principal types of non-store retailers are catalog houses, itinerant retailers, and vending-machine operators.

Mail-order selling

The largest catalog houses in the world—Sears, Roebuck and Company and Montgomery Ward & Company Inc.—were founded in the United States during the latter part of the 19th century. They began by selling exclusively by mail order through catalogs, although a century later both firms did more than three-quarters of their business in their retail stores. They also sold through catalog order stations in their retail stores, catalog stations located at other points, and by telephone. Other department stores and larger specialty stores frequently supplement their store operations with mail and telephone orders, often in response to advertising.

Following World War II, mail-order specialists grew rapidly in West Germany, Great Britain, and a few other highly industrialized countries. The firms of Quelle and Neckermann in West Germany also do a considerable part of their business in retail stores, in catalog service outlets, and by telephone. Development of a mail-order operation requires an efficient transportation network and a secure means of transmitting payments; in many countries these are lacking.

Itinerant retailers contact their customers in homes, at places of work, or in streets and other public places. Some sell by soliciting customers on a door-to-door basis; others work through merchandise "clubs" or personal introductions. In less-developed countries they hawk their wares in the streets.

Vending-machine operations are usually carried on in sites owned by other businesses (including retail stores), by institutions, and by transportation agencies. In the United States, vending machines deal mainly in food for immediate consumption. In other countries—Sweden, for example—vending machines have developed as a supplementary channel to retail stores where hours of business are restricted by law. High costs of manufacturing, installation, and operation have limited the expansion of vending-machine retailing.

Wholesaling. Wholesaling consists of selling to other business firms, either for resale or for business use, usually in bulk and at less than retail prices. A wholesaler is an independent merchant operating one or more wholesale establishments. Manufacturers may engage in wholesaling through their sales branches, and retailers may also do so internally through chain-store warehouses, although they are not classed as wholesalers.

Kinds of wholesalers

Wholesalers are of two kinds: full-service and limited-service. Full-service wholesalers are generally the more important in terms of sales volume; they perform a variety of services for their customers: stocking inventories, operating warehouses, supplying credit, employing salespeople to aid customers, and making deliveries. Full-service wholesalers who carry broad assortments of merchandise such as groceries are called general-line wholesalers; those who sell only narrow lines of goods, such as coffee and tea, or cigarettes, are called specialty wholesalers.

Limited-service wholesaling has emerged in West Germany, the United States, and other industrialized countries as a way of reducing the costs of service. Examples of limited-service wholesalers are "cash-and-carry" outlets, usually handling only fast-moving merchandise; "drop shippers," who ship goods directly from the manufacturer to the retailer; "truck jobbers," who sell and deliver directly from their vehicles, often for cash; and the "rack jobber," who handles the requirements of supermarkets in certain nonfood lines, such as housewares or toiletries.

In less-developed countries, wholesalers often dominate trade; they are the main elements in the distribution systems of many countries in Latin America, the Far East, and tropical Africa. In those countries the business activities of wholesalers may extend to manufacturing or retailing or branch out into nondistributive ventures, such as real estate, moneylending, or transportation. Until the latter half of the 1950s Japan was a country of wholesaling where even fairly large manufacturers and retailers relied principally on wholesalers as their intermediaries, but in more recent years, as in the United States almost a century earlier, Japanese wholesalers have decreased in importance. Even in the most highly industrialized nations, however, wholesalers remain essential to the operations of large numbers of small retailers.

In Communist countries, wholesaling performs basically the same functions as in other economies. Some capital goods pass directly from the producer to the user; in other cases, goods may move through one or more layers of wholesalers. The latter are usually supervised by trade ministries at the national or regional level.

Brokers and agents

Manufacturers may use agents in marketing their products. Agent middlemen, unlike merchant wholesalers, do not take title to the goods they sell; nor, as a rule, do they take physical possession of them. The three principal types of agent middlemen are brokers, manufacturers' agents, and sales agents.

Brokers, most commonly found in the food industries, may represent either a buyer or a seller, but most frequently the latter. Relationships are usually temporary, and brokers often represent a number of manufacturers of noncompeting products on a commission basis.

Manufacturers' agents, on the other hand, represent their principals on a continuous basis. As a rule, they sell only part of a manufacturer's output, sometimes in areas where the manufacturer cannot support full-time salespeople. Compensation is usually by commission.

Sales agents sell all of a manufacturer's output and generally have considerable freedom in setting prices. Sometimes they function in effect as a manufacturer's marketing department, even though they work on a straight commission basis. Sales agents often feed back market and product information to their principals and play an active role in product development.

Marketing goods to industry. Many goods are sold for use in producing other goods and services. These include buildings, land, machinery, component parts, semifinished goods (such as textiles, steel sheets, and other products that will undergo further processing), raw materials, and such supplies as coal, oil, or janitorial services.

Industrial channels

The marketing of goods and services to manufacturers differs markedly from marketing to final consumers. Industrial purchasers often employ professional purchasing agents backed up by engineers, testing laboratories, and production specialists. Such purchasers are generally familiar with competitive products and are prepared to negotiate. They are also more concentrated geographically: the purchasing decisions of automobile manufacturers are made in centres such as Detroit, Paris, and Turin, Italy; data-processing equipment for large-scale financial opera-

tions is sold mainly in such cities as Tokyo, London, New York, Geneva, and Zürich.

The demand for industrial goods is further distinguished by its dependence upon the demand for consumer goods that are the final products. Thus, the prosperity of the steel industry depends upon conditions in the automobile industry. Similarly, the industrial demand for screws, nuts, bolts, drafting supplies, and circuits is largely derived from the purchases of the general public.

Customers for industrial goods may be classified into three groups: users, original-equipment manufacturers, and middlemen. Users make use of the goods in their businesses—for example, the manufacturer who purchases a metal-stamping press to produce parts for his finished goods. Original-equipment manufacturers incorporate the purchased goods into their final products—for example, the manufacturer of television receivers who buys his tubes and transistors. Industrial middlemen are essentially wholesalers who distribute industrial goods to users, to original equipment manufacturers, and to other middlemen. Industrial goods wholesalers include mill-supply houses, steel warehouses, machine-tool dealers, paper jobbers, and chemical distributors; different industries often have different names for similar functions.

Marketing farm products. Channels of distribution for agricultural products follow the same general scheme as those for other goods. Individual farmers, however, often produce in relatively small quantities for the market—*e.g.*, grapes in France or milk in Wisconsin—and therefore the production of many farmers often has to be assembled as the first stage in marketing it. This is done by three principal types of middlemen: wholesale merchants, who buy and resell to processors or other merchants; merchandise agents and brokers, who do not take title to the goods and usually do not have possession of them; and such special types as commission buyers, auctioneers, and livestock concentration yards.

Assemblers of agricultural products

The auction is one of the more colourful features of agricultural marketing. Major auctions, as a rule, are devoted to a particular product indigenous to a region, such as cut flowers in San Remo, Italy, cheese in Alkmaar, Neth., and tobacco in Winston-Salem, N.C. Although sellers at the auctions are usually the producers, they may be assemblers reselling to a higher echelon of assemblers or to industrial users, retailers, or even consumers.

Farmers' cooperative associations are found throughout the world. Governments have encouraged them in order to stabilize the supply of agricultural products and raise farm productivity and incomes. In addition to marketing the farms' output, these cooperatives often purchase supplies, obtain loans, encourage research, and offer a host of services ranging from irrigation to insurance. Many cooperatives pool their members' products and market them as a single batch. Swedish cooperatives handle almost all milk production, most of the output of livestock and eggs, and a large percentage of the timber. In the United States in the 1980s, cooperatives marketed about 30 percent of farm production and handled about 20 percent of the farm supply business.

Less-developed countries have given particular encouragement to the development of agricultural cooperatives. Most sugarcane in India is sold through cooperative societies sponsored and administered by the state governments. The commodity marketing boards of West African countries are government-supervised organizations regulating the buying and selling of peanuts, palm kernels and oil, cocoa, cottonseed and lint, and other exported commodities. The boards attempt to secure the most favourable terms for the purchase, grading, export, and sales of the commodities. Cooperatives are also used in Communist countries to market the farms' products and to handle many of the functions of supply in rural districts.

Economic and social aspects of marketing

As marketing has become more influential and pervasive in industrialized countries, its development has taken two different routes. On the one hand, channels of distribution have been compressed—that is, the number of intermediaries between the original producer and the final consumer has decreased. This has come about through a process of vertical integration up and down the marketing structure; mergers, expansion, cooperation, voluntary ties, and long-term contractual relationships among distributors have all worked to this effect. On the other hand, new marketing institutions and new ways of operating have developed. Industrialization has encouraged diversity in marketing. Developments in mass communications, for example, have increased the number of ways in which advertising and promotion may be carried on. While in some respects products have become more uniform, producers make greater efforts to cater to the individualized tastes of consumers and to tailor their products for specific segments of a market.

Developments in the industrial countries

Marketing has not been without its critics. In the more industrialized countries, objections have been raised to the practice of creating "forced obsolescence" of products through superfluous changes in their design; to the economic waste involved in heavy expenditures on marketing and advertising, much of it aimed at differentiating essentially similar products; and to the use of hard-sell techniques to induce consumers to buy. Along with this has gone a more general type of criticism directed at the ends of economic activity, questioning the social desirability of promoting the use of tobacco, alcohol, conventional automobiles, supersonic aircraft, and even industrial growth in general. One response to pervasive criticism has been the growth of regulation and bureaucratic control by governments; another has been self-policing by trade associations. Consumer-protection legislation enacted in many countries has in many cases been occasioned by marketing practices that attracted particular notice.

At the same time dramatic changes in the opposite direction have taken place in the centrally planned economies, partly as a result of a recognition of the harmful effects of strict bureaucratic control on economic performance. Economic units are increasingly free to decide what they will produce, at least at the margin, after they have fulfilled their requirements under the state plan, and to dispose of the product as they choose. The most dramatic example of this gradual change of policy has occurred in rural China. By allowing peasants to grow what they wish and to sell it in the open market, that nation has become practically self-sufficient in food for the first time in decades, if not millennia. Institutions corresponding to advertising and market-research agencies have come into being in the Soviet Union and eastern Europe. Packaging of consumer goods has improved and branding has become much more common.

Advertising in particular has attracted criticism from a wide variety of viewpoints and in response has developed a well-articulated rationale in terms of both its social and its economic utility. (D.Ca./P.J.S.L.)

Advertising has been most highly developed in countries having an economic system that is basically a free-market system, in which consumer choices largely determine what goods and services will be produced. The many suppliers competing for consumer favour use advertising to influence these choices by informing consumers that a given product exists, that it offers certain satisfying qualities, that it is priced at a certain amount, and that it is available at certain retail outlets. Advertising performs these communication functions quickly and efficiently. Although personal selling can be appropriate for reaching relatively small, well-defined, highly concentrated markets such as industrial buyers, advertising is the only feasible means for communicating with millions of ultimate consumers.

Advertising in free-market economies

Manufacturers are willing to invest in research and development of new products because they know they have the means to introduce a new or improved product quickly to a large market, thereby increasing their chances of building a profitable volume of sales and recovering their initial investment in a reasonably short period of time.

The fact that large sums of money are spent on advertising and the knowledge that consumers ultimately pay this cost have led some people to believe prices would be lower without advertising. Defenders of advertising argue that this overlooks the fact that advertising facilitates mass

production with its concomitant lower unit costs and that a free-market system counts on competition to pass along the benefit of such cost reductions in the form of lower prices.

The prices consumers pay are an indicator of the value they place on the product, and value is something more than the sum total of all the production and marketing costs. Value to the consumer is the sum total of his perceptions of the product and the gratifications he attaches to it. Anything that enhances his perceptions adds value. Advertising adds value by endowing a product with subjective qualities. Consumers who pay a price 20 percent higher for a well-known advertised brand over an unadvertised brand are indicating the greater confidence, and therefore higher value, that they place on the advertised brand.

The use of advertising to build a brand preference among consumers and the fact that few firms can afford costly national advertising programs giving them the power to dominate particular product categories leads to the assumption that advertising tends to build monopolies. When a manufacturer adopts a trademark or brand name, he benefits from a common-law right to the exclusive use of that mark or name. In effect he is granted a monopoly to sell the product bearing that brand name. Whatever reputation he is able to build for the brand accrues solely to that brand. Through advertising he seeks to differentiate his brand from competing brands, to embellish it with attractions so that a large number of consumers will prefer it over all others. He thus gains a degree of monopoly power. He can charge a premium price. He can count on some stability in sales. Whatever control he gains, however, is based on the loyalty of his customers to the brand. If the price differential becomes too great, he risks losing customers to competing brands. Brand monopolies, therefore, are subject to the whims of consumers and the machinations of competitors.

The high cost of using national media such as television on a large scale restricts the use of such media to the few advertisers who can afford them. Their ability to bear the high cost of introducing new products nationally and to outspend small competitors is thus assumed to discourage would-be entrants and to make it difficult for others to survive. It is further assumed that large firms gain a larger share of market because they can afford to gain a larger share of the public mind and that they can afford to hire the most sophisticated marketing managers, thereby widening their competitive advantage. Undoubtedly, large firms with great financial resources can do many things—including large-scale advertising—that small firms cannot do. The opportunity to advertise, however, gives the small businessman a chance to get started. If he has a good product or service fairly priced, he can advertise and build a market. He obviously cannot match the giant firms on a national scale, but he might be able to outspend them in his local market area. As his business grows, he can use advertising to enter new markets and to expand his own competitive capabilities.

Effects of advertising on editorial policy

The fact that the press in most of the non-Communist world is usually supported financially by advertising raises some questions about the freedom of the press from advertiser control. Many newspapers and magazines get 60 to 70 percent of their revenue from advertising. Commercial television and radio are financed by advertising. The notion that "he who pays the fiddler calls the tunes" underlies the suspicion that advertisers influence news coverage by suppressing and distorting the news to their own advantage. There are some studies, however, that suggest that, if advertisers do dictate policy, it seems to occur among the weaker or smaller newspapers or radio stations. A financially stable newspaper is generally able to resist pressure from advertisers. Moreover, the advertiser's dependence on the press may be a deterrent to his intrusion on press freedom. He needs their circulation and audiences to advertise effectively. It is at least arguable that, if he influenced news coverage, the press would risk losing public confidence, which in turn would reduce circulation and jeopardize the effectiveness of the press as an advertising medium. (V.R.F./P.J.S.L.)

BIBLIOGRAPHY. Among the most widely used texts are PHILIP KOTLER, *Marketing Management: Analysis, Planning, and Control,* 5th ed. (1984); E. JEROME MCCARTHY and WILLIAM D. PERREAULT, JR., *Basic Marketing: A Managerial Approach,* 8th ed. (1984); ROY T. SHAW, RICHARD J. SEMENIK, and ROBERT H. WILLIAMS, *Marketing: An Integrated Analytical Approach,* 4th ed. (1981); THEODORE N. BECKMAN, WILLIAM R. DAVIDSON, and W. WAYNE TALARZYK, *Marketing,* 9th ed. (1973); JAMES M. CARMAN and KENNETH P. UHL, *Phillips and Duncan's Marketing: Principles and Methods,* 7th ed. (1973); EDWARD W. CUNDIFF and RICHARD R. STILL, *Basic Marketing: Concepts, Decisions, and Strategies,* 2nd ed. (1971); MARTIN ZOBER, *Principles of Marketing* (1971); and FREDERICK D. STURDIVANT *et al., Managerial Analysis in Marketing* (1970). See also VICTOR P. BUELL (ed.), *Handbook of Modern Marketing,* 2nd ed. (1986).

Works of a more conceptual and theoretical nature include WROE ALDERSON, *Dynamic Marketing Behavior: A Functionalist Theory of Marketing* (1965); and DAVID B. MONTGOMERY and GLEN L. URBAN, *Management Science in Marketing* (1969).

Books on international and comparative marketing include ERDENER KAYNAK (ed.), *Global Perspectives in Marketing* (1985); FRANKLIN R. ROOT, *International Trade & Investment,* 5th ed. (1984); JOHN FAYERWEATHER, *International Marketing,* 2nd ed. (1970); GORDON E. MIRACLE and GERALD S. ALBAUM, *International Marketing Management* (1970); and DAVID CARSON, *International Marketing: A Comparative Systems Approach* (1967).

Works on specific aspects of marketing include ANDRÉ GABOR, *Pricing: Principles and Practices* (1977, reprinted 1985); YORAM J. WIND, *Product Policy: Concepts, Methods, and Strategy* (1982); WILLIAM J. STANTON and RICHARD H. BUSKIRK, *Management of the Sales Force,* 6th ed. (1983); PAUL E. GREEN and DONALD S. TULL, *Research for Marketing Decisions,* 4th ed. (1978); and JAMES F. ENGEL, ROGER D. BLACKWELL, and PAUL W. MINIARD, *Consumer Behavior,* 5th ed. (1986). Information on retailing may be found in DANNY N. BELLENGER and JAC L. GOLDSTUCKER, *Retailing Basics* (1983); and WILLIAM H. BOLEN, *Contemporary Retailing,* 2nd ed. (1982).

Works on advertising in particular include C.H. SANDAGE, VERNON FRYBURGER, and KIM ROTZOLL, *Advertising Theory and Practice,* 11th ed. (1983), a comprehensive analysis of the field, including institutions, strategies, and procedures; WALTER TAPLIN, *Advertising: A New Approach* (1960, reissued 1963), a nontechnical description of how advertising affects and is affected by consumer behaviour; HARPER W. BOYD, JR., and SIDNEY J. LEVY, *Promotion: A Behavioral View* (1967), a popularized translation of behavioral science concepts to marketing communications; and DARRELL BLAINE LUCAS and STEUART HENDERSON BRITT, *Measuring Advertising Effectiveness* (1963, reprinted 1985), a review of popular copy-testing techniques and media research methods. Notable works in criticism of the advertising industry include VANCE PACKARD, *The Hidden Persuaders,* rev. ed. (1980, reprinted 1981); and SAMM SINCLAIR BAKER, *The Permissible Lie: The Inside Truth About Advertising* (1968, reprinted 1971). Historical treatments of advertising can be found in STEPHEN FOX, *The Mirror Makers: A History of American Advertising and Its Creators* (1984, reprinted 1985); DANIEL POPE, *The Making of Modern Advertising* (1983); and T.R. NEVETT, *Advertising in Britain: A History* (1982).

Current trends are reported in several trade journals, including *Marketing News* (biweekly); *Journal of Retailing* (quarterly); and *Advertising Age* (semiweekly). More scholarly research journals include the *Journal of the Academy of Marketing Science* (quarterly); *International Advertiser* (bimonthly); *Journal of Advertising Research* (bimonthly); and two journals from the American Marketing Association, *Journal of Marketing* (quarterly), which includes a review of marketing literature, and *Journal of Marketing Research* (quarterly).

(D.Ca./V.R.F./J.M.-Br./P.J.S.L.)

Markets

Markets in the most literal and immediate sense are places in which things are bought and sold. In the modern industrial system, however, the market is not a place; it has expanded to include the whole geographical area in which sellers compete with each other for customers. Alfred Marshall, whose *Principles of Economics* (first published in 1890) was for long an authority for English-speaking economists, based his definition of the market on that of the French economist A. Cournot:

> Economists understand by the term Market, not only particular market place in which things are bought and sold, but the whole of any region in which buyers and sellers are in such free intercourse with one another that the prices of the same goods tend to equality easily and quickly.

To this Marshall added:

> The more nearly perfect a market is, the stronger is the tendency for the same price to be paid for the same thing at the same time in all parts of the market.

The concept of the market as defined above has to do primarily with more or less standardized commodities such as wool or wheat or shoes or automobiles. The word is also used to deal with, for example, the market for real estate or for old masters; and there is the "labour market," although a contract to work for a certain wage is not quite the same thing as the sale of a packet of goods. There is a connecting idea in all of these various usages—namely, the interplay of supply and demand.

Most markets consist of groups of intermediaries between the first seller of a commodity and the final buyer. There are all kinds of intermediaries, from the brokers in the great produce exchanges down to the village grocer. They may be mere dealers with no equipment but a telephone, or they may provide storage and perform important services of grading, packaging, and so on. In general, the function of a market is to collect products from scattered sources and channel them to scattered outlets. From the point of view of the seller, dealers channel the demand for his product; from the point of view of the buyer, they bring supplies within his reach.

There are two main types of markets for products, in which the forces of supply and demand operate quite differently, with some overlapping and borderline cases. In the first, the producer offers his goods and takes whatever price they will command; in the second, the producer sets his price and sells as much as the market will take. In addition, along with the growth of trade in goods, there has been a proliferation of financial markets, including securities exchanges and money markets.

The article is divided into the following sections:

The market in economic doctrine and history 552
  Market theory 552
    The abstract nature of traditional market theory
    Modifications of the theory
  The historical development of markets 553
    The origin of markets
    Markets under Socialism
Markets for primary products 554
  Commodity markets 554
  Futures markets 554
    Economic functions of the futures contract
    The theory and practice of hedging
    Important futures markets
Markets for manufactures 556
Financial markets 556
  Securities trading 557
    Types of corporate securities
    The marketing of new issues
    The machinery of securities trading
    The structure of demand for securities
    Dynamics of the securities markets
  Money market 562
    Banks and the money market
    The international money market
    The U.S. money market
    The British money market
    The money markets of other countries
The markets and social welfare 566
  The politics of the market 566
  Market Socialism 567

The market in economic doctrine and history

MARKET THEORY

The abstract nature of traditional market theory. The key to the modern concept of the market may be found in the famous observation of the 18th-century British economist Adam Smith that "The division of labour depends upon the extent of the market." He foresaw that modern industry depended for its development upon an extensive market for its products. The factory system developed out of trade in cotton textiles, when merchants, discovering an apparently insatiable worldwide market, became interested in increasing production in order to have more to sell. The factory system led to the use of power to supplement human muscle, followed in turn by the application of science to technology, which in an ever-accelerating spiral has produced the scope and complexity of modern industry.

19th-century views

The economic theory of the late 19th century, which is still influential in academic teaching, was, however, concerned with the allocation of existing resources between different uses rather than with technical progress. This theory was highly abstract. The concept of the market was most systematically worked out in a general equilibrium system developed by the French economist Léon Walras, who was strongly influenced by the theoretical physics of his time. His system of mathematical equations was ingenious, but there are two serious limitations to the mechanical analogy upon which they were based: it omitted the factor of time—the effect upon peoples' present behaviour of their expectations about the future; and it ignored the consequences for the human beings concerned of the distribution of purchasing power among them. Though economists have always admitted the abstract nature of the theory, they generally have accepted the doctrine that the free play of market forces tended to bring about full employment and an optimum allocation of resources. On this view, unemployment could only be caused by wages being too high. This doctrine was still influential in the Great Depression of the 1930s.

The turn toward realism

Modifications of the theory. The change in view that was to become known as the Keynesian Revolution was largely an escape to common sense, as opposed to abstract theory. In a private-enterprise economy, investment in industrial installations and housing construction is aimed at profitability in the future. Because investment therefore depends upon expectations, unfavourable expectations tend to fulfill themselves—when investment outlay falls off, workers become unemployed; incomes fall, purchases fall, unemployment spreads to the consumer goods industries, and receipts are reduced all the more. The operation of the market thus generates instability. The market may also generate instability in an upward direction. A high level of effective demand leads to a scarcity of labour; rising wages raise both costs of production and incomes so that there is a general tendency to inflation.

While the English economist John Maynard Keynes was attacking the concept of equilibrium in the market as a whole, the notion of equilibrium in the market for particular commodities was also being undermined. Traditional theory had conceived of a group of producers as operating in a perfect market for a single commodity; each produced only a small part of the whole supply; for each, the price was determined by the market; and each maximized its profits by selling only as much as would make marginal cost equal to price—that is to say, only so much that to produce a little more would add more to costs than it would to proceeds. Each firm worked its plant up to capacity; *i.e.*, to the point where profitability was limited by rising costs. This state of affairs, known as "perfect competition," is quite contrary to the general run of business experience, particularly in bad times when under-capacity working is prevalent. A theory of imperfect competition was invented to reconcile the traditional theory with under-capacity working. This in its turn was attacked as unrealistic. The upshot was a general recognition that strict profit maximizing is impossible in conditions of uncertainty; that prices of manufactures are generally formed by adding a margin to direct costs, large enough to yield a profit at less than capacity sales; and that an increase in capacity generally has to be accompanied by some kind of selling campaign to ensure that it will be used at a remunerative level.

Once it is recognized that competition is never perfect in reality, it becomes obvious that there is great scope for individual variations in the price policy of firms. No precise generalization is possible. The field is open for study of what actually happens, and exploration is going on. Meanwhile, however, textbook teaching often continues to seek refuge in the illusory simplicity of the traditional theory of market behaviour.

THE HISTORICAL DEVELOPMENT OF MARKETS

Economies without markets

History and anthropology provide many examples of economies based neither on markets nor on commerce. An exchange of gifts between communities with different resources, for example, may resemble trade, particularly in diversifying consumption and encouraging specialization in production, but subjectively it has a different meaning. Honour lies in giving; receiving imposes a burden. There is competition to see who can show the most generosity, not who can make the biggest gain.

Another kind of noncommercial exchange was the payment of tribute, or dues, to a political authority, which then distributed what it had collected. On this basis, great, complex, and wealthy civilizations have arisen in which commerce was almost entirely unknown: the whole network of supply and distribution was operated through the administrative system. Herodotus remarked that the Persians had no marketplaces.

The distinguishing characteristic of commerce is that goods are offered not as a duty or for prestige or out of neighbourly kindness but in order to acquire purchasing power. It is clearly a convenience to all parties to have a single generally established currency-commodity. Once a commodity is acceptable as money, its use to store purchasing power overshadows its use for its original purpose; it ceases to be a commodity like any other and becomes the very embodiment of value.

The origin of markets. Markets as centres of commerce seem to have had three separate points of origin. The first was in rural fairs. A typical cultivator fed his family and paid the landlord and the moneylender from his chief crop. He had sidelines that provided salable products, and he had needs that he could not satisfy at home. It was then convenient for him to go to a market where many could meet to sell and buy.

The second point was in service to the landlords. Rent, essentially, was paid in grain; even when it was translated into money, sales of grain were necessary to supply the cultivator with funds to meet his dues. Payment of rent was a one-way transaction, imposed by the landlord. In turn, the landlord used the rents to maintain his warriors, clients, and artisans, and this led to the growth of towns as centres of trade and production. An urban class developed with a standard of life enabling its members to cater to each other as well as to the landlords and officials.

International markets

The third, and most influential, origin of markets was in international trade. From early times merchant adventurers (the Phoenicians, the Arabs) risked their lives and their capital in carrying the products of one region to another. The importance of international trade for the development of the market system was precisely that it was carried on by third parties. Within a settled country, commercial dealings were restrained by considerations of rights, obligations, and proper behaviour. In medieval Europe, for example, dealings were regulated in the main by the concept of the "just price," that is, a system of valuations that assured the producers and merchants an income sufficient to maintain life at a level suited to their respective positions in society. But in trade in which the dealer is not subject to any obligation at either end, no holds are barred; purely commercial principles have free play. It was in trade (for instance, the export of English wool to the weavers of Italy) that the commercial principle undermined feudal conceptions of rights and duties. As Adam Smith observed, a great leap occurred when trade released the forces of industrial production.

Throughout history the relations between the trader and the producer have changed with the development of technique and with changes in the economic power of the parties. The 19th century was the heyday of the import–export merchant. Traders from a metropolitan country could establish themselves in a foreign centre, become experts on its needs and possibilities, and deal with a great variety of producers and customers, on a relatively small scale with each. With the growth of giant corporations, the scope of the merchant narrowed; his functions were largely taken over by the sales departments of the industrial concerns. Nowadays it is common to hold international fairs at which industrial products are displayed for inspection by customers, a grand and glorified version of the village market; the business, however, consists in placing orders rather than buying on the spot and carrying merchandise home. The function of the independent wholesaler, like that of the merchant, has declined as great retail businesses have grown to a scale whereby they can deal directly with manufacturers; but specialized exchanges for primary commodities are still important.

Markets under Socialism. Markets are essential to the free enterprise system; they grew and spread along with it. The propensity "to truck, barter, and exchange one thing for another" (in Adam Smith's words) was exalted into a principle of civilization by the doctrine of laissez-faire, which taught that the pursuit of his self-interests by the individual would be to the benefit of society as a whole. In the Soviet Union and other Socialist countries, a different kind of economy exists and a different ideology is dominant. There are two interlocking systems in the economy of the Soviet Union: one for industry and one for agriculture; and the same pattern is followed, with variations, in the other Socialist countries. Industrially, all equipment and materials are owned by the state, and production is directed according to a central plan. In theory, payments to workers are thought of as their share of the total production of the economy; in practice, however, the system of wages is very much like that in capitalist industry except that rates as a rule are set by decree and the managers of enterprises have little scope for bargaining. Workers may move around looking for jobs, but there is no "labour market" in the capitalist sense. Materials and equipment are distributed among enterprises by the state planning offices. (Faulty planning has given rise to intermediaries who operate between enterprises, but this is not at all the same thing as the highly developed markets in materials, components, and equipment that exist under capitalism.)

Markets in Communist-governed countries

Consumption goods, on the other hand, are distributed to Soviet households through a retail market. Though some Socialist idealists, regarding buying and selling as the essence of capitalism, have advocated that money should be abolished altogether, in a large community it has proved to be most convenient to provide incomes in the form of generalized purchasing power and to allow each to choose what he pleases from whatever goods are available. Classical economists usually assert that the advantage of the retail market system is that it runs itself without excessive regulation; the housewife who goes shopping is in charge

of her own money and need account to no one for what she does with it. Retail markets in Soviet-type economies are said to differ from those in capitalist economies in that while in both systems the buyer is in this sense a principal, the seller in the Soviet system is an agent. Retailers and manufacturers all serve as agents of the same authority—the central plan. Rather than making it their business to woo and cajole the customer, sellers throw supplies into the shops in a somewhat arbitrary way and the customer searches for what he wants.

Soviet agriculture is organized on principles quite different from those operative for manufacturing. Collective farms, though managed in an authoritarian way, are like cooperatives in which members share in the income of their farm in respect to the "work points" each can earn. The value of a work point is affected by the prices set for the products of the farm, and these are politically, rather than only economically, determined. In the Western industrial economies, there is also a political element involved in the setting of agricultural prices; generally the problem here is to prevent excess production from driving prices too low. For the Soviets, the problem was the opposite. There, agricultural output failed to expand rapidly enough to keep pace with the requirements of the growing industrial labour force, and prices were therefore kept down so that they would not be unfavourable to the industrial sector. At the same time, individual members of the collective farms were permitted to sell the produce of their household plots on a free market. In this specific market, the peasant was as much a principal as the buyer.

In China, cooperative farms established after 1949 were much more genuinely cooperatives than were those in the Soviet Union, and trade with the cities in China is organized through a kind of Socialist wholesaling. City authorities place contracts with neighbouring farms, specifying prices, varieties, quantities, and delivery dates, and then direct the supplies to retail outlets, which are part of the Socialist economy. A similar system controls trade in manufactured consumer goods. Through the retail shops, the authorities keep a finger on the pulse of demand and guide supply as far as possible to meet it by the contracts that they place with the Socialist manufacturers. By adapting to its own requirements a particular element in the market system—wholesale trade—the Chinese economy seems to have avoided some of the difficulties that the Soviets have encountered.

The kibbutzim in Israel

An example of socialism without a formal market was seen in the early days of the cooperative settlements known as kibbutzim in Israel, where cultivators shared the proceeds of their work without any distinction of individual incomes. (Because a kibbutz could trade with the surrounding market economy, its members were not confined to consuming only the produce of their own soil.) At the outset some of the kibbutzim carried the objection to private property so far that a man who gave a shirt to the laundry received back just some other shirt. But to dispense altogether with market relationships is apparently possible only in a small community in which all share a common ideal, and the austere standards of the original kibbutzim have softened somewhat with growing prosperity; but they still maintain a small-scale example of economic efficiency without commercial incentives.

Markets for primary products

COMMODITY MARKETS

The behaviour of primary commodity markets

The general run of agricultural commodities is produced under competitive conditions by relatively small scale cultivators scattered over a large area. The final purchasers are also scattered, and centres of consumption are distant from regions of production. The dealer, therefore, since he is indispensable, is in a stronger economic position than the seller. This situation is markedly true when the producer is a peasant who lacks both commercial knowledge and finance so that he is obliged to sell as soon as his harvest comes in; it is true also, though to a lesser extent, of the capitalist plantation for which the only source of earnings is a particular specialized product. In this kind of business, both demand and supply are said to be inelastic in the short run—that is, a fall in price does not have much effect in increasing purchases and a rise in price cannot quickly increase supplies. Supplies are subject to natural variations, weather conditions, pests, and so forth; and demand varies with the level of activity in the centres of industry and with changes in tastes and technical requirements. Under a regime of unregulated competition such markets are, therefore, tormented with continual fluctuations in prices and volume of business. Though dealers may mitigate this to some extent by building up stocks when prices are low and releasing them when demand is high, such buying and selling often turns into speculation, which tends to exacerbate the fluctuations.

The behaviour of primary commodity markets is a serious matter when whole communities depend upon a single commodity for income or for employment and wages. The agricultural communities that form part of an industrial economy are therefore generally sheltered from the operation of supply and demand by government regulations of various types, price supports, or tariff protection. Though some attempts have been made to control world commodity markets, these are generally more talk than performance. Some nations, Australia for example, have been able to make enough profit from primary commodity exports to attract capital into the development of industry; but most of the so-called developing countries find their export earnings insecure and insufficient. Their spokesmen complain that the world market system operates in favour of the industrialized nations. (J.Ro.)

FUTURES MARKETS

Origins and development

From very early times, and in many lines of trade, buyers and sellers have found it advantageous to enter into contracts—termed futures contracts—calling for delivery of a commodity at a later date. Dutch whalers in the 16th century entered into forward sales contracts before sailing, partly to finance their voyage and partly to get a better price for their product. Potato growers in Maine in the U.S. from early times made forward sales of potatoes at planting time. The European futures markets arose out of import trade. Cotton importers in Liverpool, for example, entered forward contracts with U.S. exporters from around 1840. With the introduction of the fast transatlantic Cunard mail services, it became possible for cotton exporters in the United States to send samples to Liverpool in advance of the slow cargo ships, which carried the bulk of the cotton. Futures trading within the United States in the form of "to arrive" contracts appears to have commenced before the railroad days (1850s) in Chicago. Merchants in Chicago who bought wheat from outlying territories were not sure when they would obtain delivery or what their quality would be. Under these conditions, the introduction of "to arrive" contracts enabled the sellers to get a better price for their product and buyers to avoid serious price risk.

Futures trading of this sort in grains, coffee, cotton and oilseeds also arose in other centres such as Antwerp, Amsterdam, Bremen, Le Havre, Alexandria, and Ōsaka between the 17th and the middle of the 19th centuries. In the process of evolution, "to arrive" contracts became standardized with respect to grade and delivery period, with allowances for grade adjustment when the delivered grade happened to be different. These developments helped to enlarge the volume of trade, encouraging more trading by merchants who dealt in the physical commodity and also the entry of speculators, who were interested not in the commodity itself but in the favourable movement of its price in order to make profits. The larger volume of trading lowered the transaction costs, and by stages the trading became impersonal. The rise of the clearing house depersonalized the buyer-seller relations completely, giving rise to the present form of futures trading.

Economic functions of the futures contract. Commodity futures markets provide insurance opportunities to merchants and processors against the risk of price fluctuation. In the case of a trader, an adverse price change brought either by supply or demand change affects the total value of his commitments; and the larger the value of his inventory, the larger the risk to which he is exposed. The futures

market provides a mechanism for the trader to lower the per unit inventory risk on his commitments in the cash market (where actual physical delivery of the commodity must eventually be made) through what is known as hedging. A trader is termed a hedger if his commitments in the cash market are offset by opposite commitments in the futures market. An example would be that of a grain elevator operator who buys wheat in the country and at the same time sells a futures contract for the same quantity of wheat. When his wheat is delivered later to the terminal market or to the processor in a normal market, he buys back his futures contract. Any change of price that occurred during the interval should have been cancelled out by mutually compensatory movements in his cash and futures holdings. The hedger thus hopes to protect himself against loss resulting from price changes by transferring the risk to a speculator who relies upon his skill in forecasting price movements.

For a better understanding of the process involved, the distinctive features of the cash market and the futures market should be made clear. The cash market may be either a spot market concerned with immediate physical delivery of the specified commodity or a forward market, where the delivery of the specified commodity is made at some later date. Futures markets, on the other hand, generally permit trading in a number of grades of the commodity to protect hedger sellers from being "cornered" by speculator buyers who might otherwise insist on delivery of a particular grade whose stocks are small. Since a number of alternative grades can be tendered, the futures market is not suitable for the acquisition of the physical commodity. For this reason, physical delivery of the commodities in fulfillment of the futures contract generally does not take place, and the contract is usually settled between buyers and sellers by paying the difference between the buying and selling price. Several futures contracts in a commodity are traded during a year. Thus, five wheat contracts, July, September, December, March, and May, and six soybean contracts, September, November, January, March, May and July are traded on the Board of Trade of the City of Chicago. The length of these contracts is for a period of about 10 months, and a contract for "September wheat" or "September soybean" indicates the month the contract matures.

Hedging as insurance

Though hedging is a form of insurance, it seldom provides perfect protection. The insurance is based on the fact that the cash and futures prices move together and are well correlated. The price spread between the cash and futures, however, is not invariant. The hedgers, therefore, run the risk that the price spread, known as the "basis," could move against them. The possibility of such an unfavourable movement in the basis is known as basis risk. Thus hedgers, through their commitment in the futures market, substitute basis risk for the price risk they would have taken in carrying unhedged stocks. It must be emphasized, however, that risk reduction is not the final objective with merchants and processors; what they seek to do is to maximize profits.

The availability of capital for financing the holding of inventories depends on whether they are hedged or not. The bankers' willingness to finance them increases with the proportion of the inventory that is hedged. For example, the banks may advance loans to the extent of only 50 percent of the value of unhedged inventories and 90 percent if they are all hedged, a difference explained by the fact that hedging reduces the risk on which the amount of the loan and the interest rate depend. Merchants and processors can therefore derive a twofold advantage from futures trading; they can insure against price decline and they can secure larger and cheaper loans from the banks.

The theory and practice of hedging. There are two rival hypotheses concerning the motives for and costs of hedging. The first of these, advanced by John Maynard Keynes and J.R. Hicks, suggests that risk reduction is the prime motive for hedging and that hedgers pay a risk premium to speculators for assuming risk. The Keynes-Hicks hypothesis states that under normal conditions in commodity markets, when demand, supply, and spot prices are expected to remain unchanged for some months to come and there is uncertainty in traders' minds regarding these expectations, the futures price, say, for one month's delivery is bound to be below the spot price that traders expect to prevail one month later. This condition exists because inventory holders would be ready to hedge themselves from the risk of price fluctuations by selling futures to speculators below the expected spot price. By selling futures below the expected spot price, according to the theory, inventory holders who hedge pay a risk premium to speculators.

The rival hypothesis of Holbrook Working maintains that hedging is done with the expectation of a profit from a favourable change in the spot-futures price relation, to simplify business decisions, and to cut costs, and not for the sake of reducing risk alone. Hedgers, according to Working, are arbitrageurs; *i.e.*, they take advantage of a temporary price difference between two markets to buy in one and sell in the other. They thus speculate on the basis and assume risk.

A compromise between these rival theories and a more balanced view regarding the need for hedging and the scope of hedging activities is that hedging is motivated by the desire to reduce risks, as suggested by the Keynes-Hicks theory, but that the levels of inventory held by merchants and processors are determined by expected hedging profits, as Working has emphasized.

The long and short hedger

There are two categories of hedgers in the futures market: they are called short and long hedgers. Short hedgers are merchants and processors who acquire inventories of the commodity in the spot market and who simultaneously sell an equivalent amount or less in the futures market. The hedgers in this case are said to be long on their spot transaction and short in the futures transaction. Wheat merchants or wheat flour mills who either have 100,000 bushels of wheat as inventory or have bought it for later delivery are said to short hedge if they sell 100,000 bushels of wheat in futures contracts. By holding inventories, both merchants and processors can make their purchases when it is most opportune and lower their transaction costs through fewer transactions. Another advantage to the processing firm in holding inventory is that it makes it possible to avoid interruption in production. It must be borne in mind that short hedgers do not normally deliver the physical commodity in fulfillment of the futures contract. They "lift the hedge" by repurchasing the futures contract at the prevailing futures price when they sell the raw material or the processed good in the spot market.

The merchants and processors do not generally hedge all their inventories for the sake of reduced risk. The decision on what part of inventories to hedge is based on their expectations relating to return from holding hedged and unhedged inventories in storage, given the cost incurred in both forms of inventory holding. The return per unit inventory to merchants and processors on their hedged inventories, when liquidated, is the change in the spot price less the change in the futures price and the storage costs. Their return on per unit unhedged inventory is the change in spot price less storage costs.

Long hedgers, in contrast, are merchants and processors who have made formal commitments to deliver a specified quantity of raw material or processed goods at a later date at a price currently agreed upon and who do not now have the stocks of the raw material necessary to fulfill their forward commitment. The parties who have made the commitment generally seek to hedge against the risk of price rise in the raw material between the time of making the forward contract and the time of acquiring the raw material stocks for fulfilling the contract. The hedging is done by buying futures contracts of the raw material equal in quantity to what is needed to fulfill the forward commitment.

The question arises under what circumstances the long hedger might prefer the purchase of futures to the alternative of immediately buying the raw material through spot or forward purchase to meet the obligations of his forward sale. He may prefer buying futures to buying in the cash market (spot or forward) if current cash prices are high because of scarcity. Generally there is an increase in the amount of long hedging when, as the season advances,

spot prices rise, inventory holdings fall, and the new crop is not yet available. Long hedging is not as risk-reducing as it may appear at first sight. The long hedger processor, for example, who buys raw material futures to satisfy his forward commitment of the processed good may find that the raw material delivered to him in futures is not of suitable grade and quality to meet the obligations of the forward sale. Quite often, therefore, he may sell his futures contract and purchase raw material of the grade needed. If the spot price of the raw material moves unfavourably relative to the price of the processed good sold forward by him, the long hedger actually increases the risk by buying futures instead of buying the raw material in the cash market. Long hedging, unlike short hedging, may serve to increase risk, and the total risk on long hedging increases with the size of the commitment.

The volume of short hedging tends to be large when stocks in commercial hands are large and when the cash price is below the futures price; a reversal in this situation brings a decline. Conversely, the volume of long hedging is large when stocks are small and the cash price is above the futures price. Short hedging has a marked seasonal pattern, reaching a peak when commercial stocks are largest and the basis is favourable and then declining as the season advances. The seasonal pattern is less marked in long hedging. Generally there is an excess of short over long hedging during the bulk of the crop year.

The speculator's role

Apart from hedgers, the futures market includes speculators, and these can also be classified in two categories, namely, long and short speculators. The long speculators are those who expect the price to rise above the current level and assume risks by purchasing futures contracts. Short speculators are those who expect the price to fall. They sell futures contracts. In a futures market the total short selling position, made up of short hedgers and short speculators, and the total long buying position, made up of long hedgers and long speculators, must always be equal. Any excess of short over long hedging must be balanced by an equal excess of long over short speculation. Since short hedging exceeds long hedging for most of the crop year, hedgers are generally short and speculators, therefore, are generally long.

Futures markets have flourished and become important in commodities where sizeable inventories have to be stored and carried forward for meeting the consumption needs of the entire season. Successful futures trading requires a large volume with low transaction costs and that spot and futures prices be well correlated in order to make hedging effective.

Important futures markets. Based on the number and volume of commodities in which active futures trading exists, the United States occupies first place. The Chicago Board of Trade, the largest of the world's futures markets in terms of volume and value of business, is the centre for trading in wheat, corn, oats, rye, soybeans, soybean oil, and soybean meal. About 30 commodities in all are traded on organized exchanges in the United States. The wheat market in Minneapolis, the cotton and wool markets in New York City, and the markets in frozen pork bellies and live hogs in the midwestern United States are among them. The number of commodities in which futures trading takes place are far fewer outside the United States.

Chicago Board of Trade

There are futures markets for wool in London, Paris, and Sydney; for cotton in Liverpool and Bombay; for sugar in London and Paris; for jute goods in Calcutta; for black pepper in Cochin, India; and for turmeric in Sāngli, India. As a result of government controls on futures markets and also of international commodity agreements, the volume of futures trading in several countries has been adversely affected. The commodity markets in Europe, with few exceptions, have been dormant since the end of World War II. Many of the Indian commodity markets, such as those in gur, jute, and oilseeds, which were once active, have met the same fate. The recurrent arguments in the United States, India, and elsewhere against the futures markets are that they encourage speculation and that the participation of speculators causes price instability. These arguments have led to the demand that markets be controlled or prohibited from functioning. To refute such allegations requires a comparison between price variations in the presence and absence of speculation, which is impossible for commodities that have futures markets, since it is not meaningful to say for these markets what the price would have been in the absence of speculation. (L.S.V.)

Markets for manufactures

The market for manufactured goods is what economists call "imperfect," because each company has its own style, its own reputation, and its own locations; and all of the arts of advertisement and salesmanship are devoted to making it even more imperfect by attracting buyers to particular brand names. Even small businesses that depend upon outside channels of retail distribution may have the final say in what prices they will charge, and great corporations can differentiate their goods in order to create demand for them.

In this type of market, supply normally is very elastic—that is, responsive to demand—in the short run. Stocks or inventories are held at some point in the chain of distribution; while stocks are running down or building up, there is time to change the level of production, and once a price has been set, it is rarely altered in response to moderate changes in demand. Even in a deep slump, defensive rings may often be formed to prevent price cutting.

In the long run, as well as in the short, supply is responsive to demand in the market for manufactures. It is easier to change the composition of a firm's output than it is to change the production of a mine or a plantation. And when changes in demand are not too rapid, gross profits from one plant can be siphoned off and invested in something quite different. When business is good, moreover, there is continual new investment so that productive capacity is adapted to meeting changing requirements. Workers themselves may not even be aware of changes in the final commodities to which their work contributes, and the level of wages for any grade of factory labour is very little affected by the fortunes of a particular market.

Financial markets

Along with a growth of trade in goods, there has been a proliferation of financial markets. A stock exchange is an organized market for dealing in the securities of businesses and governments. Currently, dealings in securities that came into being in the past predominate over dealings in new issues, and the greater part of industrial investment is financed by retention of profits. Instead of serving mainly as a channel for lending to industry, the chief functions of a stock exchange are to provide a convenient way for owners of property, inherited or newly saved, to place their wealth in income-yielding form, and to provide liquidity and security for them by making a market in which financial assets can readily be realized or redeployed. Dealings in a commodity market, as noted earlier, necessarily contain a speculative element; but in the market for securities, speculation overshadows all else. The art of speculation, as Keynes said, is "to anticipate what average opinion expects average opinion to be." The notorious instability of stock exchanges, with their disturbing effects upon the availability of finance for business and upon the nominal wealth of shareholders, may have repercussions in the market for real goods and services.

Stock markets and money markets

Another kind of financial market is the so-called money market. The money market is not an organized exchange; nor is it confined to money in the ordinary sense. The phrase refers to what might be called wholesale transactions in money and short-term credit carried on mainly by large commercial banks. Through the banking system, government authorities (say, the Bank of England or the United States Federal Reserve Board) have some power to control the market for money. But the more the authorities attempt to exercise their powers, the faster the money market develops new organs for bringing lenders and borrowers together and new types of transferrable obligations that are "almost money," in order to escape from control. (An example of this is the huge Eurocurrency and Eurobond markets in which money and financial instruments

are bought and sold independently of national monetary authorities.) (J.Ro.)

SECURITIES TRADING

Securities are written evidences of ownership giving their holders the right to demand and receive property not in their possession. The most common types of securities are stocks and bonds, of which there are many particular kinds designed to meet specialized needs. This article deals mainly with the buying and selling of securities issued by private corporations. (The securities issued by governments are discussed in the article GOVERNMENT FINANCE.)

Types of corporate securities. Corporations create two kinds of securities: bonds, representing debt, and stocks, representing ownership or equity interest in their operations. (In Great Britain, the term stock ordinarily refers to a loan, whereas the equity segment is called a share.)

Bonds. The bond, as a debt instrument, represents the promise of a corporation to pay a fixed sum at a specified maturity date, and interest at regular intervals until then. Bonds may be registered in the names of designated parties, as payees, though more often, in order to facilitate handling, they are made payable to the "bearer." The bondholder usually receives his interest by redeeming attached coupons.

Since it could be difficult for a corporation to pay all of its bonds at one time, it is common practice to pay them gradually through serial maturity dates or through a sinking fund, under which arrangement a specified portion of earnings is regularly set aside and applied to the retirement of the bonds. In addition, bonds frequently

Calling bonds

may be "called" at the option of the company, so that the corporation can take advantage of declining interest rates by selling new bonds at more favourable terms and using these funds to eliminate older outstanding issues. In order to guarantee the earnings of investors, however, bonds may be noncallable for a specified period, perhaps for five or 10 years, and their redemption price may be made equal to the face amount plus a "premium" amount that declines as the bond approaches its maturity date.

The principal type of bond is a mortgage bond, which represents a claim on specified real property. This protection ordinarily results in the holders' receiving priority treatment in the event that financial difficulties lead to a reorganization. Another type is a collateral trust bond, in which the security consists of intangible property, usually stocks and bonds owned by the corporation. Railroads and other transportation companies sometimes finance the purchase of rolling stock with equipment obligations, in which the security is the rolling stock itself.

Although in the United States the term debentures ordinarily refers to relatively long-term unsecured obligations, in other countries it is used to describe any type of corporate obligation, and "bond" more often refers to loans issued by public authorities.

Corporations have developed hybrid obligations to meet varying circumstances. One of the most important of these is the convertible bond, which can be exchanged for common shares at specified prices that may gradually rise over time. Such a bond may be used as a financing device to obtain funds at a low interest rate during the initial stages of a project, when income is likely to be low, and encourage conversion of the debt to stock as earnings rise. A convertible bond may also prove appealing during periods of market uncertainty, when investors obtain the price protection afforded by the bond segment without materially sacrificing possible gains provided by the stock feature; if the price of such a bond momentarily falls below its common-stock equivalent, persons who seek to profit by differentials in equivalent securities will buy the undervalued bond and sell the overvalued stock, effecting delivery on the stock by borrowing the required number of shares (selling short) and eventually converting the bonds in order to obtain the shares to return to the lender.

Income bonds

Another of the hybrid types is the income bond, which has a fixed maturity but on which interest is paid only if it is earned. These bonds developed in the United States out of railroad reorganizations, when investors holding defaulted bonds were willing to accept an income obligation in exchange for their own securities because of its bond form; the issuer for his part was less vulnerable to the danger of another bankruptcy because interest on the new income bonds was contingent on earnings.

Still another hybrid form is the linked bond, in which the value of the principal, and sometimes the amount of interest as well, is linked to some standard of value such as commodity prices, a cost of living index, a foreign currency, or a combination of these. Although the principle of linkage is old, bonds of this sort received their major impetus during the inflationary periods after World Wars I and II. In recent years they have had the most use in countries in which the pressures of inflation have been sufficiently strong to deter investors from buying fixed-income obligations.

Stock. Those who provide the risk capital for a corporate venture are given stock, representing their ownership interest in the enterprise. The holder of stock has certain rights that are defined by the charter and bylaws of the corporation as well as by the laws of the country or state in which it is chartered. Typically these include the right to share in dividends and other distributions, to vote for directors and fundamental corporate changes, and to inspect the books of the corporation, and, less frequently, the "pre-emptive right" to subscribe to any new issue of stock. The stockholder's interest is divided into units of participation, called shares.

A stock certificate ordinarily is given as documentary evidence of share ownership. Originally this was its primary function; but as interest in securities grew and the capital market evolved, the role of the certificate gradually changed until it became, as it is now, an important instrument for the transfer of title. In some European countries the stock certificate is commonly held in bearer form and is negotiable without endorsement. To avoid loss, the certificates are likely to be entrusted to commercial banks or a clearing agency that is able to handle much of the transfer function through offsetting transactions and bookkeeping entries. In the United States, certificates usually are registered in the name of the owner or in a "street name"—the name of the owner's broker or bank; the bank may for legal reasons use the name of another person, known as a "nominee." When a certificate is held in the name of a broker or bank nominee, the institution is able to make delivery more readily and the transfer process is facilitated. Investors, for legal or personal reasons, may prefer to keep the certificates in their own names.

A corporation may endow different kinds or classes of stock with different rights. Preferred stock has priority with respect to dividends and, if the corporation is dissolved, to the division of assets. Dividends on preferred stock usually are paid at a fixed rate and are often cumulated in the event the corporation finds it necessary to omit a distribution. In the latter circumstance the full deficiency must be cleared before payments may be made on the common shares. Participating preferred stock, in addition to stipulated dividends, receives a share of whatever earnings are paid to the common stock. Participation is usually resorted to as an inducement to investors when the corporation is financially weak. Although a preferred issue has no maturity date, it may be given redemption terms much like those of a bond, including a conversion privilege and a sinking fund. Preferred stockholders may or may not be allowed to vote equally with common stockholders on some or all propositions or more characteristically may vote only upon the occurrence of some prescribed condition, such as the default of a specified number of dividend payments.

Common stock, in some countries called ordinary shares, represents a residual interest in the earnings and assets of a corporation. Whereas distributions to bonds or preferred stock are ordinarily fixed, dividends paid on common stock are set at the time of payment by the directors and tend to vary with earnings. The market price of common stock is likely to move in a relatively wide range, depending on investors' expectations of earnings in the future.

Options. An option contract is an agreement enabling the holder to buy a security at a fixed price for a limited period of time. One form of option contract is the stock

Warrants

purchase warrant, which entitles the owner to buy shares of common stock at designated prices and according to a prescribed ratio. Warrants are often used to enhance the salability of a senior security, and sometimes as part of the compensation paid to bankers who market new issues.

Another use of the option contract is the employee stock option. This is used to compensate key executives and other employees; it is normally subject to a variety of restrictions and is generally nontransferrable. Stock rights, like warrants, are transferrable privileges permitting stockholders to buy another security or a portion thereof at a specified price for an indicated period of time. The stock right allows stockholders to subscribe to additional shares of stock in proportion to their present holdings. Stock rights usually have a shorter life-span than warrants, and their subscription price is below, rather than above, the market price of the common stock.

The marketing of new issues. The marketing of securities is an essential link in the mechanism that transfers capital funds from savers to users. The transfer may involve intermediaries such as savings banks, insurance companies, or investment trusts. The ultimate user of the funds may be a corporation or any of the various levels of government from municipalities to national states.

The growth of public debt throughout the world has made governments increasingly important participants in the markets for new securities. They have had to develop financing techniques with careful attention to their influence on the markets for nongovernmental securities. The treasuries must carefully study interest rates, yield patterns, terms of financing, and the distribution of holdings among investors.

Local governments are usually subject to various statutory restrictions that must be carefully observed when offering a new issue for sale. Local government bonds are distributed through investment bankers who buy them and reoffer them to the public at higher prices and correspondingly lower yields. Sometimes the terms of the offer are negotiated. In the United States, however, a more prevalent means of selling state and local bonds is through competitive bidding, in which the issuer announces a contemplated offering of bonds for a designated amount, with specified maturity dates, and for certain purposes. Syndicates of investment bankers are formed to bid on the issue, and the award is made to the group providing the most favourable terms. The winning syndicate then reoffers the bonds to the public at prices carefully tailored to be competitive with comparable obligations already on the market and to provide a suitable profit margin.

Marketing corporation securities

The financial manager of a company requiring additional funds has a number of alternative courses of action open to him. He may do all of his financing through commercial banks by means of loans and revolving credit arrangements that, in essence, are formalized lines of credit. Or, he may prefer to raise capital through the sale of securities. If he chooses to do the latter, he may undertake a private sale with an institutional investor such as an insurance company, permitting him to avoid both the complicated procedures of a public distribution and the risks of unsettled market conditions. On the other hand, a private placement of this sort deprives the issuer of the favourable publicity flowing from a successful public offering; it may not afford sufficient resources for very large firms with continuing demands for capital; and it involves rather restrictive legal requirements.

A company that elects to float its securities publicly in the capital market will ordinarily utilize the services of an investment banker. The investment banker may buy the securities from the issuer and seek a profit by selling them at a higher price to the public, thereby assuming the market risks. If the issue is large, the originating investment banker may invite other houses to join with him in purchasing the issue from the company, while to facilitate disposal he may form a selling group to take over the issue from the buying firms for resale to the public. In lieu of buying the securities from the issuer, the investment banker may act as an agent and receive a commission on the amount sold. If the issuer negotiates the selection of an investment banker, the banker will serve as financial counsel and offer advice on the timing and terms of the new issue. If the selection is by competitive bidding, the relationship is likely to be more impersonal.

An accepted principle of modern finance is that investors are entitled to knowledge about the issuer in order to appraise the quality of the securities offered. A number of countries now require issuers to file registration statements and provide written prospectuses.

The security markets of Europe do not have the aggressive investment-banking machinery developed in the United States. European commercial banks, on the other hand, play a much more important role in financing the needs of industry than do commercial banks in the United States and Great Britain.

In the 1960s, a number of industrial nations faced increasing difficulties in meeting their financing needs through local capital markets. Several issuers began to float securities that were payable in any of 17 different European currencies. This marked what might be called the beginning of an "international securities" market. Efforts were also made to issue bonds on a parallel basis in different countries with each portion denominated in the currency of the country in which it was sold. For various legal and technical reasons, these methods did not attract a wide following.

Another factor that hastened the growth of a European securities market was the balance of payments problem confronting the United States in the 1960s. Certain legislative enactments substantially shut the capital market of the U.S. to foreign issuers; and other restraints were imposed on foreign lending by United States financial institutions and on direct foreign investment by United States corporations. As a result, a number of multinational corporations headquartered in the United States were forced to seek financing in overseas securities markets for the expanding business of their foreign subsidiaries. United States and foreign investment bankers joined in syndicates to float these securities. The process was facilitated by the growth of an international market in Eurodollars, representing claims on dollars deposited in European banks. The bulk of the new bonds offered abroad were denominated in Eurodollars.

Eurodollars market

During the period 1957–65, when this new European market came into being, the volume of foreign bonds issued publicly rose from $492,600,000 to $1,489,500,000. The principal and most consistent borrowers were in Canada, Australia, Japan, Norway, Israel, Denmark, and New Zealand. In all of these countries, the major borrower was the national government, except in Canada, where the political subdivisions were the major borrowers. In West Germany, Great Britain, and the United States, the only borrowers in international markets were private units.

The machinery of securities trading. Organized securities markets and stock exchanges are a product of economic development. In the early years of economic growth, most of a country's industrial units are small and their capital requirements relatively modest. The rate of saving is low, and institutions for channelling private savings into investment are generally lacking. As the economy progresses and national income grows, new institutions enter the financial picture to direct the mounting volume of savings into productive outlets. The appearance of growing numbers of individual and institutional investors creates a need for trading markets to speed up transactions and enable stockholders swiftly and easily to convert their holdings to cash.

At this stage of development, corporations usually meet less of their financing needs through direct sales of securities in the new issue market and obtain a larger percentage through reinvesting their own earnings. This plowing back of earnings is not insensitive to the judgment of investors: if the prospects of a company are good, investors bid up the price of its shares in the trading market and show a willingness to forego dividends for the possibility of long-term capital gains achieved through internal growth. Thus, when a company is able to finance its expansion by means of reinvested earnings rather than by new stock issues, the trading segment becomes the more important aspect of the capital market.

The development of stock exchanges

History. Stock exchanges grew out of early trading activities in agricultural and other commodities. Traders in European fairs in the Middle Ages found it convenient to use credit, which required the supporting documents of drafts, notes, and bills of exchange. The French stock exchange may be traced as far back as the 12th century, when trading occurred in commercial bills of exchange. To regulate these incipient markets, Philip the Fair (1268–1314) created the profession of *courratier de change,* the forerunner of the modern French stockbroker, or *agent de change.* At about the same period in Bruges, then a prosperous centre of the Low Countries, merchants took to gathering in front of the house of the Van der Buerse family to engage in trading. From this custom, the name of the family became identified with trading, and eventually "bourse" came to signify a stock exchange. From similar roots in trade and commerce, the institutional beginnings of stock exchanges appeared during the 16th and 17th centuries in other great trading centres throughout the world—Amsterdam, Great Britain, Denmark, Germany.

The growth of trade created a need for banks and insurance companies. Political developments caused governments to seek new sources of funds. This combination of expanding activity and intermittent capital shortages stimulated the early issuers of securities—governments, banks, insurance companies, and some joint-stock enterprises, particularly the great trading companies. From the existing exchanges for commercial bills and notes, it was an easy and logical transition to the establishment of stock exchanges for securities. By the early 1600s, shares of the Dutch East India Company were being traded in Amsterdam; in 1773, London stock dealers who had previously been meeting in coffeehouses moved into their own building; and by the 19th century, trading in securities on a formal basis was common in the industrialized nations.

The evolution of stock exchanges continued. In Great Britain, progress has for the most part been internal and voluntary; the London Stock Exchange has regulated its own activities. The French stock exchanges, in contrast, are directly subject to law, and the operations of the *agents de change* have been affected by national decrees. At one time, there were three markets for securities in Paris: an official market called the Parquet (the "floor"); a semiofficial market, the Coulisse (the "wing"); and the Hors Côté (the "outside"), an unregulated market in unlisted securities. In 1929, the Hors Côté was subjected to official regulation and in the following decade its activities were absorbed into the Coulisse, which in turn was combined with the Parquet in a reorganization in 1961. In Belgium the exchanges have had a mixed history. Strict governmental controls were imposed in 1801 and not removed until 1867. Following the economic crisis of 1929–34, the pendulum swung the other way, and the exchanges were once more placed under the control of central authority. In Switzerland, the exchanges have been governed by cantonal (state) law.

Historical events have left their mark upon the development of stock exchanges in some countries. Mining, rather than trade and commerce, was the impelling influence in the establishment of stock exchanges in South Africa and Canada. In Germany, the Berlin Stock Exchange lost its dominant role after World War II, and its position was assumed by exchanges in Frankfurt and Düsseldorf. The Japanese securities markets were revolutionized following World War II, when a new securities law was enacted patterned after the U.S. model. A campaign to distribute stock formerly held by the large *zaibatsu* (family-owned combines) and semigovernment corporations greatly increased public stock ownership, which in turn contributed to the considerable growth of trading on the nine Japanese exchanges. Another post-World War II development was the interest of the governments of developing countries in the use of stock exchanges to facilitate external financing.

Securities markets in the United States began with speculative trading in issues of the new government. In 1791, the country's first stock exchange was established in Philadelphia, then the leading city in domestic and foreign trade. An exchange in New York was set up in 1792, when 24 merchants and brokers decided to charge commissions while acting as agents for other persons and to give preference to each other in their negotiations. They did much of their trading under a tree at 68 Wall Street. Government securities formed the basis of the early trading. Stocks of banks and insurance companies added to the volume of transactions. The building of roads and canals brought more securities to the market. In 1817 the New York brokers decided to organize formally as the New York Stock and Exchange Board. Thereafter, the stock market grew with the industrialization of the country. In 1863, the New York Stock Exchange adopted its present name. During the Civil War additional exchanges were organized, one of them the forerunner of the present American Stock Exchange, the second largest stock market in the country.

New York Stock Exchange

Organization. All stock exchanges perform similar functions with respect to the listing, trading, and clearing of securities. They differ in their administrative machinery for handling these functions.

The London Stock Exchange, the largest in the world in terms of the number and variety of domestic and international securities traded, is an independent institution not subject to governmental regulation. It resembles a private club with its own constitution and operating rules, administered by a council that, except for the government broker who is an ex officio nonvoting member, is elected by the membership. Operating responsibility is vested in the secretary and his staff.

In the United States, as in Great Britain, the government does not participate directly in the operations of the exchanges. Since 1933, however, Congress has enacted half a dozen measures that in one way or another affect the securities market. The most important are the Securities Act of 1933, which is primarily concerned with the new-issue market, and the Securities Exchange Act of 1934, which is directed toward the trading market. The latter requires that every stock exchange register with the Securities and Exchange Commission (SEC) as a national securities exchange, unless it is exempted because of the limited volume of its transactions, and that it conform to certain rules in its trading practices. This relationship between the stock exchanges and the SEC is peculiar to the United States; it involves a sort of administrative partnership between the exchanges as private associations—but functioning as quasi-public institutions—and the government. The exchanges generally may adopt policies and issue regulations governing their own operations but are subject to SEC intercession in the event that the Commission believes modifications are required in the public interest.

Most European exchanges are also subject to some form of governmental regulation. The Amsterdam Stock Exchange is a private organization, relatively free to regulate the activities of the market, but the Minister of Finance exercises some supervision under existing legislation. The Zürich exchange is governed by a board of elected members that determines general policy and coordinates the work of the exchange's committees. Of these, the Zürich Cantonal Committee, which directs dealings on the floor of the exchange, is chaired by the head of the Finance Department of the State of Zürich. Although the Frankfurt Stock Exchange is under the direct administration of a Board of Governors elected by its own members, the rules of the exchange must be approved by the authorities of the state of Hesse; and the official specialists, each responsible for the trading in certain securities, are appointed by the Minister of Finance in the state of Hesse. In Brussels the Ministry of Finance is involved in the appointment of members to major committees, while a governmental representative is attached to each exchange to supervise the observance of all rules and laws; in addition, the Banking Commission that is nominated by the government has substantial powers over the admission of securities to public trading. The policy-making Exchange Commission of the Paris bourse is headed by the Governor of the Banque de France, while the regular members are chosen by the Ministry of Finance; the *agents de change* who supervise the trading process are semigovernmental officials. New members of the Italian stock exchanges are appointed by the government from a list of candidates as a result of competitive examinations and therefore have some public status.

Membership requirements of stock exchanges

Since 1953, membership in the New York Stock Exchange has been limited to 1,366. Only individuals may be members, but they may be partners or stockholders of organizations that do business with the public. Their organizations in such cases are known as member firms. The exchange supervises and regulates member firms in what it considers to be the public interest: a majority of the owners must be engaged primarily in the business of brokers or dealers in securities; exchange approval is required of any shareholder with a 5 percent interest in a member corporation; and all principal officers and directors who are active in the firm must be members or allied members of the exchange. An allied member is subject to the rules of the exchange but does not have the right to engage in transactions on the floor.

To become a member, an individual must acquire, with the approval of the board of governors, a "seat" from a present member or from the estate of a deceased one. Before granting approval, the exchange will investigate such matters as the applicant's past record and financial standing; he will also have to pass an examination demonstrating his knowledge of the securities field.

There are several kinds of brokers on the floor of the exchange. They include the commission broker who executes customer orders placed at or near the current market price; the specialist in one or more issues who, as a broker, executes limited orders for other members and, as a dealer, buys and sells securities for his own account; floor brokers, or "two-dollar" brokers, who execute orders for other brokers at a commission but have no contact with the public; brokers associated with odd-lot firms, who undertake to buy or sell in quantities other than the standard 100-share lot; and "registered traders" who buy and sell for their own account.

To become a member of the London Stock Exchange, an individual must acquire a "nomination" from a retiring member at a price that varies with demand and supply. Every applicant must be approved by at least three-quarters of the stock exchange council. A member may be a broker, dealing as an agent of the public, or a jobber, dealing for his own account with other brokers or jobbers.

Membership in the Paris stock exchange is limited to 85 *agents de change* who supervise activity while the actual work of trading and executing orders is done by their employees and those of the exchange. To become an *agent de change,* an applicant must meet prescribed standards of education and experience as well as pass a written examination. He must be nominated by a retiring member or the heirs of a deceased member and make a deposit guaranty. He is formally appointed by the Minister of Finance.

Other European exchanges set eligibility requirements of character, experience, and financial standing, and some have educational requirements as well. In Brussels, in addition to the completion of six consecutive years in a broker's office, a candidate must have a degree in commercial science or economics and pass a professional examination. In Germany, Switzerland, and Sweden, the brokerage business is dominated by banks.

Members of Japanese exchanges must be corporations doing a business in securities. There are two kinds of members: regular members, who buy and sell for customers or for their own accounts, and *saitori,* who act principally as intermediaries for regular members.

Trading procedures. Most stock exchanges are auction markets, in which prices are determined by competitive bidding. In very large, active markets, the auction is continuous, occurring throughout the day's trading session and for any security in which there is buying and selling interest. In smaller markets the names of the listed stocks may be submitted in some form of rotation, with the auction occurring at that time; this process is described as a "call market."

Trading methods on all the exchanges in the United States are similar. In a typical transaction for a security listed on the New York Stock Exchange, a customer gives an order to an employee in a branch or correspondent office of a member firm, who transmits it either indirectly through the firm's New York office or, as is becoming increasingly common, directly to a receiving clerk on the floor of the exchange. The receiving clerk summons the firm's floor broker, who takes the order, goes to the post where the stock is traded, and participates in an auction procedure as either buyer or seller. If the order is not a market order calling for immediate action, the broker turns it over to an appropriate specialist who will execute it when an indicated price is reached.

Floor broker

As in any auction market, securities are sold to the broker bidding the highest price and bought from the broker offering the lowest price. Since the market is continuous, buyers and sellers are constantly competing with each other. In the New York Stock Exchange, the specialist plays an important role. As a principal, he has the responsibility of buying and selling for his own account, thereby providing a stabilizing influence; as an agent, he represents other brokers on both sides of the market when they have orders at prices that cannot be readily executed.

With the growing demand for stocks on the part of institutions such as insurance companies, mutual funds, pension funds, and so forth, the size of orders consummated on the New York Stock Exchange has grown. The common way of handling these big blocks on the floor of the exchange has been to break them into smaller orders executed over a period of time. Another method is to assemble matching orders in advance and then "cross" them, executing the purchase or sale at current prices in accordance with prescribed rules; since the broker initially may have obtained the matching orders off the floor, this procedure assumes some of the aspects of a negotiated rather than a pure auction transaction. It is only a step from this to so-called block positioning, in which the broker functions as a principal and actually buys the block from the seller and distributes the securities over a period of time on the floor of the exchange.

When none of these methods appears feasible, the exchange permits certain special procedures. A *secondary distribution* of stock resembles the underwriting of a new issue, the block being handled by a selling group or syndicate off the floor after trading hours, at a price regulated by the exchange. In an *exchange distribution* a member firm accumulates the necessary buy orders and then crosses them on the floor. This is distinguished from an ordinary "cross" because the selling broker may provide extra compensation to his own registered representatives and to other participating firms. A *special offering* is the offering of a block through the facilities of the exchange at a price not in excess of the last sale or the current offer, whichever is lower, but not below the current bid unless special permission is obtained. The terms of the offer are flashed on the tape. The offerer agrees to pay a special commission. A *specialist block purchase* permits the specialist to buy a block outside the regular market procedure, at a price that is somewhat below the current bid.

Brokers and jobbers

Trading on the London Stock Exchange is carried on through a unique system of brokers and jobbers. A broker acts as an agent for his customers; a jobber, or dealer, transacts business on the floor of the exchange but does not deal with the public. A customer gives an order to a brokerage house, which relays it to the floor for execution. The receiving broker goes to the area where the security is traded and seeks a jobber stationed in the vicinity who specializes in the particular issue. The jobber serves only in the capacity of a principal, buying and selling for his own account and dealing only with brokers or other jobbers. The broker asks the jobber's current prices without revealing whether he is interested in buying or selling. The broker may seek to narrow the spread between the bid and ask quotations or he may approach another jobber handling the same issue and undertake the same bargaining process. Eventually, when satisfied that he has obtained the best possible price for his client, the broker will complete the bargain.

A broker is compensated by the commission received from the customer. The jobber seeks to maximize his profitable business by adjusting his buying and selling prices. As the ultimate dealer in the London market, the jobber's activities provide a stabilizing factor, but unlike the specialist on the New York Stock Exchange, the jobber is under no obligation to help support prices. The grow-

ing importance of institutional customers has increased the size of transactions in the London market as it has in the U.S., and therefore the jobber has been compelled to risk larger sums. To offset this risk, arrangements for a particularly large order may be negotiated beforehand and the transaction put through the floor as a matter of procedure, with the jobber accepting a minimum "turn." Although the jobbing system provides a continuous market, it does not employ the auction bidding of the New York exchange.

The trading procedures of other major exchanges throughout the world employ the principles that have been described above, although they vary in their application of them. In the exchanges of Paris, Brussels, Copenhagen, Stockholm, and Zürich, some form of auction system is employed: prices are established through bids and offers made on specific securities at particular periods of time. In Tokyo, trading is continuous and orders are consummated through the *saitori* members, who keep order sheets on all transactions. Unlike the specialist on the New York exchange or the jobber in London, however, the *saitori* does not trade for his own account but serves only as an intermediary between regular members. In Amsterdam, trading in active securities is done directly between members during designated trading periods; specialists function as intermediaries between buyers and sellers.

Types of orders. The simplest method of buying stock is through the *market order.* This is an order to buy or sell a stated amount of a security at the most advantageous price obtainable after the order reaches the trading floor. A *limit* (or *limited*) *order* is an order to buy or sell a stated amount of a security when it reaches a specified price or a better one if it is obtainable after the order comes to the trading floor. In the Amsterdam market, the device of the "middle price" is used: an investor who gives a limit order before the opening will have it executed at the day's median level, or at a price that is better than the limit, whichever is found to be more advantageous to the client.

There are other more specialized types of orders. A *stop order* or *stop-loss order* is an order to purchase or sell a security after a designated price is reached or passed, when it then becomes a market order. It differs from the limit order in that it is designed to protect the customer from market reversals; the stop price is not necessarily the price at which the order will be executed, particularly if the market is changing rapidly. This type of order does not lend itself to the London jobbing system.

An important method of trading in stock is through the buying and selling of options. The most common option contracts are puts and calls. *A put* is a contract that permits the holder to deliver to the purchaser a specified number of shares of stock at a fixed price within a designated period of time, say six months; a *call* entitles him to buy shares from the seller within a given period. For example, a person who buys a stock hoping to sell it later at a higher price may also buy a put as a hedge against a fall in price. The put enables him to sell the stock at the price for which he bought it. If the stock rises he need not use the option and loses only the price of its purchase. Option trading is common in Brussels, Paris, London, and the United States.

Unorganized, or over-the-counter, markets

In the early days of securities trading, stocks and bonds were often bought at private banking houses in the same way that commodities might be purchased over the counter of a general store. This was the origin of the term "over-the-counter." It is used today to mean all securities transactions that are handled outside the exchanges. Increasingly, this market is being subjected to regulation. The extent and nature of the over-the-counter market varies throughout the world. In the United Kingdom, there is no over-the-counter market as such. In the Netherlands, transactions are illegal if they do not involve a member of the Amsterdam exchange or one of its provincial branches as an intermediary, except with the permission of the Ministry of Finance. On the Paris bourse, one post is provided for trading in unlisted issues. In Belgium, the stock exchange committee organizes, at least once a month, public sales of stocks that are not officially quoted. In Japan a second security section has been introduced into the major exchanges to provide more effective trading procedures for over-the-counter transactions.

In the United States, the over-the-counter market includes most federal, state, and municipal issues as well as a large variety of corporate stocks and bonds. The National Quotation Bureau, which compiles over-the-counter prices, has furnished quotations on approximately 26,000 over-the-counter stocks. In early 1971, a major development occurred with the introduction of current, computerized quotations on a number of active stocks.

Transactions in the over-the-counter market are executed through a large number of broker-dealers with a complex network of private wires and telephone lines. Their operations are subject to the rules of the National Association of Securities Dealers, Inc., a self-regulating body created in 1939. In 1964 the Congress extended to the larger over-the-counter companies the same requirements as to periodic reporting, proxy solicitation, and insider trading that are applied to dealers in listed stocks.

The over-the-counter market is a negotiated market, as distinguished from the auction markets for listed securities. An investor desiring to trade an over-the-counter security gives his order to a broker functioning as a retailer, who ordinarily shops among various firms to obtain the best possible price.

Because of the difficulty that institutions often experience in disposing of large blocks of listed securities on the exchanges, nonmember firms have set up over-the-counter markets in these issues—principally in those listed on the New York Stock Exchange. Although such transactions are conducted within the framework of the over-the-counter market, their prices are tied to those on the Exchange. Accordingly, this form of trading has been labelled the "third market." There is now also a "fourth market," consisting of direct transactions between investors without an intermediary. This market also had its origins in the need of the institutions to find ways of executing large transactions. Impetus to such direct dealings has been given by the development of computerized systems to bring together large traders.

The structure of demand for securities. Interest in the ownership of securities has increased greatly in recent decades. Inflationary tendencies have directed attention to stocks as a means of offsetting rising prices. Stock exchanges have cultivated investors through public relations programs. Government regulation has strengthened public confidence in stock trading procedures. Many governments have given support to the capital markets in order to facilitate business financing.

The distribution of stock ownership

In the United States, in addition to the millions of individuals who own shares of publicly held corporations, many others own shares indirectly through institutions that are large holders of stock, such as investment companies and pension funds. It is difficult to obtain figures for other countries. A major problem of developing countries has been the absence of investors able and willing to buy shares. Many investors in these countries have preferred to place their funds in tangible assets such as land.

Institutions such as insurance companies, mutual funds, pension funds, foundations, and universities have grown very important in the security markets of the United States. Because of their financial responsibilities to others, these institutions have characteristically followed conservative investment policies stressing the purchase of fixed-income securities. The long-term trend toward inflation, along with mounting stock prices, has led the institutions to look more favourably upon common stocks.

Among the most rapidly growing institutions are the mutual funds. Technically, these are known as open-end investment companies because the number of their shares outstanding constantly changes as new shares are sold to investors and old ones redeemed.

Dynamics of the securities markets. Over a long-term period, the movements of stock prices and of general business indicators tend to parallel each other. In studies of business cycles, it has been found that stock prices tend to reach their cyclical peaks and troughs somewhat ahead of general business indicators, and these are therefore generally classified as "leading indicators."

Long-term and short-term movements of stock prices

The price of a stock reflects the present value of expected future earnings, and the profits of a firm are strongly influenced by the general level of economic activity. The tendency of stocks to lead business may be attributable to investors' preoccupation with the future. Over the years, the trend of stock prices has been upward; since World War II the upward cycles has tended to be of longer duration, while declines have been relatively shorter. Within the generally expansionary movement, changes in share values of specific companies have been mixed, some showing striking long-term gains while others have suffered losses.

Dramatic events sometimes have a special influence on the psychology of investors, driving stock prices down despite improving business conditions. For example, between the fall of 1940 and the spring of 1942, the period immediately prior to and after the entry of the United States into World War II, U.S. stock prices dropped swiftly despite a continued revival of economic activity. In other cases the reasons for a fall in stock prices are not easy to discover.

Stock prices also experience daily changes of substantial size. Technical market analysts attempt to predict these changes by studying patterns in stock prices. Many theorists, however, claim that in a highly competitive market, prices fluctuate primarily as a result of new information that is not likely to appear in any organized fashion; they maintain that successive price movements are independent of each other and take place in a random fashion.

As investors' expectations change over time, their attitudes toward different types of stock change. Buoyant investors lean toward growth stocks, the value of which is expected to increase rapidly; when uncertainty prevails, the preference is for more conservative issues with stable records of earnings. Within any given period, investors' choices of particular stocks vary with their judgments of the related companies.

(S.M.R.)

MONEY MARKET

The functions of a money market

Every country with a monetary system of its own has to have some kind of market in which dealers in bills, notes, and other forms of short-term credit can buy and sell. The "money market" is a set of institutions or arrangements for handling what might be called wholesale transactions in money and short-term credit. The need for such facilities arises in much the same way that a similar need does in connection with the distribution of any of the products of a diversified economy to their final users at the retail level. If the retailer is to provide reasonably adequate service to his customers, he must have active contacts with others who specialize in making or handling bulk quantities of whatever is his stock-in-trade. The money market is made up of specialized facilities of exactly this kind. It exists for the purpose of improving the ability of the retailers of financial services—commercial banks, savings institutions, investment houses, lending agencies, and even governments—to do their job. It has little if any contact with the individuals or firms who maintain accounts with these various retailers or purchase their securities or borrow from them.

The elemental functions of a money market must be performed in any kind of modern economy, even one that is largely planned or socialist, but the arrangements in socialist countries do not ordinarily take the form of a market. Money markets exist in countries that use market processes rather than planned allocations to distribute most of their primary resources among alternative uses. The general distinguishing feature of a money market is that it relies upon open competition among those who are bulk suppliers of funds at any particular time and among those seeking bulk funds, to work out the best practicable distribution of the existing total volume of such funds.

In their market transactions, those with bulk supplies of funds or demands for them, rely on groups of intermediaries who act as brokers or dealers. The characteristics of these middlemen, the services they perform, and their relationship to other parts of the financial mechanism vary widely from country to country. In many countries there is no single meeting place where the middlemen get together, yet in most countries the contacts among all participants are sufficiently open and free to assure each supplier or user of funds that he will get or pay a price that fairly reflects all of the influences (including his own) that are currently affecting the whole supply and the whole demand. In nearly all cases, moreover, the unifying force of competition is reflected at any given moment in a common price (that is, rate of interest) for similar transactions. Continuous fluctuations in the money market rates of interest result from changes in the pressure of available supplies of funds upon the market and in the pull of current demands upon the market.

Banks and the money market. *Commercial banks.* Commercial banks are at the centre of most money markets, as both suppliers and users of funds, and in many markets a few large commercial banks serve also as middlemen. These banks have a unique place because it is their role to furnish an important part of the money supply. In some countries they do this by issuing their own notes, which circulate as part of the hand-to-hand currency. More often, however, it is checking accounts at commercial banks that constitute the major part of the country's money supply. In either case, the outstanding supply of bank money is in continual circulation, and any given bank may at any time have more funds coming in than going out, while at another time the outflow may be the larger. It is through the facilities of the money market that these net excesses and shortages are redistributed, so that the banking system as a whole can at all times provide the means of payment required for carrying on each country's business.

In the course of issuing money the commercial banks also actually create it by expanding their deposits, but they are not at liberty to create all that they may wish whenever they wish, for the total is limited by the volume of bank reserves and by the prevailing ratio between these reserves and bank deposits—a ratio that is set by law, regulation, or custom. The volume of reserves is controlled and varied by the central bank (such as the Bank of England, the Bank of France, or the Federal Reserve System in the U.S.), which is usually a governmental institution, is always charged with governmental duties, and almost invariably carries out a major part of its operations in the money market.

Volume of reserves

Central banks. The reserves of the commercial banks, which are continually being redistributed through the facilities of the money market, are in fact mainly deposit balances that these commercial banks have on the books of the central bank or notes issued by the central bank, which the commercial banks keep in their own vaults. As the central bank acquires additional assets, it pays for them by crediting depositors' accounts or by issuing its own notes; thus the potential volume of commercial bank reserves is enlarged. With more reserves, the commercial banks can make additional loans or investments, paying for them by entering credits to depositors' accounts on their books. And in that way the money supply is increased. It may be reduced by reversing the sequence. The central bank can sell some of its marketable assets in the money market or in markets closely interrelated with the money market; payment will be made by drawing down some of the commercial bank reserve balances on its books; and with smaller reserves remaining, the commercial banks will have to sell or reduce some of their investments or their loans. That, in turn, results in a shrinkage of the outstanding money supply. Central bank operations of this kind are called open-market operations.

The central bank may also increase bank reserves by making loans to the banks or to such intermediaries as bill dealers or dealers in government securities. Reduction of these loans correspondingly reduces bank reserves. Although the mechanics of these lending procedures vary widely among countries, all have one feature in common: the central bank establishes an interest rate for such borrowing—the bank rate or discount rate—pivotally significant in the structure of money market rates.

Money market assets may range from those with the highest form of liquidity—deposits at the central bank—through bank deposits to various forms of short-term paper such as treasury bills, dealers' loans, bankers' accep-

tances, and commercial paper, and including government securities of longer maturity and other kinds of credit instruments eligible for advances or rediscount at the central bank. Although details vary among countries, the touchstone of any money market asset other than money itself is its closeness—*i.e.*, the degree of its substitutability for money. So long as the institutions making use of a money market regard a particular type of credit instrument as a reasonably close substitute—that is, treat it as "liquid"—and so long as the central bank acquiesces in or approves of this approach, the instrument is in practice a money market asset. Thus no single definition or list can apply to the money markets of all countries nor will the list remain the same through the years in the money market of any given country.

The international money market. Each central bank usually holds some form of reserve that is acceptable in settling international transactions. International monetary reserves are mainly gold, or "money market assets" in some country whose currency is widely used, such as the United States dollar. The monetary laws of all countries provide for the establishment of some kind of parity between their currencies and those of other countries. This parity may be defined either in terms of gold or in relation to a key currency such as the British pound sterling or the United States dollar, which in turn has a fixed parity with gold. A country maintains the "convertibility" of its currency by standing ready to buy and sell gold or other currencies in exchange for its own at prices within a fixed and rather narrow "spread" above or below the "exchange rate" for its own currency that is implied by the declared parity.

The buying and selling of currencies

Because world trade continually gives rise to various needs for payment in various currencies, an international money market must exist so that traders with an excess of one currency can use it to buy another currency for which they have a need. Within the scope of convertibility arrangements, this trading in currencies is carried out by skilled intermediaries, usually banks or specialized foreign exchange brokers and dealers. Trading in currencies is extensive both for immediate use ("spot") and for future ("forward") delivery. Quotations vary according to changes in supply and demand, over the range between the upper and lower buying and selling prices set by official parity. If no parity has been set quotations may fluctuate widely. If a currency is subject to exchange controls, there may be two or more quotations for different uses of the same nominal currency.

Changes in a country's balance of payments may affect the usefulness or prestige of its currency. A sustained and substantial balance of payments deficit (outpayments larger than inpayments), for example, will result in continuous large increases in the world supply of its currency, possibly leading to some decline in its acceptability abroad and to a loss of international monetary reserves. At the same time, an outward drain may reduce the reserves of the commercial banks (the base for the domestic money supply), unless the central bank takes offsetting action.

Since 1944 most of the countries that have domestic money markets or that play a role in the international money market have been joined together in the International Monetary Fund, which represents a pooling of part of the foreign exchange reserves (including gold) of more than 100 member countries. Drawings on the pool may be made by member countries to meet some of the reserve drains arising from balance of payments deficits and in amounts related to the quota that each has subscribed.

The internal money markets of a surprisingly high proportion of the countries of the world are quite rudimentary. The work of the money market in these countries is done largely by transfers of deposit balances, government securities, or foreign exchange among a few banks and between them and the central bank. But in nearly all such cases there is genuine discontent with the rigidity of these limited facilities and a desire to develop a structure, as well as instruments and procedures, which would provide the open-market attributes of the arrangements that have evolved in the leading countries. Several of the more fully developed money markets are described below.

The U.S. money market. The domestic money market in the United States carries out the largest volume of transactions of any such market in the world; its participants include the most heterogeneous group of financial and nonfinancial concerns to be found in any money market; it permits trading in an unusually wide variety of money substitutes; and it is less centralized geographically than the money market of any other country. Although there has always been a clustering of money market activities in New York City and much of the country's participation in the international money market centres there, a process of continuous change during the 20th century has produced a genuinely national money market.

The Federal Reserve System

By 1935 the financial crises of the Great Depression had resulted in a basic revision of the banking laws. All gold had been withdrawn from internal circulation in 1933 and was henceforth held by the U.S. Treasury for use only in settling net flows of international payments among governments or central banks; its price was raised to $35 per ounce, and the U.S. dollar became the key currency in an international gold bullion standard. Domestically, the changes included legislative recognition of the primary importance of unified open-market operations by the Federal Reserve System and delegation to the board of governors of the Federal Reserve System of authority to raise or lower the ratios required between reserves and commercial bank deposits. Although about half of the 30,000 separate banks existing in the early 1920s had disappeared by the mid-1930s, the essential character of commercial banking in the U.S. remained that of a "unit" (or single-outlet) banking system in contrast to those of most other countries, which had a small number of large branch-banking organizations.

The unit banking system. This system has led inevitably to striking differences between money market arrangements in the United States and those of other countries. At times, some smaller banks almost inevitably find that the wholesale facilities of the money market cannot provide promptly the funds needed to meet unexpected reserve drains, as deposits move about the country from one bank to another. To provide temporary relief, pending a return flow of funds or more gradual disposal of other liquid assets in the money market, such banks have the privilege, if they are members of the Federal Reserve System, of borrowing for reasonable periods at their own Federal Reserve bank. At times some large banks, which serve as depositories for part of the liquid balances of many of the smaller ones (including those that are not members of the Federal Reserve System) also find that demands converging on them are much greater than expected. These large banks, too, can borrow temporarily at a Federal Reserve bank if other money market facilities are not adequate to their needs. Because these borrowing needs are unavoidably frequent in a vast unit banking system and, as a rule, do not indicate poor management, the discount rate charged by the Federal Reserve banks on such borrowing is not ordinarily put at punitive or severe penalty levels—thus, contrary to practice in many other countries, the central bank does not always maintain its interest rate well above those prevailing on marketable money market instruments. To avoid abuse, there is continuous surveillance of the borrowing banks by the Federal Reserve banks.

Along with this practice of borrowing at a Federal Reserve bank has developed the market for "federal funds." This specialized part of the money market provides for the direct transfer to a member bank of balances on the books of a Federal Reserve bank in return for payment of a variable rate of interest called the "federal funds rate." These funds are immediately available. There are transactions, too, in funds that are on deposit at commercial banks—by means of loans between banks, or through loans by one large depositor to another. Because these must be collected through a clearing process, they are usually called "clearinghouse funds."

Money market instruments. Transactions in federal funds and clearinghouse funds are further supplemented by transactions in which either kind of money is exchanged for some other liquid, money market instrument, most frequently government securities. The magnitude of

the market for government securities became so great after World War II that it overshadowed all other elements of the money market. Trading in outstanding "governments" is virtually all done through dealers who buy and sell for their own account at prices which they quote on request (standing ready to "bid" for or to "offer" any outstanding issue). Most of these dealers have head offices located in New York City, but all are engaged in nationwide operations. Their transactions and the lending arrangements through which they finance their own inventories of government securities have evolved into a particularly sensitive indicator of the pressures of supply and demand on the money market from day to day. The most common form of dealer financing is the repurchase agreement, through which dealers sell parts of their inventory temporarily, subject to repurchase.

Closely interrelated, often through trading operations conducted by the same dealers, are the much smaller markets for bank drafts, bills of exchange, and commercial paper. Alongside these other markets and actually somewhat larger in outstanding volume are the markets for securities issued by various "agencies" created by federal statute, such as the Federal Home Loan banks and Federal Land banks. Another money market instrument is the negotiable time certificate of deposit (CD), issued in large volume by commercial banks, which first became significant in 1962. While the owner of a time CD cannot withdraw his deposit before the maturity date initially agreed upon, he can sell it at any time in a secondary market that is conducted by government securities dealers.

Certificates of deposit

The Federal Reserve System conducts day-to-day operations in the money market on its own initiative in order to assist the smooth working of the nation's financial machinery and to exert a general influence aimed at fostering economic growth and limiting economic instability. Its transactions include substantial outright purchases or sales of government securities, relatively small purchases and run-offs of bankers' acceptances, and a considerable volume of loans made for a few days at a time to dealers in government securities or acceptances in the form of repurchase agreements. While it is still the commercial banks as a group that have the greatest continuing need for the combined facilities of the nationwide money market, there is frequent and continuous participation by a great variety of institutional investors who channel the public's savings into various uses and who must always also make some provision for their own liquidity.

Perhaps the most unusual feature in the composition of the U.S. money market is the great importance attained by nonfinancial business concerns and local units of government since World War II. Corporate treasurers and the treasurers of many states and local political subdivisions and authorities have become so keenly sensitive to the profitable possibilities of managing their own liquid holdings instead of relying on the commercial banks as most had done formerly that this group at times provides nearly as large a part of the volatile financing needs of government securities dealers, for example, as comes from the banks. Moreover, banks outside New York City sometimes supply more of the financing needed by these dealers than do the traditional "money market banks" in New York City. The nationwide character of the money market is also shown by the participation of nearly 200 banks in the federal-funds market—banks that are widely scattered among all Federal Reserve districts, although the bulk of all transactions is executed through facilities located in New York.

While the U.S. money market has become truly national, it still needs a final clearing centre upon which the net impact of changes in overall supply or demand can ultimately converge and where the final balancing adjustments of the market as a whole can be accomplished. In filling that need, New York City continues to be the centre of the national money market.

The British money market. *The discount houses.* In Great Britain the money market consists of a number of linked markets, all of them concentrated in London. The 12 specialist banks known as discount houses have the longest history as money market institutions; they have their origin in the London bill broker who in the early 19th century made the market in inland commercial bills. By selling bills through this market, the growing industrial and urban areas were able to draw upon the surplus savings of the agricultural areas. Quite early many bill brokers began to borrow money from banks in order to buy and hold these bills, instead of simply acting as brokers, and thus became the first discount houses. Since then the major assets held by the discount houses have at different times been commercial bills (first inland bills as described above and later bills financing international trade), treasury bills, and short-dated government bonds. During the 1960s there was a resurgence of the commercial bill, which finally became the discount houses' largest single class of asset, only to be overtaken later by the certificate of deposit.

Important changes were introduced into the British monetary system in 1971, but money at call with the discount houses retained its role as a reserve asset. Such is the safety and liquidity of call money that, despite the fractionally lower rate on it compared with other reserve assets, the banks hold about half of their required reserves in this form. This in turn provides the discount houses with a large pool of funds, which they invest in relatively short-dated assets, of which the most important is sterling certificates of deposit, followed by commercial bills, local authority securities, and treasury bills. This pattern of assets is greatly influenced by the fact that all call loans to the discount houses are secured loans, parcels of assets being deposited *pro rata* with the lending banks as security, and the assets held by the discount houses must therefore be suitable for use as such security.

They also need to hold a substantial proportion of assets that are rediscountable at the Bank of England in case of need, and the Bank of England limits their holding of assets other than public sector debt to a maximum of 20 times their capital resources.

Operations in money

On the liabilities side of the discount house's balance sheet, operating in call money is part of its day-to-day work. A bank that expects to make net payments to other banks during the day (for example, in settlement of checks paid by its customers to their customers) will probably call in some of its call loans, and by convention this is done before noon. Since the banks that have called in money then pay it to other banks, these other banks will have an equal amount to re-lend to the discount houses in the afternoon. Thus the discount houses can "balance their books"—borrow enough money in the afternoon to replace the loans called from them in the morning.

It is not uncommon for perhaps £100,000,000 to be called from, and re-lent to, the discount houses on an active day.

There is one main reason why the money position may not balance in this way. The British government accounts are kept with the Bank of England, which does not lend at call as other banks do. Thus net payments into or out of these government accounts will cause a net shortage or surplus of money for the discount houses in the afternoon and will tend to cause money rates to rise or fall. The Bank of England can allow such shortages or surpluses to affect interest rates, or it can offset them by buying or selling bills or by lending overnight to the discount houses at market rates. Even if the Bank of England does not act in this way to meet a shortage of funds, the discount houses are always finally able to secure the funds they need by their right to borrow from the Bank of England (the lender of last resort) against approved security at the "minimum lending rate" (the penalty rate).

Dealing in assets

On the assets side of their balance sheets, the discount houses are active dealers in a number of the assets they hold. They make the market in sterling certificates of deposit and in commercial bills, quoting buying and selling rates for different maturities. They also quote selling rates for treasury bills that they acquire at the weekly tender in competition with each other and with any other banks that may tender, including the Bank of England. Most of these other banks tender for treasury bills in order to hold them to maturity, but the discount houses sell theirs on the average when only a few weeks of the bills' 91-day life has passed. A large proportion of these bills is sold to the

clearing banks, which do not tender on their own account. The Bank of England minimum lending rate is normally determined for each week 0.5–0.75 percent above the average treasury bill rate at the previous Friday's tender. The bank, however, has the power to fix it at a different level if it so wishes, and this has been done.

Other markets. Important changes have also occurred outside the discount market described above; after the mid-1950s there was steady growth in public borrowing by local authorities. This led to an active local authority loan market conducted through a number of brokers, where money can be lent on deposit for a range of maturities from two days up to a year (and indeed for longer periods). Much more rapid was the growth after the mid-1960s of the interbank market, in which banks borrow and lend unsecured for a range of maturities from overnight upward. This market also is conducted through brokers, often firms that also operate in the local authority and other markets; a number of these firms of brokers are subsidiaries of discount houses.

In addition to the markets mentioned, there is the gilt-edged (government bond) market on the stock exchange; short-dated bonds are held by the discount houses and by banks and other money market participants, as are short-dated local authority stocks and local authority "yearling" (very short-dated) bonds. With flexibility of bank deposit rates (at least for deposits of large denomination), both banks and nonbank transactors are faced with a wide and competitive range of sterling money market facilities in London.

Finally, mention should be made of the Eurodollar market, because London is its centre; this is an entrepôt market with a very large volume of business in U.S. dollar balances, conducted through brokers (often the same firms that operate in the sterling markets), and U.K. banks are active participants. However, owing to exchange control there has been little significant interaction between the Eurodollar market and the U.K. domestic money market.

The money markets of other countries. *The Canadian money market.* The Canadian money market was substantially broadened in 1954 with the introduction of day-to-day bank loans against Government of Canada treasury bills and other short-term government and government-guaranteed securities. Treasury bills of 91 days' and 182 days' maturity are issued weekly with the occasional offering of a longer maturity of up to one year. Government of Canada bonds and Government of Canada guaranteed bonds are issued at less regular intervals.

Government of Canada bonds

Groups involved in the money market are the following: the government, as the issuer of the securities; the Bank of Canada, acting as issuing agent for the government and as a large holder of market material; the chartered banks, as large holders and as distributors and potential buyers and sellers of bills and bonds at all times; the security dealers, as carriers of inventories and traders in such securities; and the public (mainly provincial and municipal governments and larger corporations), as short-term investors.

Treasury bills are sold by competitive tender in which the Bank of Canada, the chartered banks, and a small number of investment dealers participate. Bonds are normally issued at a price at which the yield is in line with outstanding comparable issues.

The central bank, through its tender at the weekly treasury-bill sale, active manipulation of its own bill and bond portfolio, and regulation of the money supply, has workable instruments for active monetary control. For both banks and qualified dealers, the Bank of Canada acts as lender of last resort. The rate is set slightly above the average rate of the last treasury bill auction to discourage regular borrowing.

The West German money market. In West Germany, where the money market developed strongly after World War II, transactions are to a large extent confined to interbank loans. In addition, insurance companies and other nonbank investors are also important lenders of short-term funds. Treasury bills and other short-term bills and notes from government agencies (railways and post) were gaining in importance by the 1960s, whereas in 1955 certain nonmarketable securities (the so-called equalization claims, created during the 1948 currency reform) held by the Bundesbank were transformed into short-term marketable securities in order to obtain suitable market material for the open-market operations of the Bundesbank. Banks are not used to dealing in short-term government securities between each other. They generally either hold these securities to maturity or resell them to the central bank at its buying rates, so that a true money market has not developed.

The market for commercial paper is of some significance, and dealing in it takes place from time to time between banks, especially in times of tight market conditions. Comprehensive regulations have been given through the Bundesbank about the rediscountability of the several kinds of commercial paper.

The influence of the Bundesbank on the monetary situation through open-market operations by the 1960s was greatly hampered by the vast liquidity of the banking system as a consequence of the persistence of Germany's favourable balance of payments situation.

The French money market. The French money market is fairly well established, but its size is restricted by the fact that in France currency still plays an important role in the money supply, whereas by regulations the nonfinancial private sector of the economy is excluded from dealing in the market. Banks as well as a few public or semipublic agencies working in the financial sphere and intermediaries—brokers and discount houses—constitute the market. Transactions take place in commercial paper and in treasury bills. The monetary authorities maintain a special bookkeeping system for all the treasury bills held by banks and other financial institutions, under which such bills are not represented by actual certificates but by entries in special accounts administered by the Banque de France for the treasury.

The central bank's open-market operations, which were normally limited to smoothing out disturbances in the local money market, have gained importance in recent years. Open-market transactions are effected to keep domestic money market rates in line with international rates, in an effort to prevent unwanted capital flows. The possibilities of the central bank's influencing the monetary situation through the money market are limited to the large government needs for short-term funds, no market for long-term government borrowing being established.

The Japanese money market. In Japan's rapidly growing economy the demand for funds, both short-term and long-term, has been persistently strong. Commercial banks and other financial institutions have therefore had an important role. The monetary authorities (the Ministry of Finance and the Bank of Japan) have been unwilling to allow market forces to equilibrate demand and supply in many financial markets for fear that interest rates would become excessively high. Most interest rates have been set administratively at levels high by international comparison (until the late 1960s) but lower than market forces would have dictated. Monetary policy is implemented by controls on both the availability of credit and its cost.

Under these circumstances Japan has had a very restricted money market. The market for short-term government securities is negligible; the low, pegged interest rate means that the Bank of Japan is the main buyer and that open-market operations are impossible. Transactions in commercial paper are minimal, being discouraged because they would tend to undermine the structure of interest rates and financial institutions.

Call money market in Japan

Only the call money market is well developed. It is restricted to transactions among financial institutions. The interest rate on call money has been relatively free, and persistently above most other short-term and long-term rates. Although small amounts are lent overnight, most are "unconditional loans" (repayment after one day's notice, with a minimum of two days) or "over-month-end-loans" (repayment on a fixed day the following month). The pattern of flows is rather stable, despite seasonal and cyclical fluctuations. City banks are the major borrowers; they have a strong demand for loans by large enterprises and use call funds as a major source of liquidity. Major lenders are local banks, trust banks, credit associations,

and agricultural cooperatives, which collect individual urban and rural savings and are attracted by the high yields, liquidity, and low risk of call loans relative to other uses. Call brokers help make a market, though most funds flow directly from one financial institution to another. About three-quarters of the funds flow through the Tokyo market, and there are also call markets in Ōsaka and Nagoya.

Money markets in developing countries. Well-developed money markets exist in only a few high-income countries. In other countries money markets are narrow, poorly integrated, and in many cases virtually nonexistent. Despite the many differences among countries, one can say in general that the degree of development of a country's financial system, including its money markets, is directly related to the level of its economy. Most very-low-income countries have limited financial systems in which money markets play no role. In many former colonies, notably in Africa, expatriate commercial banks had substituted for a local money market; the banks met fluctuations in loan demand by changing their balances at head offices in London or elsewhere. More recently, government policies have encouraged these banks to develop domestic channels for temporary surpluses and deficits. Persistent inflation has been another factor inhibiting the growth of money markets in developing countries, notably in Latin America.

Most developing countries, except those having socialist systems, have the encouragement of money markets as a policy objective, if only to provide outlets for short-term government securities. At the same time many of these governments pursue low-interest-rate policies in order to reduce the cost of government debt and to encourage investment. Such policies discourage saving and make money market instruments unattractive. Nevertheless, a demand for short-term funds and a supply of them exist in all market-oriented economies. In many developing countries these pressures have led to "unorganized money markets," which are often highly developed in urban areas. Such markets are unorganized because they are outside "normal" financial institutions; they manage to escape government controls over interest rates; but at the same time they do not function very effectively because interest rates are high and contacts between localities and among borrowers and lenders are limited. In all developing countries traditional forms of moneylending continue, particularly for agriculture and small enterprise.

(H.T.P./R.V.R./R.F.G.A./C.G.)

The markets and social welfare

The branch of academic teaching called the "economics of welfare" is nowadays generally expounded by starting with a presumption in favour of laissez-faire and free competition and then following it by a list of reservations and exceptions that destroy its validity. One is first asked to imagine a particular group of individuals, each with his tastes for a predetermined list of specific commodities, with a predetermined endowment of labour power, with specified equipment and stocks, and with a given body of technical knowledge. These individuals can produce various alternative combinations of commodities. If resources were fully utilized, it would be impossible to produce more of any one commodity without producing less of others. One could then speak of an "opportunity cost" of each commodity—that is, the cost of producing a little more of it in terms of the amounts of other commodities that would have to be sacrificed.

Static balance of interests

On the other side of the market, the same individuals appear as consumers. If their tastes are all alike, it can be argued that there is one set of outputs and prices at which the relative subjective valuations are proportional to relative opportunity costs. This situation is the optimum in the sense that any move away from it would reduce the total satisfaction. The argument is not so simple when different consumers have different tastes. At any one pattern of prices established in the market (with full utilization of resources), each consumer can adjust his purchases so as to maximize his satisfaction at those prices. Then, starting from that set of prices, it can be shown that no one consumer could be made better off, by changing the composition of output, without making some other worse off. This is a powerful argument for the status quo, wherever it may happen to be. At every other point, there would be a different status quo, with a pattern of prices suiting some consumers better and some worse than the one that was chosen at the beginning of the argument.

The first general objection that is now admitted to this scheme (over and above some technical points within its own terms) it that is is purely static. Actual life is lived in the stream of time. In any economy, tastes, commodities, resources, and technology are continuously changing and modifying each other.

Second, there is a deep-seated inconsistency in the assumptions: when every individual pursues his own advantage, atomistic competition cannot persist, for any group of sellers or of buyers can gain by acting together and sharing the benefits among individual members of the group. Perfect competition, like free trade among nations, can persist only when there is some rule of behaviour that overrides pure self-interest.

Third, it is necessary to consider the distribution of consumption between individuals who make up the market. The purchasing power of each is limited by the receipts that he draws from his sales. His receipts depend on the amount of his original endowment and the price of the product he offers to sell. It is obvious that the distribution of the benefits of the market between individuals is quite arbitrary, and the market optimum is therefore not the same as their welfare optimum. When this objection to the laissez-faire principle is admitted, it is often suggested that inequalities might be corrected by a system of bounties and taxes; but this has never been taken seriously as a practical scheme.

A fourth set of objections comes under the heading of differences between private and social costs. The classic example is the smoke nuisance, which imposes costs upon the public that the factory concerned cannot be charged for. Examples of this phenomenon on a devastating scale are now daily coming to notice. At the same time there are many socially beneficial activities for which it is impossible to collect adequate payment from individuals in a society of unequal wealth and income—not only performances of grand opera but the whole of education and the health service.

Thus the doctrine that the free play of individual interests in a competitive market maximizes welfare for society as a whole has been demolished by its own exponents. Yet these notions still have an important influence on the formation of ideology. The rationalization grew out of the 19th-century utilitarian philosophers' hedonistic calculus, which regarded all human life as governed by the pursuit of pleasure and avoidance of pain. This train of thought identifies pleasure with consumption, leaving out of account the whole sphere of conditions in which work is carried on, and finally it sees consumption in terms of spending money. The whole economic basis of life is thus reduced to terms of commercial transactions. As the American critic Thorstein Veblen pointed out,

> so great and pervading a force has this habit of pecuniary accountancy become that it extends, often as a matter of course, to many facts which properly have no pecuniary bearing and no pecuniary magnitude, as, for example, works of art, science, scholarship, and religion.

THE POLITICS OF THE MARKET

The doctrines of laissez-faire are attractive in many ways. If the economy is a self-regulating mechanism and if economics is a system of scientific laws, then moral and political problems are excluded. Questions of social justice do not arise. The function of government is to be strictly neutral between interested parties. But when people come to recognize that the market, by its very nature, is necessarily a scene of conflicting interests, every element in it becomes a moral and political problem. This is distressing because one can no longer rely upon "principles of economics" to provide safe and simple rules for finding the correct solutions.

Laissez-faire as an ideal

The intrusion of politics into economic affairs increased dramatically in all the Western industrial nations after

1945. Capitalism took on a new shape, with certain characteristics that distinguish it from all former systems. The reasons for this mutation in the free market economy are numerous. Perhaps the most important was the experience of the great slump of the 1930s. After the war, public opinion, business interests, politicians, and administrators were united in resolving that massive unemployment must not be allowed to recur. Indeed, those who earlier had been the strongest adherents of laissez-faire became the strongest supporters of policies to maintain full employment, arguing that this is the best way to preserve the free enterprise system. Some observers of modern capitalism have maintained that the real purpose of the full-employment policy is to maintain profitability for the great corporations, but it cannot be denied that the results are important for all classes.

The traditional doctrines of the economists are still influential. Even John Maynard Keynes, whose critique of the traditional economics had so much to do with creating the new, gave his blessing to the old conception of the free market. He argued that if our central controls succeed in establishing full employment, there is no objection to be raised against the orthodox theory of the manner in which private self-interest will determine what in particular is produced, or how the value of the final product will be distributed.

Evidently this view owes more to sentiment than to logic. It is impossible to have a policy designed to maintain effective demand in the abstract. Every policy must have some concrete content. The instruments in the hands of government—monetary policy, the exchange rate, taxation, government expenditure—impinge in specific ways upon specific interests. Taxation may be designed to foster either investment or consumption; if the latter, it may be done either by adhering to the principle of "to him that hath shall be given" (that is, by a proportional reduction in direct tax rates) or by favouring lower income groups. Government investment cannot be neutral either; there is no criterion for allocating funds in a neutral manner between, say, armaments and hospitals.

While economic affairs grow ever more overtly political, at the same time commercial influences spread into new spheres. Old notions of loyalty, service, and proper behaviour are undermined by commercial principles. Continuous inflation and rapid technical change make it necessary for every group to defend itself against the erosion of its relative income and status. It would appear that the new capitalism, for all its benefits, may have been developing internal contradictions of its own that threaten its future stability.

MARKET SOCIALISM

Communist economic reforms

The approach to the market has changed in the Soviet sphere also. The Soviet style of administration and ideology was formed in the struggle for rapid and massive industrialization under Stalin's five-year plans. By the 1960s the Soviet economy had become too large and complex to be administered in the highly centralized fashion of earlier years. Some economists in Russia and eastern Europe, seeing that capitalism was providing its working class with a rising level of real earnings and an enticing variety of goods and services, came to the conclusion that the secret of its success lay in the price mechanism and the interplay of supply and demand. The notion of "market Socialism" thus came into vogue. The rigid, overcentralized administration of industry by a system of directives had, admittedly, led to considerable waste and inefficiency, and there was great scope for improvement in management.

Profits in managed economies

But the notion that a general solution could be found by introducing the criterion of profitability into Soviet enterprises turned out to be illusory. The secret of private enterprise (both for good and for ill) is that it is private; the manager of a capitalist concern acts as a principal and is able to use initiative and judgment. A Socialist manager that is told to make profits is still an agent acting under instructions, necessarily anxious to appear correct and to avoid responsibility. The profit goal may be somewhat less confusing and contradictory than the old multitude of targets handed down from above, but it is still arbitrary because it depends upon a system of prices fixed by decree. When the enterprise is granted the freedom to set its own prices (as in Yugoslavia), it is back to the anomalies of imperfect competition and of instability of markets. There is a school of thought that holds that the way ahead for the Soviet type of economy lies not in less planning but in more—in improving efficiency by intelligent application of devices such as linear programming with the assistance of modern computers. But mathematical methods can apply only in the strictly limited sphere of using preexisting resources for predetermined ends. It is obvious that the mathematicians cannot instruct the planners on what new products to produce. There remains the problem of finding some means of bringing the needs and tastes of would-be consumers to bear upon both the design and the composition of the flow of production intended for their benefit.

Meanwhile, many of the new and developing nations find that the opening up of their resources and their markets turns out to be more advantageous for established businesses based on successful capitalist industry than for struggling newcomers in their own countries. The economic performance of the so-called developing countries does not give much support to the doctrine that the free play of market forces can be relied upon to maximize human welfare.

(J.Ro.)

BIBLIOGRAPHY

General texts: ADAM SMITH'S famous discussion of the division of labour is found in the first chapter of *An Inquiry into the Causes of the Wealth of Nations* (1776). For ALFRED MARSHALL'S approach to the market, see his *Principles of Economics,* 9th ed., 2 vol. (1961). The instability of markets for primary commodities is the theme of the Argentinian economist RAUL PREBISCH, *Towards a New Trade Policy for Development* (1964). The development of the general equilibrium approach to markets by LEON WALRAS *et al.* is well recounted by JOSEPH SCHUMPETER in *A History of Economic Analysis* (1954). A critical account of the theory of imperfect competition is presented in JOAN ROBINSON, *Economics of Imperfect Competition,* preface to the 2nd ed. (1969). A slightly different approach is that of EDWARD H. CHAMBERLIN, *Theory of Monopolistic Competition,* 7th ed. (1956). Economies without markets are described in KARL POLANYI, *Primitive, Archaic, and Modern Economies* (1968), a collection of essays of great interest and originality. The classic appraisal of the market from the standpoint of social welfare is ARTHUR CECIL PIGOU, *The Economics of Welfare,* 4th ed. (1962). Appraisals of welfare economics include: I.M.D. LITTLE, *A Critique of Welfare Economics,* 2nd ed. (1957); J. de V. GRAAFF, *Theoretical Welfare Economics* (1957); and MAURICE DOBB, *On Economic Theory and Socialism: Collected Papers* (1955). THORSTEIN VEBLEN'S critique of the market ideology is most directly expressed in *The Place of Science in Modern Civilization, and Other Essays* (1919, reprinted 1966). The best short introduction to the Keynesian Revolution is by MICHAL KALECKI, *Studies in the Theory of Business Cycles, 1933–1939* (1966). These essays, published unnoticed in Polish, preceded the great work of JOHN MAYNARD KEYNES, *The General Theory of Employment, Interest and Money* (1935–36). ANDREW SHONFIELD, *Modern Capitalism* (1966), is a study of the ways in which various countries have adapted their economic administration to modern requirements. A more critical view of modern capitalism is that of JOHN KENNETH GALBRAITH, *The New Industrial State* (1967). A Marxist view is set forth by PAUL BARAN and PAUL SWEEZY in *Monopoly Capital* (1968). There is no satisfactory single source on the attempts at economic reform in the Soviet Union and other countries with socialist economies. A summary is given in MICHAEL ELLMAN, *Economic Reform in the Soviet Union* (1969). The economic problems of the poor countries are examined by GUNNAR MYRDAL in *The Challenge of World Poverty* (1969), a summary of his monumental work *Asian Drama* (1968). Relevant sources include MARTIN MAYER, *The Fate of the Dollar* (1980); M.S. MENDELSOHN, *Money on the Move: The Modern International Capital Market* (1980); PAUL DAVIDSON, *International Money and the Real World* (1982); JULIAN WALMSLEY, *A Dictionary of International Finance* (1979).

Commodities and futures: An extensive summary of available work on commodities and futures trading is provided by JAMES B. WOY, *Commodities Futures Trading: A Bibliographic Guide* (1976). Recent work on the specific methods and rules of the major world exchanges is included in BRIAN REIDY and JOHN EDWARDS, *Guide to World Commodity Markets,* 2nd ed. (1979). Basic works on commodity futures trading include GERALD GOLD, *Modern Commodity Futures Trading,* 4th ed., rev.

(1966, © 1959); BRUCE G. GOULD, *The Dow-Jones-Irwin Guide to Commodities Trading* (1973); and RALPH M. AINSWORTH, *Basic Principles of Commodity Futures Speculation* (1979). Legal aspects of futures trading are addressed in the Corporate Law & Practice Handbook Series, *Commodities and Futures Trading, 1977* (1977). G. BLAU, "Some Aspects of the Theory of Futures Trading," *Rev. Econ. Stud.*, 12:1–30 (1944–45), a discussion of the purpose and organization of futures markets, reasons for carrying commodities over time, the relation between cash and futures prices, and the role of arbitrage; M.J. BRENNAN, "The Supply of Storage," *American Economic Review,* 48:50–72 (1958), theory of supply of storage in relation to risk premium; P.H. COOTNER, "Speculation and Hedging," *Food Res. Inst. Stud.*, suppl. to vol. 7, pp. 65–105 (1967), theory and practice of hedging and speculation; R.W. GRAY, "The Search For a Risk Premium," *J. Polit. Economy,* 69:250–260 (1961), surveys empirical tests of the Keynes-Hicks hypothesis; J.R. HICKS, *Value and Capital,* 2nd ed., pp. 135–142 (1946), theory of "normal backwardation" also known as the Keynes-Hicks hypothesis; H.S. HOUTHAKKER, "Scope and Limits of Futures Trading," in *The Allocation of Economic Resources,* pp. 134–159 (1959), discussion of the nature of uncertainty faced by individuals and groups, cash and futures markets, futures trading and hedging, and conditions of futures trading; "Can Speculators Forecast Prices?" *Rev. Econ. and Statistics,* 39:143–151 (1957), empirical study of profits and losses for large speculators, large hedgers, and small traders in the U.S.; J.M. KEYNES, *A Treatise on Money,* vol. 2 (1930), pp. 142–144, theory of "normal backwardation" formulated by Keynes; C.S. ROCKWELL, "Normal Backwardation, Forecasting and the Returns to Commodity Futures Traders," *Food Res. Inst. Stud.*, suppl. to vol. 7, pp. 107–130 (1967), which provides empirical evidence on the returns to futures traders in large and small markets; L.G. TELSER, "Futures Trading and the Storage of Cotton and Wheat," *J. Polit. Economy,* 66:233–255 (1958), theory of storage in futures markets; "The Supply of Speculative Services in Wheat, Corn and Soybeans," *Food Res. Inst. Stud.*, suppl. to vol. 7, pp. 131–176 (1967), empirical investigation of returns to speculators and hedgers in the wheat, corn, and soybeans markets in the United States; H. WORKING, "Theory of the Inverse Carrying Charge in Futures Markets," *J. Farm Econ.*, 30:1–28 (1948), theory of inverse carrying charge seeking to explain futures market conditions when deferred futures are below the price of near futures, and futures price is below the spot price; "Futures Trading and Hedging," *Am. Econ. Rev.*, 43:314–343 (1953), theory that hedging is done for a variety of different reasons and not for the sole purpose of reducing risk; and "New Concepts Concerning Futures Markets and Prices," *Am. Econ. Rev.*, 52:431–459 (1962), which presents new concepts relating to futures markets that describe the role of hedging and the functions of futures markets. See also BARRY B. BOSWORTH and ROBERT Z. LAWRENCE, *Commodity Prices and the New Inflation* (1982); and ROBERT L. ROTHSTEIN, *Global Bargaining: UNCTAD and the Quest for a New International Economic Order* (1979).

Securities trading: W.J. BAUMOL, *The Stock Market and Economic Efficiency* (1965), an interesting effort to apply economic theory to the modern securities market; J.I. BOGEN (ed.), *The Financial Handbook,* 4th ed. (1968), a comprehensive compilation of all aspects of finance with good sections on securities trading; H. BULLOCK, *The Story of Investment Companies* (1959), an account of the development of mutual funds; W.L. CARY, *Cases and Materials on Corporations,* 4th ed. (1969), an excellent work on the law of corporations; V.P. CAROSSO, *Investment Banking in America* (1970), a relatively thorough history of investment banking; J.F. CHILDS, *Long Term Financing* (1961), a standard work providing the corporate manager with the essentials necessary to make long-range financing decisions; J.B. COHEN and E.D. ZINBARG, *Investment Analysis and Portfolio Management* (1967), a comprehensive investment text covering some of the more recent advances in the management of investments; A.S. DEWING, *Financial Policy of Corporations,* 2 vol., 5th ed. (1953), a classic treatise on financial policy, particularly useful for historical and statistical purposes; W.J. and D.K. EITEMAN, *Leading World Stock Exchanges* (1964), a survey of the practices of some of the major stock exchanges in the world; I. FRIEND *et al.*, *Investment Banking and the New Issues Market* (1967), a review of all aspects of the investment banking business with particular attention to current problems; B. GRAHAM, D.L. DODD, and S. COTTLE, *Security Analysis,* 4th ed. (1962), a classic work that led to the development of the field of security analysis; J.W. HAZARD and M. CHRISTIE, *The Investment Business* (1964), a readable abridgement of the landmark U.S. Securities and Exchange Commission's Special Study of the Securities Markets; INVESTMENT BANKERS ASSOCIATION OF AMERICA, *Fundamentals of Municipal Bonds,* 7th ed. (1969), a basic introduction to the field of municipal bonds; R.W. JENNINGS and H. MARSH (eds.), *Securities Regulation,* 2nd ed. (1968), a leading textbook dealing with the legal background of securities regulation; G.L. LEFFLER and L.C. FARWELL, *The Stock Market,* 3rd ed. (1963), a good reference book on this topic; L. LOSS, *Securities Regulation,* 2nd ed., 3 vol. (1961), 3 vol. suppl. (1969), a classic work that delves into all aspects of the U.S. federal regulation of securities and securities markets; H.G. MANNE (ed.), *Economic Policy and the Regulation of Corporate Securities* (1969), a compilation of articles by economists and other specialists on various aspects of the securities markets; S.M. ROBBINS, *The Securities Markets: Operations and Issues* (1966), an analysis of the current problems affecting the securities market; H.C. SAUVAIN, *Investment Management,* 3rd ed. (1967), a standard investment text; I.O. SCOTT, *The Government Securities Market* (1965), a recent general review of this market; D.E. SPRAY (ed.), *The Principal Stock Exchanges of the World* (1964), a good general survey of world stock exchanges; R. SOBEL, *The Big Board* (1965), a readable history of the New York Stock Exchange; F. ZARB and G. KEREKES (eds.), *The Stock Market Handbook* (1970), a relatively complete reference manual dealing with the securities industry. Later monographs include ROBERT LAMB and STEPHEN P. RAPPAPORT, *Municipal Bonds: The Comprehensive Review of Tax-Exempt Securities and Public Finance* (1980); PAUL STONHAM, *Major Stock Markets of Europe* (1982); JOEL SELIGMAN, *The Transformation of Wall Street: A History of the Securities and Exchange Commission and Modern Corporate Finance* (1982); PHYLLIS S. PIERCE (ed.), *The Dow Jones Averages, 1855–1980* (1982); DAVID N. DREMAN, *The New Contrarian Investment Strategy* (1983); and NANCY H. ROTHSTEIN and JAMES M. LITTLE (eds.), *The Handbook of Financial Futures: A Guide for Investors and Professional Financial Managers* (1984).

Money market: For the United States money market, see ROBERT V. ROOSA, *Federal Reserve Operations in the Money and Government Securities Markets* (1956); and MILTON FRIEDMAN and ANNA J. SCHWARTZ, *A Monetary History of the United States, 1867–1960* (1963). For Great Britain, W.M. SCAMMELL, *The London Discount Market* (1968), is a good general treatment, while R.S. SAYERS, *Gilletts in the London Money Market 1867–1967* (1968), is a useful case study. See also the bibliography in D.R. CROOME and H.G. JOHNSON (eds.), *Money in Britain, 1959–1969* (1970). German works include HELMUT LIPFERT, *Der Geldmarkt* (1962); and JOACHIM VON SPINDLER, *Geldmarkt, Kapitalmarkt, internationale Kreditmärkte* (1960). The best single source for Japan is HANNAN EZEKIEL, "The Call Money Market in Japan," *International Monetary Fund Staff Papers,* 13:26–49 (1966). HUGH T. PATRICK, *Monetary Policy and Central Banking in Contemporary Japan* (1962), includes an explanation of the nondevelopment of money markets in Japan, and an analysis of the instruments of monetary policy in that environment. THE BANK OF JAPAN, *The Japanese Financial System* (1969), is a brief description of financial institutions, financial markets, and characteristics of the financial structure. For undeveloped countries, SATYENDRANATH SEN, *Central Banking in Undeveloped Money Markets,* 3rd ed. (1961), is a sensible, general study, but rather dated. A good discussion of policy is EDWARD NEVIN, *Capital Funds in Underdeveloped Countries* (1961), although in some respects it is out of date. JOSEPH O. ADEKUNLE, "The Demand for Money: Evidence from Developed and Less Developed Countries," *International Monetary Fund Staff Papers,* vol. 15 (1968), is a good theoretical and empirical paper relating money both to other financial assets and to real assets. YOSHIO SUZUKI, *Money and Banking in Contemporary Japan,* trans. from the Japanese (1980), analyzes Japan's participation in international capital markets; ADAM ZWASS, *Money, Banking, and Credit in the Soviet Union and Eastern Europe* (1979), examines trade between East and West; and DAVID M. DARST, *The Handbook of the Bond and Money Markets* (1981), is a practical guide.

Marseille

Founded more than 2,500 years ago, the port city of Marseille (English conventional spelling Marseilles) has a history of vigorous independence asserted against central authority in a variety of forms. It retained its status as a free city even after falling to Julius Caesar's troops in the 1st century BC, and after centuries of decline it was revived and allowed great independence under the local control of the viscounts of Provence in the 10th–14th centuries. After Provence joined the Kingdom of France in the 15th century, Marseille retained a separate administration and continually engaged in spirited revolt against kings or governments that threatened its liberties. It was for this reason that in 1800, when France was divided into the present administrative *départements,* Marseille was only reluctantly granted its status as capital of the Bouches-du-Rhône.

Frenchmen elsewhere, convinced that the Mediterranean climes of Provence could never be fully integrated into either the French realm or the Gallic spirit, long looked upon Marseille as a sort of folkloric institution: a place of comic anecdote and dialect, with a seasoning of picturesque criminality; a place where the citizens played a peculiar form of outdoor bowling known as *pétanque* and concocted the glorious garlic- and saffron-flavoured fish stew known as bouillabaisse.

By whatever proportion fact may have been coloured with myth in its image, Marseille undoubtedly forms a major element in the economic and social structure of France. It is the second largest city proper and the largest commercial seaport of France, and the associated development at Fos-sur-Mer, about 23 miles to the northwest, has bid since the 1970s to rival Rotterdam as a transit port for all of Europe. Under the Socialist mayor Gaston Defferre, whose administration, from 1953 until his death in 1986, was the longest in its history, Marseille developed as a prosperous modern city.

Marseille remains the capital of the Bouches-du-Rhône *département* and is the capital of the Provence–Côte d'Azur region as well. It is situated on the Mediterranean's Gulf of Lion within a semicircle of limestone hills and lies 536 miles (863 kilometres) south-southeast of Paris by rail and 218 miles southeast of Lyon. The city proper has an area of about 320 square miles (830 square kilometres). The metropolitan area (1,100 square miles) covers two-thirds of the *département.* (B.E./R.C.Bu.)

This article is divided into the following sections:

Physical and human geography 569
- The landscape 569
  - The city site
  - Climate
  - The city layout
- The people 571
- The economy 571
  - Industry
  - Commerce and finance
  - Transportation
- Administration and social conditions 571
  - Government
  - Services
  - Health
  - Education
- Cultural life 572

History 572
- The early period 572
  - Antiquity and the Middle Ages
  - Uneasy union with France
  - Era of expansion
- The modern city 573

Bibliography 573

Physical and human geography

The character of Marseille has been determined to a great extent by geographic location. Its natural harbour, sheltered by a semicircle of hills and close to the estuary of the Rhône River, offered its first settlers the prospect of linking the Mediterranean seaways with northern Europe across a land that, in classical times, was made largely impassable by forests. The trading port founded by Greeks from the city of Phocaea in about 600 BC was to attract both settlers and visitors. The first of these account for a heterogeneous population and the second for services designed to cater to seamen and merchants. Marseille has the oldest chamber of commerce in France, established in 1599. It is a city of mosques and synagogues, besides many varieties of Christian churches. Its bars and brothels have been a magnet for dishonest dealings, and its waterfront still evokes the romance of a gateway to distant lands.

The city's most enduring characteristic, however, has been its readiness to welcome change. Its architecture preserves few vestiges of the past. Some landmarks, such as the transporter bridge that crossed the Old Port and the Panier district north of the harbour, were destroyed by the German occupation forces in 1943 and 1944. But more change has been wrought by the Marseillais themselves. Despite the legend that attaches to them, they are an unsentimental people open to new ideas.

For centuries, Marseille's mixed population and its inclination to political dissidence had made the city seem both foreign and marginal to French life and culture. After World War II, however, it was able to develop as a major European port and industrial centre. The city that had been the starting point for past colonial enterprises bore a large share of the aftermath of French colonialism. It proved remarkably successful in absorbing new waves of immigrants, notably the former European colonists who crossed the Mediterranean from North Africa after Algeria won its independence in 1962. The industrial complex at Fos-sur-Mer also attracted many thousands of migrant workers from North Africa in the 1960s, and for a time the city overcame the practical problem of housing them and providing for their material needs. But the economic crisis of the 1970s put a stop to industrial expansion, and, as a result of the rise in unemployment and increased racist hostility to immigrants, Marseille faced renewed difficulties in providing for its mixed population. The main cause for optimism lay in the history of a city apparently little concerned with the external monuments to its past, but which had repeatedly demonstrated its ability to adapt to the needs of a diverse population and to make good use of its human, as well as its natural, resources.

THE LANDSCAPE

The city site. Marseille lies in a sheltered depression surrounded by hills, which have inhibited the development of suburbs. The Old Port (Vieux-Port) is a natural harbour and one of the most westerly of the inlets along the rocky coastline characteristic of the northeastern Mediterranean; farther west, beyond the large tidal lake called the de Berre Lagoon (Étang de Berre), the shoreline flattens out. There the sandy dunes of the Gulf of Fos and the Camargue region in the Rhône's delta were less attractive to early mariners and were only later seen as offering possibilities for development.

Marseille's natural port was extended in Roman times and again from the 16th century onward to accommodate increased traffic and larger ships. By the 19th century the Old Port was insufficient. An artificial basin at La Joliette, built on the bay just outside of the Old Port, began operation in the mid-1840s, and five additional basins were subsequently built along a five-mile stretch of the bay. Further expansion was also undertaken to the west of the city, with the creation in 1863 of Port-Saint-Louis-du-Rhône. Eventually, in 1965, work began on the development of the port complex at Fos-sur-Mer.

Marseille's hinterland consists of a chain of mountains, known as the Étoile Chain, which leads northward toward Aix-en-Provence, formerly Marseille's rival as capital of the region, and to Mount Sainte-Victoire. The slopes around Aix are devoted to vineyards, which produce the wines of the Côtes de Provence ("Hills of Provence").

Climate. The climate of Marseille is not typical of the Mediterranean region as a whole. The lowest rainfall and highest temperatures are found in the hot, dry months of summer, but rainfall reaches a peak in spring and autumn, rather than in winter. The coldest months are December and January, when there is some frost, but otherwise the winter is mild. During the summer months the temperature rises to levels that would be unpleasant were it not for the sea breeze. In winter, Marseille is particularly subject to the dry, cold northwest wind known as the Mistral, which blows down the Rhône Valley, at times with considerable force.

The city layout. The popular area of Marseille was the seedy district, north of the Old Port, known as the Panier, which was destroyed in 1943. The more prosperous middle-class districts developed in the 19th century to the south around the rue Paradis and the avenue du Prado. The period following World War II saw various schemes to develop the city, including Le Corbusier's Unité d'Habitation, an 18-story residential block that expressed the architect's ideal of urban family lodging. The block was intended, when completed in 1952, to be one of six such units. It is now surrounded by luxury apartment buildings. Less attractive are the high-rise, working-class housing blocks developed after 1960 and the shantytowns that have grown up on the outskirts of the urban area to house immigrant workers. From the 1970s onward the municipal authorities undertook a massive program of rebuilding in various parts of the city, restoring some of the Panier and turning the district around the Old Port into a largely pedestrian area. (R.C.Bu.)

From the historic centre of Marseille at the Old Port, the thoroughfare of La Canebière climbs eastward up the hill; its name is a corruption of a Latin word for hemp, recalling Marseille's importance as a source of hemp and supplier of hemp rope in the Middle Ages. Thronged by people from around the world, La Canebière is the main commercial street in Marseille. Its starting point is marked by one of the most imposing public buildings in the city, the Bourse, which houses the Chamber of Commerce and a maritime museum.

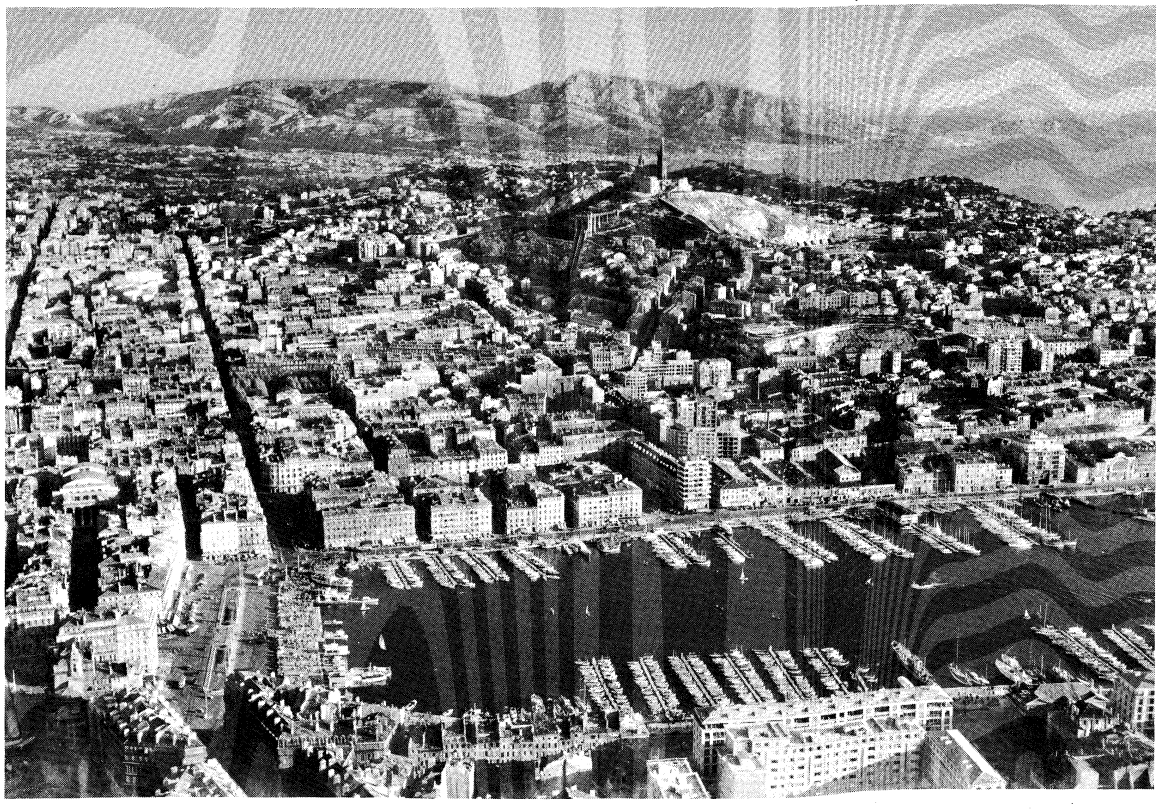
Behind the Bourse, building operations in 1967 uncovered a section of the Hellenistic ramparts of Massalia. Excavated by archaeologists, the site, dating from the 3rd and 2nd centuries BC, was found to consist of walls and towers and three sections of Roman road. The ancient port was also excavated. Nearby, close to City Hall on the edge of the port, is the Museum of Roman Warehouses (Musée des Docks Romains), which displays storage jars and other remains of commerce under Roman domination and traces the subsequent history of the port.

The port entrance is guarded by the Fort Saint-Jean, a 13th-century command post of the Knights Hospitaller of St. John of Jerusalem; some ruins remain, along with a tower built in the mid-15th century by King René of Provence. The extant fortress, dating from the 17th century, was part of a nationwide system of defenses. The other side of the harbour entrance is occupied by Fort Saint-Nicolas. In the harbour itself lie the Frioul Islands, on which the city has developed a large centre for water sports. Between these islands and the mainland is the Château d'If, a fortified island where Alexandre Dumas's fictional count of Monte Cristo and large numbers of all-too-real political prisoners were incarcerated.

The Château d'If

Other historic buildings are located around the Old Port. In the Place de la Major, the old cathedral of la Major, built on the ruins of a temple of Diana, dates from the 11th century; it was partially dismantled to make way for the eight-domed structure that in 1852 replaced it as the

By courtesy of the French Government Tourist Office



The Old Port and the sanctuary of Notre-Dame-de-la-Garde (centre), Marseille.

city's cathedral. The dome and supporting arches of the old cathedral are fine examples of Provençal Romanesque stonemasonry.

Nearby is the Old Charity Hospital (Hospice de la Vieille Charité), built between 1660 and 1750. The interior courtyard surrounds a chapel by Pierre Puget, regarded as the most powerful of French Baroque sculptors. Close by is the Hôtel Dieu, the oldest hospital in the city, built at the end of the 16th century. The principal building, by Jacques Hardouin-Mansart, was erected 200 years later and still serves its original function. Almost next door, the bell tower of the vanished church of Accoules, a 14th-century spire mounted on a 12th-century tower, marks the centre of Old Marseille.

On the opposite side of the port stands the crenellated, square-towered basilica of Saint-Victor, dating from the 11th to the 14th century; it once was attached to an abbey founded about 413 by St. John Cassian to commemorate a 3rd-century martyr and patron saint of sailors and millers. When Saint-Victor was built, the abbey was a temporal power of considerable extent, ruling properties in Spain, Sardinia, and the hinterlands of France.

High on the hill over the south side of the Old Port stands the celebrated Notre-Dame-de-la-Garde, a sanctuary honoured from the 8th century. Its present structure was built in 1853–64; its steeple, crowned by a 30-foot (nine-metre) gilded statue of the Virgin, rises 150 feet over the hillside. (B.E./R.C.Bu.)

THE PEOPLE

Marseille's population, drawn from all parts of the Mediterranean and from elsewhere in Europe and Africa, has always been mixed, so that it has never been possible to talk of a "typical" Marseillais. In 1880, for example, more than one in six of the inhabitants of the city were foreign. New residents have created a diverse pattern, sometimes concentrated in certain districts, such as the Muslim quarter that grew up during the 1970s north of La Canebière, and sometimes specializing in particular trades or professions. Certain groups—Jews, Greeks, Armenians—have their own community leaderships, which have semiofficial recognition.

Prominent among the new additions to the population are young technicians drawn from other parts of France to staff the expanding technical centres. Former colonials have had a strong impact on the community, and Marseille has always attracted Corsicans (including the Bonaparte family during the French Revolution). Manual labour is increasingly performed by North Africans or Africans who arrive from former colonies.

THE ECONOMY

Industry. Marseille is a major manufacturing centre with a wide range of industries, the most important of which are petroleum refining, shipbuilding, and metalworking. The oldest industry, founded in the 15th century, is the manufacture of soap from olive oil produced in the surrounding district. This, like the production of sugar, has long been overtaken by metallurgy, the production of building materials, and the chemical industry, originally related to the needs of the wine producers of the area and now also associated with the import of petrochemicals through the installations at Fos.

The port development around the de Berre Lagoon and at Fos-sur-Mer attracted much new industrial development after 1966. The port facilities include terminals for oil tankers, minerals, natural gas, container traffic, and general merchandise. The industrial zone of Fos–Lavéra–Berre comprises steelworks, chemical factories, a refinery, and a metal construction plant at Fos; a petrochemical plant and a refinery at Lavéra; and a petrochemical complex and refineries at the de Berre Lagoon. The majority of these installations use raw materials that enter through the port of Marseille and are expedited, after processing, by highway, canal, railroad, pipeline, or sea.

The port complex of Marseille–Fos is the largest in France and among the largest in Europe. Its dockyards account for two-thirds of the French industry for repair and maintenance of ships, with several dry docks (one floating), as well as specialized installations for cleaning and maintenance of oil tankers. The port complex is administered by the Port Autonome de Marseille ("Autonomous Port of Marseille"), a financially autonomous state enterprise that is responsible for the construction, administration, and maintenance of the industrial zones at Fos and Lavéra and the port facilities at Marseille, Lavéra, Caronte, Fos, and Port-Saint-Louis-du-Rhône. In addition to administering Marseille–Fos, the Port Autonome de Marseille also provides advice, information, and planning services for port authorities around the world. It is financed by rents, taxes, and fees for services, with state aid for investment in construction of harbours and quays.

Commerce and finance. The site of Marseille, on the Mediterranean and close to the estuary of the Rhône River, has long been recognized as a particularly favourable juncture of trade routes. The commercial and passenger ports are served by major freeways to northern Europe, Italy, and Spain; by railroad (Marseille–Lyon–Paris) via the Rhône Valley; by canal; and by pipelines for oil, natural gas, and ethylene to other parts of France and to West Germany and Switzerland. While the traffic in oil and petrochemicals declined steadily during the 1970s, as a result of the oil crisis and conversion to alternative fuels, trade in other goods, particularly mineral ores, has increased. The port of Marseille was handling traffic of well over a million passengers and a third of a million automobiles by the mid-1980s.

Because of its geographic position and its commercial importance, Marseille has long been able to attract foreign capital. In the mid-19th century two major banks were established: the Société de Crédit Foncier de Marseille in 1852 and the Société Marseillaise de Crédit Industriel et Commercial et de Dépôts in 1865. Local money was directed both into industrial projects in the region and into foreign ventures such as the Suez Canal Company. But the disintegration of the French empire during the 20th century helped to accelerate the decline of Marseille as a financial centre, and development projects in the region now depend on investment by the regional and national governments.

Marseille remains the site of one of France's stock exchanges. The Marseille exchange trades particularly in minerals and chemical products. (R.C.Bu.)

Transportation. For the 100 years preceding 1953, no important thoroughfare had been constructed in Marseille. To cope with the engulfing tide of automobiles, the city government inaugurated a program of street widening and freeway and tunnel construction that, together with the building program, has altered (and, some feel, banalized) the appearance of Marseille. The first line in the subway system was opened in 1977. A network of trains and highways now connects Marseille with other European cities. Marignane Airport, 18 miles south of the city, provides air transportation. (B.E./R.C.Bu.)

Innovations in transportation facilities

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The city government consists of a popularly elected municipal council. The council keeps very much alive the historical tradition of local independence in spite of the intimate involvement of many national ministries in the financing and planning of projects throughout the area.

The city is divided into 16 *arrondissements,* but for the purposes of local government these are grouped into six *secteurs,* which elect mayors. In addition to the six city halls, one for each *secteur,* there are two "mini city halls" in each *arrondissement.* The city mayor is assisted by a local government of 30 *adjoints,* each with responsibility for a particular facet of government, such as town planning, culture, finance, employment, or transport, and by delegate councillors who assist the *adjoints* or undertake more detailed responsibilities.

Services. The city's *adjoints* oversee the main urban services administered by the local authorities: lighting, refuse disposal, relations with the police and fire services, and so on. With an unemployment rate above the national average and a large population of immigrant workers, Marseille has faced acute problems in the provision

of public services, although it has also shown remarkable success in overcoming them. The most acute problem which remains unsolved is that of sewage disposal: in the mid-1980s sewage was still being emptied directly into the Mediterranean. In summer the region is particularly threatened by forest fires, and Marseille is the centre from which the firefighting is coordinated. A fleet of specially equipped airplanes is stationed at Marignane Airport.

Health. Because of the nature of its population, Marseille also faces particular health problems. Both the crime rate and the accident rate (including that for industrial accidents) are above the national average. On the other hand, suicides and mental illness are below the average for the rest of France.

There are two teaching hospital complexes in Marseille, the North Hospital and the Timone Hospital. The centre for the study of tropical medicine at the Michel Lévy Hospital is well known. A computer centre in the suburb of Luminy links the region's hospitals.

Education. In 1968 the University of Provence was divided into three, with the universities of Aix–Marseille I and II in Marseille and the University of Aix–Marseille III in Aix. The university is further divided into teaching and research units, which include the scientific unit at Luminy, the regional institute for sports and physical education, the unit for the study of marine sciences and the environment, and the Institute of Fluid Mechanics. The administrative headquarters of the university are near the Old Port. At Luminy are located the National School of Architecture, the Regional School of Fine Arts, and the Marseille School of Business Administration.

(R.C.Bu.)

CULTURAL LIFE

Marseille has several museums, including a very popular Children's Museum. The Museum of Old Marseille was installed in 1960 next to the City Hall in Diamond House (La Maison Diamantée), so called because of its 16th-century facade of projecting diamond-shaped stone lozenges. The Cantini Museum, close to the rue Paradis, east of the Old Port, has a fine collection of Oriental art, of local pottery, and of modern paintings and sculptures.

In 1975 the mayor of Marseille signed a "Cultural Charter" providing for state participation in subsidizing the opera and other cultural ventures. Both state and local subsidies for major cultural projects are administered by the Municipal Office of Culture, which helped to set up two theatres opened in 1980 on the site of a former fish market, a media centre for various types of spectacle, and the Roland Petit Dance School. The city has tried to reflect the diverse cultures of Marseille by encouraging exchanges with artists and companies from Algeria and other North African countries. During the 1930s, Marcel Pagnol founded a film studio in Marseille that made the city, for a time, the only centre of the industry outside the Paris region.

As building has increased within Marseille, more attention has been paid to the conservation and development of the municipality's parks and playgrounds. In addition to a municipal sports centre with a 50,000-seat stadium and an outdoor theatre, Marseille has public swimming pools and a number of sports fields and public beaches. The parks of the Château du Pharo, Château Borély, and Palais Longchamp are extensive.

History

THE EARLY PERIOD

The oldest of the large French cities, Marseille was founded as Massalia (Massilia) by Greek mariners from Phocaea in Asia Minor about 600 BC. Archaeological finds exhibited in the Museum of Antiquities in the 18th-century Château Borély suggest that Phoenicians had settled there even earlier.

Antiquity and the Middle Ages. The Massalians spread trading posts inland as well as along the coasts, westward to Spain, and eastward to Monaco, founding the present cities of Arles, Nice, Antibes, Agde, and La Ciotat. Their coins have been found across France and through the Alps as far as the Tirol. In the 4th century BC a Massalian, Pytheas, visited the coasts of Gaul, Britain, and Germany, and a Euthymenes is said to have navigated the west coast of Africa as far south as Senegal.

When their great trade rivals, the Carthaginians, fought the Romans in the Punic Wars, Marseille supported the Romans and received help in subduing the native tribes of Liguria. When Pompey and Julius Caesar clashed, Marseille took Pompey's side and subsequently fell to Caesar's lieutenant Trebonius in 49 BC. Although stripped of dependencies, it was permitted to retain its status as a free city in recognition of past services. For some time the city remained the last centre of Greek learning in the West, but, eventually, it declined almost to extinction. After centuries of invasion and epidemic, it became little more than a huddle of nearly abandoned ruins.

Decline and resuscitation

In the 10th century, under the protection of successive viscounts of Provence, the area was repopulated, and it found new prosperity as a shipping and staging point for the Crusades. Gradually, the town bought up the rights of the viscounts, and, at the beginning of the 13th century, it formed a republic around the Old Port, though the upper part of the city and its southern suburb remained under ecclesiastical jurisdiction. The counts of Provence allowed the city great independence, and only in 1245 and 1256 did Charles of Anjou force acknowledgement of his sovereignty. After Marseille was sacked by Alfonso V of Aragon in 1423, King René of Provence, whose winter residence was there, restored prosperity to the city.

Uneasy union with France. When Provence, including Marseille, became part of the Kingdom of France in 1481, the city preserved a separate administration directed by royal officials. During the 16th-century wars of religion, Marseille was fanatically Roman Catholic and long refused to recognize Henry IV as king because, until his conversion to Catholicism and accession to the French throne, he had been leader of the Protestants. During the Fronde, a movement in 1648–53 that opposed royal absolutism, the city sought to conserve its ancient liberties and rose against Louis XIV, who in 1660 came in person, breached the walls, and subdued the revolt. To discourage further manifestations of independence, the King planted Fort Saint-Nicolas at the southern extremity of the Old Port. In the same year, the city pushed inland to the west beyond its walls. A few buildings constructed in this expansion still survive in the area around the Cours Belsunce—a major thoroughfare—and the Préfecture.

Marseille joined enthusiastically in the French Revolution. Some 500 volunteers marching to Paris in 1792 sang "The War Song of the Rhine Army," which had been composed in Strasbourg in the late 18th century. The song, which thrilled the crowds along the route of march, was renamed "La Marseillaise" and became the national anthem of France. As the early federalist concepts were washed away in the blood of the Reign of Terror, however, the city revolted against the ruling Convention. Quickly mastered by force of arms, it was officially designated as "the city without a name." When its commerce was almost destroyed by the maritime blockade of the Continent directed against Napoleon, Marseille became bitterly anti-Bonapartist and hailed the Bourbon restoration. Under Napoleon III, however, it remained stubbornly republican.

Marseille and the Revolution

Era of expansion. During the second half of the 19th century, Marseille was expanded as the "port of empire," gaining impetus from the elimination of the Barbary pirates (1815–35), the conquest of Algeria (1830), and the inauguration of the Suez Canal (1869). The great avenues and many of the monuments of the city were constructed in this period. A serious water problem was solved by a project (1837–48) that brought water from the Durance River. The distribution reservoir above the city was disguised as the Longchamp Palace, containing the Museum of Natural History and the Museum of Fine Arts. The Château du Pharo, one of the city's principal landmarks, was built as a villa for Napoleon III and Eugénie at the edge of the bay beyond the Old Port, but it was never occupied by the imperial couple. The Bourse, an imposing colonnaded structure on La Canebière, was built in 1852–60 to house the Chamber of Commerce. (B.E./R.C.Bu.)

THE MODERN CITY

The city, occupied by the German Army from November 1942 until August 1944, continued to be an active centre of the French Resistance movement, which partly explains the German decision to dynamite the Panier district and the Old Port in 1943. Further destruction was caused by German mines in August 1944.

The postwar history of Marseille is the story of the rebuilding of these areas damaged in the war and the development of the industrial and port complexes around Fos. Many of the best-known Marseillais industries were no longer owned by local industrialists but formed part of international and multinational concerns. With its ambition to become the centre of a large economic and geographic region, the Rhône-Alps-Côte d'Azur, and a rival to Rotterdam as a European port, Marseille's conception of itself has changed; it has lost some of its individual character in exchange for closer integration with the rest of France and with Europe. (R.C.Bu.)

BIBLIOGRAPHY

General works: DOUGLAS GOLDRING, *The South of France: The Lower Rhône Valley and the Mediterranean Seaboard from Martigues to Menton* (1952), contains a description of Marseille and environs; ALFRED FIRTH, *French Life & Landscape,* vol. 2, *Southern France* (1953), contains a brief discussion of the city and its inhabitants; PAUL CARRÈRE and RAYMOND DUGRAND, *La Région Méditerranéenne,* 2nd rev. ed. (1967), deals with the city's industrial region; ARCHIBALD LYALL, *The Companion Guide to the South of France,* 3rd ed. rev. and expanded by A.N. BRANGHAM (1972, reissued 1983), includes a discussion of the topography, principal monuments, and cultural institutions of Marseille; GERMAINE KRULL, *Marseille* (1935), is a collection of photographs with text by ANDRÉ SUARÈS; IAN B. THOMPSON, *The Lower Rhone and Marseille* (1975); M.F.K. FISHER, *A Considerable Town* (1978), presents vignettes on the city and its people; ANDRÉ REMACLE, *Marseille à coeur ouvert* (1981), gives the personal view of a Marseille journalist and deals with the character and history of the city; and GILBERT ROCHU, *Marseille, les années Defferre* (1983), is a critical analysis of the administration of the city in the period since World War II.

History: GASTON RAMBERT (ed.), *Histoire du commerce de Marseille,* 6 vol. (1949–59); RAOUL BUSQUET, *Histoire de Marseille,* rev. ed. updated by PIERRE GUIRAL (1977); LUCIEN GAILLARD, *Marseille sous l'occupation* (1982), covers the occupation years of World War II and is illustrated with contemporary photographs and reproductions of documents; and PIERRE GUIRAL and PAUL AMARGIER, *Histoire de Marseille* (1983), contains a chapter on current politics and problems.

(B.E./R.C.Bu.)

Marx and Marxism

Karl Marx, revolutionary, sociologist, historian, and economist, was the author (with Friedrich Engels) of *Manifest der Kommunistischen Partei* (1848), commonly known as *The Communist Manifesto,* the most celebrated pamphlet in the history of the socialist movement, as well as of its most important book, *Das Kapital.* These writings and others by Marx and Engels form the basis of the body of thought and belief that is known as Marxism.

This article deals with Marx's life, his thinking, his accomplishments, and the development of Marxist theory. See the article SOCIO-ECONOMIC DOCTRINES AND REFORM MOVEMENTS, MODERN, for a full treatment of socialism and communism. For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 541.

The article is divided into the following sections:

Life and works of Marx 573
- Early years 573
- Brussels period 574
- Early years in London 575
- Role in the First International 576
- Last years 576
- Character and significance 577

Marxism 577
- The thought of Karl Marx 577
  - Historical materialism
  - Analysis of society
  - Analysis of the economy
  - Class struggle
  - The contributions of Engels
- German Marxism after Engels 580
  - The work of Kautsky and Bernstein
  - The radicals
  - The Austrians
- Russian and Soviet Marxism 581
  - Lenin
  - The dictatorship of the proletariat
  - Stalin
  - Trotskyism
- Variants of Soviet Marxism 583
  - Maoism
  - Marxism in Yugoslavia and Cuba
  - Marxism in the Third World
  - Marxism in the West

Bibliography 585

Life and works of Marx

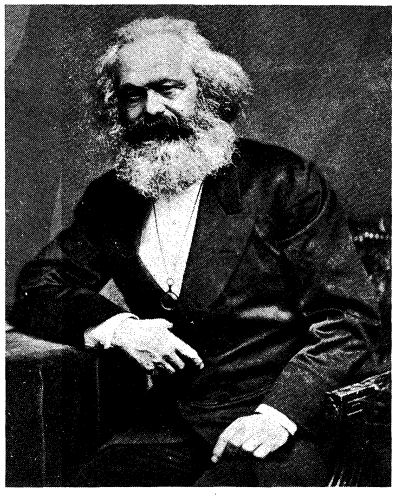
EARLY YEARS

Karl Heinrich Marx was born on May 5, 1818, in the city of Trier in the Rhine province of Prussia, now in West Germany. He was the oldest surviving boy of nine children. His father, Heinrich, a successful lawyer, was a man of the Enlightenment, devoted to Kant and Voltaire, who took part in agitations for a constitution in Prussia. His mother, born Henrietta Pressburg, was from Holland. Both parents were Jewish and were descended from a long line of rabbis, but, a year or so before Karl was born, his father—probably because his professional career required it—was baptized in the Evangelical Established Church. Karl was baptized when he was six years old. Although as a youth Karl was influenced less by religion than by the critical, sometimes radical social policies of the Enlightenment, his Jewish background exposed him to prejudice and discrimination that may have led him to question the role of religion in society and contributed to his desire for social change.

Marx was educated from 1830 to 1835 at the high school in Trier. Suspected of harbouring liberal teachers and pupils, the school was under police surveillance. Marx's writings during this period exhibited a spirit of Christian devotion and a longing for self-sacrifice on behalf of humanity. In October 1835 he matriculated at the University of Bonn. The courses he attended were exclusively in the humanities, in such subjects as Greek and Roman mythology and the history of art. He participated in customary student activities, fought a duel, and spent a day in jail for being drunk and disorderly. He presided at the Tavern Club and joined a poets' club that included some political activists. A politically rebellious student culture was, indeed, part of life at Bonn. Marx, however, left Bonn after a year and in October 1836 enrolled at the University of Berlin to study law and philosophy.

The Young Hegelians

Marx's crucial experience at Berlin was his introduction to Hegel's philosophy, regnant there, and his adherence to the Young Hegelians. At first he felt a repugnance toward Hegel's doctrines; when Marx fell sick it was partially, as he wrote his father, "from intense vexation at having to make an idol of a view I detested." The Hegelian pressure in the revolutionary student culture was powerful, however, and Marx joined a society called the Doctor Club, whose members were intensely involved in the new



Marx.

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

literary and philosophical movement. Their chief figure was Bruno Bauer, a young lecturer in theology, who was developing the idea that the Christian Gospels were a record not of history but of human fantasies arising from emotional needs and that Jesus had not been a historical person. Marx enrolled in a course of lectures given by Bauer on the prophet Isaiah. Bauer taught that a new social catastrophe "more tremendous" than that of the advent of Christianity was in the making. The Young Hegelians began moving rapidly toward atheism and also talked vaguely of political action.

The Prussian government, fearful of the subversion latent in the Young Hegelians, soon undertook to drive them from the universities. Bauer was dismissed from his post in 1839. Marx's "most intimate friend" of this period, Adolph Rutenberg, an older journalist who had served a prison sentence for his political radicalism, pressed for a deeper social involvement. By 1841 the Young Hegelians had become left republicans. Marx's studies, meanwhile, were lagging. Urged by his friends, he submitted a doctoral dissertation to the university at Jena, which was known to be lax in its academic requirements, and received his degree in April 1841. His thesis analyzed in a Hegelian fashion the difference between the natural philosophies of Democritus and Epicurus. More distinctively, it sounded a note of Promethean defiance:

> Philosophy makes no secret of it. Prometheus' admission: "In sooth all gods I hate," is its own admission, its own motto against all gods, . . . Prometheus is the noblest saint and martyr in the calendar of philosophy.

In 1841 Marx, together with other Young Hegelians, was much influenced by the publication of *Das Wesen des Christentums* (1841; *The Essence of Christianity*) by Ludwig Feuerbach. Its author, to Marx's mind, successfully criticized Hegel, an idealist who believed that matter or existence was inferior to and dependent upon mind or spirit, from the opposite, or materialist, standpoint, showing how the "Absolute Spirit" was a projection of "the real man standing on the foundation of nature." Henceforth Marx's philosophical efforts were toward a combination of Hegel's dialectic—the idea that all things are in a continual process of change resulting from the conflicts between their contradictory aspects—with Feuerbach's materialism, which placed material conditions above ideas.

In January 1842 Marx began contributing to a newspaper newly founded in Cologne, the *Rheinische Zeitung*. It was the liberal democratic organ of a group of young merchants, bankers, and industrialists; Cologne was the centre of the most industrially advanced section of Prussia. To this stage of Marx's life belongs an essay on the freedom of the press. Since he then took for granted the existence of absolute moral standards and universal principles of ethics, he condemned censorship as a moral evil that entailed spying into people's minds and hearts and assigned to weak and malevolent mortals powers that presupposed an omniscient mind. He believed that censorship could have only evil consequences.

Newspaper work

On Oct. 15, 1842, Marx became editor of the *Rheinische Zeitung*. As such, he was obliged to write editorials on a variety of social and economic issues, ranging from the housing of the Berlin poor and the theft by peasants of wood from the forests to the new phenomenon of communism. He found Hegelian idealism of little use in these matters. At the same time he was becoming estranged from his Hegelian friends for whom shocking the bourgeois was a sufficient mode of social activity. Marx, friendly at this time to the "liberal-minded practical men" who were "struggling step-by-step for freedom within constitutional limits," succeeded in trebling his newspaper's circulation and making it a leading journal in Prussia. Nevertheless, Prussian authorities suspended it for being too outspoken, and Marx agreed to coedit with the liberal Hegelian Arnold Ruge a new review, the *Deutsch-französische Jahrbücher* ("German-French Yearbooks"), which was to be published in Paris.

First, however, in June 1843 Marx, after an engagement of seven years, married Jenny von Westphalen. Jenny was an attractive, intelligent, and much-admired woman, four years older than Karl; she came of a family of military and administrative distinction. Her half-brother later became a highly reactionary Prussian minister of the interior. Her father, a follower of the French socialist Saint-Simon, was fond of Karl, though others in her family opposed the marriage. Marx's father also feared that Jenny was destined to become a sacrifice to the demon that possessed his son.

Four months after their marriage, the young couple moved to Paris, which was then the centre of socialist thought and of the more extreme sects that went under the name of communism. There, Marx first became a revolutionary and a communist and began to associate with communist societies of French and German workingmen. Their ideas were, in his view, "utterly crude and unintelligent," but their character moved him: "The brotherhood of man is no mere phrase with them, but a fact of life, and the nobility of man shines upon us from their work-hardened bodies," he wrote in his so-called "Ökonomisch-philosophische Manuskripte aus dem Jahre 1844" (written in 1844; *Economic and Philosophic Manuscripts of 1844* [1959]). (These manuscripts were not published for some 100 years, but they are influential because they show the humanist background to Marx's later historical and economic theories.)

The "German-French Yearbooks" proved short-lived, but through their publication Marx befriended Friedrich Engels, a contributor who was to become his lifelong collaborator, and in their pages appeared Marx's article "Zur Kritik der Hegelschen Rechtsphilosophie" ("Toward the Critique of the Hegelian Philosophy of Right") with its oft-quoted assertion that religion is the "opium of the people." It was there, too, that he first raised the call for an "uprising of the proletariat" to realize the conceptions of philosophy. Once more, however, the Prussian government intervened against Marx. He was expelled from France and left for Brussels—followed by Engels—in February 1845. That year in Belgium he renounced his Prussian nationality.

BRUSSELS PERIOD

Collaboration with Engels

The next two years in Brussels saw the deepening of Marx's collaboration with Engels. Engels had seen at firsthand in Manchester, Eng., where a branch factory of his father's textile firm was located, all the depressing aspects of the Industrial Revolution. He had also been a Young Hegelian and had been converted to communism by Moses Hess, who was called the "communist rabbi." In England he associated with the followers of Robert Owen. Now he and Marx, finding that they shared the same views, combined their intellectual resources and published *Die heilige Familie* (1845; *The Holy Family*), a prolix criticism of the Hegelian idealism of the theologian Bruno Bauer. Their next work, *Die deutsche Ideologie* (written 1845–46, published 1932; *The German Ideology*), contained the fullest exposition of their important materialistic conception of history, which set out to show how, historically, societies

had been structured to promote the interests of the economically dominant class. But it found no publisher and remained unknown during its authors' lifetimes.

During his Brussels years, Marx developed his views and, through confrontations with the chief leaders of the working-class movement, established his intellectual standing. In 1846 he publicly excoriated the German leader Wilhelm Weitling for his moralistic appeals. Marx insisted that the stage of bourgeois society could not be skipped over; the proletariat could not just leap into communism; the workers' movement required a scientific basis, not moralistic phrases. He also polemicized against the French socialist thinker Pierre-Joseph Proudhon in *Misère de la philosophie* (1847; *The Poverty of Philosophy*), a mordant attack on Proudhon's book subtitled *Philosophie de la misère* (1846; *The Philosophy of Poverty*). Proudhon wanted to unite the best features of such contraries as competition and monopoly; he hoped to save the good features in economic institutions while eliminating the bad. Marx, however, declared that no equilibrium was possible between the antagonisms in any given economic system. Social structures were transient historic forms determined by the productive forces: "The handmill gives you society with the feudal lord; the steammill, society with the industrial capitalist." Proudhon's mode of reasoning, Marx wrote, was typical of the petty bourgeois, who failed to see the underlying laws of history.

The Communist Manifesto

An unusual sequence of events led Marx and Engels to write their pamphlet *The Communist Manifesto.* In June 1847 a secret society, the League of the Just, composed mainly of emigrant German handicraftsmen, met in London and decided to formulate a political program. They sent a representative to Marx to ask him to join the league; Marx overcame his doubts and, with Engels, joined the organization, which thereupon changed its name to the Communist League and enacted a democratic constitution. Entrusted with the task of composing their program, Marx and Engels worked from the middle of December 1847 to the end of January 1848. The London Communists were already impatiently threatening Marx with disciplinary action when he sent them the manuscript; they promptly adopted it as their manifesto. It enunciated the proposition that all history had hitherto been a history of class struggles, summarized in pithy form the materialist conception of history worked out in *The German Ideology,* and asserted that the forthcoming victory of the proletariat would put an end to class society forever. It mercilessly criticized all forms of socialism founded on philosophical "cobwebs" such as "alienation." It rejected the avenue of "social Utopias," small experiments in community, as deadening the class struggle and therefore as being "reactionary sects." It set forth 10 immediate measures as first steps toward communism, ranging from a progressive income tax and the abolition of inheritances to free education for all children. It closed with the words, "The proletarians have nothing to lose but their chains. They have a world to win. Workingmen of all countries, unite!"

Participation in the events of 1848

Revolution suddenly erupted in Europe in the first months of 1848, in France, Italy, and Austria. Marx had been invited to Paris by a member of the provisional government just in time to avoid expulsion by the Belgian government. As the revolution gained in Austria and Germany, Marx returned to the Rhineland. In Cologne he advocated a policy of coalition between the working class and the democratic bourgeoisie, opposing for this reason the nomination of independent workers' candidates for the Frankfurt Assembly and arguing strenuously against the program for proletarian revolution advocated by the leaders of the Workers' Union. He concurred in Engels' judgment that *The Communist Manifesto* should be shelved and the Communist League disbanded. Marx pressed his policy through the pages of the *Neue Rheinische Zeitung,* newly founded in June 1849, urging a constitutional democracy and war with Russia. When the more revolutionary leader of the Workers' Union, Andreas Gottschalk, was arrested, Marx supplanted him and organized the first Rhineland Democratic Congress in August 1848. When the king of Prussia dissolved the Prussian Assembly in Berlin, Marx called for arms and men to help the resistance. Bourgeois liberals withdrew their support from Marx's newspaper, and he himself was indicted on several charges, including advocacy of the nonpayment of taxes. In his trial he defended himself with the argument that the crown was engaged in making an unlawful counterrevolution. The jury acquitted him unanimously and with thanks. Nevertheless, as the last hopeless fighting flared in Dresden and Baden, Marx was ordered banished as an alien on May 16, 1849. The final issue of his newspaper, printed in red, caused a great sensation.

EARLY YEARS IN LONDON

Expelled once more from Paris, Marx went to London in August 1849. It was to be his home for the rest of his life. Chagrined by the failure of his own tactics of collaboration with the liberal bourgeoisie, he rejoined the Communist League in London and for about a year advocated a bolder revolutionary policy. An "Address of the Central Committee to the Communist League," written with Engels in March 1850, urged that in future revolutionary situations they struggle to make the revolution "permanent" by avoiding subservience to the bourgeois party and by setting up "their own revolutionary workers' governments" alongside any new bourgeois one. Marx hoped that the economic crisis would shortly lead to a revival of the revolutionary movement; when this hope faded, he came into conflict once more with those whom he called "the alchemists of the revolution," such as August von Willich, a communist who proposed to hasten the advent of revolution by undertaking direct revolutionary ventures. Such persons, Marx wrote in September 1850, substitute "idealism for materialism" and regard

> pure will as the motive power of revolution instead of actual conditions. While we say to the workers: "You have got to go through fifteen, twenty, fifty years of civil wars and national wars not merely in order to change your conditions but in order to change yourselves and become qualified for political power," you on the contrary tell them, "We must achieve power immediately."

The militant faction in turn ridiculed Marx for being a revolutionary who limited his activity to lectures on political economy to the Communist Workers' Educational Union. The upshot was that Marx gradually stopped attending meetings of the London Communists. In 1852 he devoted himself intensely to working for the defense of 11 communists arrested and tried in Cologne on charges of revolutionary conspiracy and wrote a pamphlet on their behalf. The same year he also published, in a German-American periodical, his essay "Der Achtzehnte Brumaire des Louis Napoleon" (*The Eighteenth Brumaire of Louis Bonaparte*), with its acute analysis of the formation of a bureaucratic absolutist state with the support of the peasant class. In other respects the next 12 years were, in Marx's words, years of "isolation" both for him and for Engels in his Manchester factory.

Poverty in London

From 1850 to 1864 Marx lived in material misery and spiritual pain. His funds were gone, and except on one occasion he could not bring himself to seek paid employment. In March 1850 he and his wife and four small children were evicted and their belongings seized. Several of his children died—including a son Guido, "a sacrifice to bourgeois misery," and a daughter Franziska, for whom his wife rushed about frantically trying to borrow money for a coffin. For six years the family lived in two small rooms in Soho, often subsisting on bread and potatoes. The children learned to lie to the creditors: "Mr. Marx ain't upstairs." Once he had to escape them by fleeing to Manchester. His wife suffered breakdowns.

During all these years Engels loyally contributed to Marx's financial support. The sums were not large at first, for Engels was only a clerk in the firm of Ermen and Engels at Manchester. Later, however, in 1864, when he became a partner, his subventions were generous. Marx was proud of Engels' friendship and would tolerate no criticism of him. Bequests from the relatives of Marx's wife and from Marx's friend Wilhelm Wolff also helped to alleviate their economic distress.

Marx had one relatively steady source of earned income

in the United States. On the invitation of Charles A. Dana, managing editor of *The New York Tribune,* he became in 1851 its European correspondent. The newspaper, edited by Horace Greeley, had sympathies for Fourierism, a Utopian socialist system developed by the French theorist Charles Fourier. From 1851 to 1862 Marx contributed close to 500 articles and editorials (Engels providing about a fourth of them). He ranged over the whole political universe, analyzing social movements and agitations from India and China to Britain and Spain.

In 1859 Marx published his first book on economic theory, *Zur Kritik der politischen Ökonomie* (*A Contribution to the Critique of Political Economy*). In its preface he again summarized his materialistic conception of history, his theory that the course of history is dependent on economic developments. At this time, however, Marx regarded his studies in economic and social history at the British Museum as his main task. He was busy producing the drafts of his magnum opus, which was to be published later as *Das Kapital.* Some of these drafts, including the *Outlines* and the *Theories of Surplus Value,* are important in their own right and were published after Marx's death.

ROLE IN THE FIRST INTERNATIONAL

Marx's political isolation ended in 1864 with the founding of the International Working Men's Association. Although he was neither its founder nor its head, he soon became its leading spirit. Its first public meeting, called by English trade union leaders and French workers' representatives, took place at St. Martin's Hall in London on Sept. 28, 1864. Marx, who had been invited through a French intermediary to attend as a representative of the German workers, sat silently on the platform. A committee was set up to produce a program and a constitution for the new organization. After various drafts had been submitted that were felt to be unsatisfactory, Marx, serving on a subcommittee, drew upon his immense journalistic experience. His "Address and the Provisional Rules of the International Working Men's Association," unlike his other writings, stressed the positive achievements of the cooperative movement and of parliamentary legislation; the gradual conquest of political power would enable the British proletariat to extend these achievements on a national scale.

As a member of the organization's General Council, and corresponding secretary for Germany, Marx was henceforth assiduous in attendance at its meetings, which were sometimes held several times a week. For several years he showed a rare diplomatic tact in composing differences among various parties, factions, and tendencies. The International grew in prestige and membership, its numbers reaching perhaps 800,000 in 1869. It was successful in several interventions on behalf of European trade unions engaged in struggles with employers.

The Paris Commune

In 1870, however, Marx was still unknown as a European political personality; it was the Paris Commune that made him into an international figure, "the best calumniated and most menaced man of London," as he wrote. When the Franco-German War broke out in 1870, Marx and Engels disagreed with followers in Germany who refused to vote in the Reichstag in favour of the war. The General Council declared that "on the German side the war was a war of defence." After the defeat of the French armies, however, they felt that the German terms amounted to aggrandizement at the expense of the French people. When an insurrection broke out in Paris and the Paris Commune was proclaimed, Marx gave it his unswerving support. On May 30, 1871, after the Commune had been crushed, he hailed it in a famous address entitled *Civil War in France:*

> History has no comparable example of such greatness.... Its martyrs are enshrined forever in the great heart of the working class.

In Engels' judgment, the Paris Commune was history's first example of the "dictatorship of the proletariat." Marx's name, as the leader of The First International and author of the notorious *Civil War,* became synonymous throughout Europe with the revolutionary spirit symbolized by the Paris Commune.

The advent of the Commune, however, exacerbated the antagonisms within the International Working Men's Association and thus brought about its downfall. English trade unionists such as George Odger, former president of the General Council, opposed Marx's support of the Paris Commune. The Reform Bill of 1867, which had enfranchised the British working class, had opened vast opportunities for political action by the trade unions. English labour leaders found they could make many practical advances by cooperating with the Liberal Party and, regarding Marx's rhetoric as an encumbrance, resented his charge that they had "sold themselves" to the Liberals.

The struggle with Bakunin

A left opposition also developed under the leadership of the famed Russian revolutionary Mikhail Alexandrovich Bakunin. A veteran of tsarist prisons and Siberian exile, Bakunin could move men by his oratory, which one listener compared to "a raging storm with lightning, flashes and thunderclaps, and a roaring as of lions." Bakunin admired Marx's intellect but could hardly forget that Marx had published a report in 1848 charging him with being a Russian agent. He felt that Marx was a German authoritarian and an arrogant Jew who wanted to transform the General Council into a personal dictatorship over the workers. He strongly opposed several of Marx's theories, especially Marx's support of the centralized structure of the International, Marx's view that the proletariat class should act as a political party against prevailing parties but within the existing parliamentary system, and Marx's belief that the proletariat, after it had overthrown the bourgeois state, should establish its own regime. To Bakunin, the mission of the revolutionary was destruction; he looked to the Russian peasantry, with its propensities for violence and its uncurbed revolutionary instincts, rather than to the effete, civilized workers of the industrial countries. The students, he hoped, would be the officers of the revolution. He acquired followers, mostly young men, in Italy, Switzerland, and France, and he organized a secret society, the International Alliance of Social Democracy, which in 1869 challenged the hegemony of the General Council at the congress in Basel, Switz. Marx, however, had already succeeded in preventing its admission as an organized body into the International.

To the Bakuninists, the Paris Commune was a model of revolutionary direct action and a refutation of what they considered to be Marx's "authoritarian communism." Bakunin began organizing sections of the International for an attack on the alleged dictatorship of Marx and the General Council. Marx in reply publicized Bakunin's embroilment with an unscrupulous Russian student leader, Sergey Gennadiyevich Nechayev, who had practiced blackmail and murder.

Without a supporting right wing and with the anarchist left against him, Marx feared losing control of the International to Bakunin. He also wanted to return to his studies and to finish *Das Kapital.* At the congress of the International at The Hague in 1872, the only one he ever attended, Marx managed to defeat the Bakuninists. Then, to the consternation of the delegates, Engels moved that the seat of the General Council be transferred from London to New York City. The Bakuninists were expelled, but the International languished and was finally disbanded in Philadelphia in 1876.

Dissolution of the International

LAST YEARS

During the next and last decade of his life, Marx's creative energies declined. He was beset by what he called "chronic mental depression," and his life turned inward toward his family. He was unable to complete any substantial work, though he still read widely and undertook to learn Russian. He became crotchety in his political opinions. When his own followers and those of the German revolutionary Ferdinand Lassalle, a rival who believed that socialist goals should be achieved through cooperation with the state, coalesced in 1875 to found the German Social Democratic Party, Marx wrote a caustic criticism of their program (the so-called Gotha Program), claiming that it made too many compromises with the status quo. The German leaders put his objections aside and tried to mollify him personally. Increasingly, he looked to a European war for the overthrow of Russian tsarism, the mainstay of reaction, hoping that this would revive the political energies of

the working classes. He was moved by what he considered to be the selfless courage of the Russian terrorists who assassinated the tsar, Alexander II, in 1881; he felt this to be "a historically inevitable means of action."

Despite Marx's withdrawal from active politics, he still retained what Engels called his "peculiar influence" on the leaders of working-class and socialist movements. In 1879, when the French Socialist Workers' Federation was founded, its leader Jules Guesde went to London to consult with Marx, who dictated the preamble of its program and shaped much of its content. In 1881 Henry Mayers Hyndman in his *England for All* drew heavily on his conversations with Marx but angered him by being afraid to acknowledge him by name.

Death

During his last years Marx spent much time at health resorts and even traveled to Algiers. He was broken by the death of his wife on Dec. 2, 1881, and of his eldest daughter, Jenny Longuet, on Jan. 11, 1883. He died in London, evidently of a lung abscess, on March 14, 1883.

CHARACTER AND SIGNIFICANCE

At Marx's funeral in Highgate Cemetery, Engels declared that Marx had made two great discoveries, the law of development of human history and the law of motion of bourgeois society. But "Marx was before all else a revolutionist." He was "the best-hated and most-calumniated man of his time," yet he also died "beloved, revered and mourned by millions of revolutionary fellow-workers."

The contradictory emotions Marx engendered are reflected in the sometimes conflicting aspects of his character. Marx was a combination of the Promethean rebel and the rigorous intellectual. He gave most persons an impression of intellectual arrogance. A Russian writer, Pavel Annenkov, who observed Marx in debate in 1846 recalled that "he spoke only in the imperative, brooking no contradiction," and seemed to be "the personification of a democratic dictator such as might appear before one in moments of fantasy." But Marx obviously felt uneasy before mass audiences and avoided the atmosphere of factional controversies at congresses. He went to no demonstrations, his wife remarked, and rarely spoke at public meetings. He kept away from the congresses of the International where the rival socialist groups debated important resolutions. He was a "small groups" man, most at home in the atmosphere of the General Council or on the staff of a newspaper, where his character could impress itself forcefully on a small body of coworkers. At the same time he avoided meeting distinguished scholars with whom he might have discussed questions of economics and sociology on a footing of intellectual equality. Despite his broad intellectual sweep, he was prey to obsessive ideas such as that the British foreign minister, Lord Palmerston, was an agent of the Russian government. He was determined not to let bourgeois society make "a money-making machine" out of him, yet he submitted to living on the largess of Engels and the bequests of relatives. He remained the eternal student in his personal habits and way of life, even to the point of joining two friends in a students' prank during which they systematically broke four or five streetlamps in a London street and then fled from the police. He was a great reader of novels, especially those of Sir Walter Scott and Balzac; and the family made a cult of Shakespeare. He was an affectionate father, saying that he admired Jesus for his love of children, but sacrificed the lives and health of his own. Of his seven children, three daughters grew to maturity. His favourite daughter, Eleanor, worried him with her nervous, brooding, emotional character and her desire to be an actress. Another shadow was cast on Marx's domestic life by the birth to their loyal servant, Helene Demuth, of an illegitimate son, Frederick; Engels as he was dying disclosed to Eleanor that Marx had been the father. Above all, Marx was a fighter, willing to sacrifice anything in the battle for his conception of a better society. He regarded struggle as the law of life and existence.

Das Kapital

The influence of Marx's ideas has been enormous. Marx's masterpiece, *Das Kapital*, the "Bible of the working class," as it was officially described in a resolution of the International Working Men's Association, was published in 1867 in Berlin and received a second edition in 1873. Only the first volume was completed and published in Marx's lifetime. The second and third volumes, unfinished by Marx, were edited by Engels and published in 1885 and 1894. The economic categories he employed were those of the classical British economics of David Ricardo; but Marx used them in accordance with his dialectical method to argue that bourgeois society, like every social organism, must follow its inevitable path of development. Through the working of such immanent tendencies as the declining rate of profit, capitalism would die and be replaced by another, higher, society. The most memorable pages in *Das Kapital* are the descriptive passages, culled from Parliamentary Blue Books, on the misery of the English working class. Marx believed that this misery would increase, while at the same time the monopoly of capital would become a fetter upon production until finally "the knell of capitalist private property sounds. The expropriators are expropriated."

Marx never claimed to have discovered the existence of classes and class struggles in modern society. "Bourgeois" historians, he acknowledged, had described them long before he had. He did claim, however, to have proved that each phase in the development of production was associated with a corresponding class structure and that the struggle of classes led necessarily to the dictatorship of the proletariat, ushering in the advent of a classless society. Marx took up the very different versions of socialism current in the early 19th century and welded them together into a doctrine that continued to be the dominant version of socialism for half a century after his death. His emphasis on the influence of economic structure on historical development has proved to be of lasting significance.

The dialectical model

Although Marx stressed economic issues in his writings, his major impact has been in the fields of sociology and history. Marx's most important contribution to sociological theory was his general mode of analysis, the "dialectical" model, which regards every social system as having within it immanent forces that give rise to "contradictions" (disequilibria) that can be resolved only by a new social system. Neo-Marxists, who no longer accept the economic reasoning in *Das Kapital,* are still guided by this model in their approach to capitalist society. In this sense, Marx's mode of analysis, like those of Thomas Malthus, Herbert Spencer, or Vilfredo Pareto, has become one of the theoretical structures that are the heritage of the social scientist. (L.S.F./D.T.McL.)

Marxism

The term Marxism is used in a number of different ways. In its most essential meaning it refers to the thought of Karl Marx but is usually extended to include that of his friend and collaborator Friedrich Engels. There is also Marxism as it has been understood and practiced by the various socialist movements, particularly before 1914. Then there is Soviet Marxism as worked out by Lenin and modified by Stalin, which under the name of Marxism-Leninism became the doctrine of the communist parties set up after the Russian Revolution. Offshoots of this include Marxism as interpreted by the anti-Stalinist Leon Trotsky and his followers, Mao Zedong's (Mao Tse-tung's) Chinese variant of Marxism-Leninism, and various Third World Marxisms. There are also the post-World War II nondogmatic Marxisms that have modified Marx's thought with borrowings from modern philosophies, principally from those of Edmund Husserl and Martin Heidegger but also from Sigmund Freud and others.

THE THOUGHT OF KARL MARX

The written work of Marx cannot be reduced to a philosophy, much less to a philosophical system. The whole of his work is a radical critique of philosophy, especially of Hegel's idealist system and of the philosophies of the left and right post-Hegelians. It is not, however, a mere denial of those philosophies. Marx declared that philosophy must become reality. One could no longer be content with interpreting the world; one must be concerned with transforming it, which meant transforming both the world

itself and men's consciousness of it. This, in turn, required a critique of experience together with a critique of ideas. In fact, Marx believed that all knowledge involved a critique of ideas. He was not an empiricist. On the contrary, his work teems with concepts (appropriation, alienation, praxis, creative labour, value, etc.) that he had inherited from earlier philosophers and economists, including Hegel, Johann Fichte, Kant, Adam Smith, David Ricardo, and John Stuart Mill. What uniquely characterizes the thought of Marx is that, instead of making abstract affirmations about a whole group of problems such as man, knowledge, matter, nature, and so forth, he examines each problem in its dynamic relation to the others and, above all, tries to relate them to historical, social, political, and economic realities.

Historical materialism. In 1859, in the preface to his *Contribution to the Critique of Political Economy,* Marx wrote that the hypothesis that had served him as the basis for his analysis of society could be briefly formulated as follows:

Economic foundation of society

> In the social production that men carry on, they enter into definite relations that are indispensable and independent of their will, relations of production which correspond to a definite stage of development of their material forces of production. The sum total of these relations of production constitutes the economic structure of society, the real foundation, on which rises a legal and political superstructure, and to which correspond definite forms of social consciousness. The mode of production in material life determines the general character of the social, political, and intellectual processes of life. It is not the consciousness of men which determines their existence; it is on the contrary their social existence which determines their consciousness.

Raised to the level of historical law, this hypothesis was subsequently called historical materialism. Marx applied it to capitalist society, both in *The Communist Manifesto* and *Das Kapital* and in other writings. Although Marx reflected upon his working hypothesis for many years, he did not formulate it in a very exact manner: different expressions served him for identical realities. If one takes the text literally, social reality is structured in the following way:

1. Underlying everything as the real basis of society is the economic structure (what in late 20th-century language is sometimes called the infrastructure). This structure includes (a) the "material forces of production," that is, the labour and means of production, and (b) the overall "relations of production," or the social and political arrangements that regulate production and distribution. Although Marx stated that there is a correspondence between the "material forces" of production and the indispensable "relations" of production, he never made himself clear on the nature of the correspondence, a fact that was to be the source of differing interpretations among his later followers.

2. Above the economic structure rises the superstructure consisting of legal and political "forms of social consciousness" that correspond to the economic structure. Marx says nothing about the nature of this correspondence between ideological forms and economic structure, except that through the ideological forms men become conscious of the conflict within the economic structure between the material forces of production and the existing relations of production expressed in the legal property relations. In other words, "The sum total of the forces of production accessible to men determines the condition of society" and is at the base of society. "The social structure and the state issue continually from the life processes of definite individuals . . . as they are *in reality,* that is acting and materially producing." The political relations that men establish among themselves are dependent on material production, as are the legal relations. This foundation of the social on the economic is not an incidental point: it colours Marx's whole analysis. It is found in *Das Kapital* as well as in *The German Ideology* and the *Economic and Philosophic Manuscripts of 1844.*

Analysis of society. To go directly to the heart of the work of Marx, one must focus on his concrete program for man. This is just as important for an understanding of Marx as are *The Communist Manifesto* and *Das Kapital.* Marx's interpretation of man begins with human need. "Man," he wrote in the *Economic and Philosophic Manuscripts of 1844,*

Man as a natural being

> is first of all a *natural being.* As a natural being and a living natural being, he is endowed on the one hand with *natural powers, vital powers* . . . ; these powers exist in him as aptitudes, instincts. On the other hand, as an objective, natural, physical, sensitive being, he is a *suffering,* dependent and limited being . . . , that is, the *objects* of his instincts exist outside him, independent of him, but are the objects of his *need,* indispensable and essential for the realization and confirmation of his substantial powers.

The point of departure of human history is therefore living man, who seeks to satisfy certain primary needs. "The first historical fact is the production of the means to satisfy these needs." This satisfaction, in turn, opens the way for new needs. Human activity is thus essentially a struggle with nature that must furnish man with the means of satisfying his needs: drink, food, clothing, the development of his powers and then of his intellectual and artistic abilities. In this undertaking, man discovers himself as a productive being who humanizes himself by his labour. Furthermore, man humanizes nature while he naturalizes himself. By his creative activity, by his labour, he realizes his identity with the nature that he masters, while at the same time he achieves free consciousness. Born of nature man becomes fully human by opposing it. Becoming aware in his struggle against nature of what separates him from it, man finds the conditions of his fulfillment, of the realization of his true stature. The dawning of consciousness is inseparable from struggle. By appropriating all the creative energies, he discovers that "all that is called history is nothing else than the process of creating man through human labour, the becoming of nature for man. Man has thus evident and irrefutable proof of his own creation by himself." Understood in its universal dimension, human activity reveals that "for man, man is the supreme being." It is thus vain to speak of God, creation, and metaphysical problems. Fully naturalized, man is sufficient unto himself: he has recaptured the fullness of man in his full liberty.

Man as an alienated being

Living in a capitalist society, however, man is not truly free. He is an alienated being; he is not at home in his world. The idea of alienation, which Marx takes from Hegel and Feuerbach, plays a fundamental role in the whole of his written work, starting with the writings of his youth and continuing through *Das Kapital.* In the *Economic and Philosophic Manuscripts* the alienation of labour is seen to spring from the fact that the more the worker produces the less he has to consume, and the more values he creates the more he devalues himself, because his product and his labour are estranged from him. The life of the worker depends on things that he has created but that are not his, so that, instead of finding his rightful existence through his labour, he loses it in this world of things that are external to him: no work, no pay. Under these conditions, labour denies the fullness of concrete man. "The generic being (*Gattungwesen*) of man, nature as well as his intellectual faculties, is transformed into a being which is alien to him, into a *means of his individual existence.*" Nature, his body, his spiritual essence become alien to him. "Man is made alien to man." When carried to its highest stage of development, private property becomes "the product of alienated labour . . . the *means* by which labour alienates itself (and) the realization of this alienation." It is also at the same time "the tangible material expression of *alienated human* life."

Although there is no evidence that Marx ever disclaimed this anthropological analysis of alienated labour, starting with *The German Ideology,* the historical, social, and economic causes of the alienation of labour are given increasing emphasis, especially in *Das Kapital.* Alienated labour is seen as the consequence of market product, the division of labour, and the division of society into antagonistic classes. As producers in society, men create goods only by their labour. These goods are exchangeable. Their value is the average amount of social labour spent to produce them. The alienation of the worker takes on its full dimension in that system of market production in which part of the

Economic alienation

value of the goods produced by the worker is taken away from him and transformed into surplus value, which the capitalist privately appropriates. Market production also intensifies the alienation of labour by encouraging specialization, piecework, and the setting up of large enterprises. Thus the labour power of the worker is used along with that of others in a combination whose significance he is ignorant of, both individually and socially. In thus losing their quality as human products, the products of labour become fetishes, that is, alien and oppressive realities to which both the man who possesses them privately and the man who is deprived of them submit themselves. In the market economy, this submission to things is obscured by the fact that the exchange of goods is expressed in money.

This fundamental economic alienation is accompanied by secondary political and ideological alienations, which offer a distorted representation of and an illusory justification of a world in which the relations of men with one another are also distorted. The ideas that men form are closely bound up with their material activity and their material relations: "The act of making representations, of thinking, the spiritual intercourse of men, seem to be the direct emanation of their material relations." This is true of all human activity: political, intellectual, or spiritual. "Men produce their representations and their ideas, but it is as living men, men acting as they are determined by a definite development of their powers of production." Law, morality, metaphysics, and religion do not have a history of their own. "Men developing their material production modify together with their real existence their ways of thinking and the products of their ways of thinking." In other words, "It is not consciousness which determines existence, it is existence which determines consciousness."

In bourgeois, capitalist society man is divided into political citizen and economic man. This duality represents man's political alienation, which is further intensified by the functioning of the bourgeois state. From this study of society at the beginning of the 19th century, Marx came to see the state as the instrument through which the propertied class dominated other classes.

Ideological alienation, for Marx, takes different forms, appearing in economic, philosophical, and legal theories. Marx undertook a lengthy critique of the first in *Das Kapital* and of the second in *The German Ideology*. But ideological alienation expresses itself supremely in religion. Taking up the ideas about religion that were current in left post-Hegelian circles, together with the thought of Feuerbach, Marx considered religion to be a product of man's consciousness. It is a reflection of the situation of a man who "either has not conquered himself or has already lost himself again" (man in the world of private property). It is "an opium for the people." Unlike Feuerbach, Marx believed that religion would disappear only with changes in society.

Analysis of the economy. Marx analyzed the market economy system in *Das Kapital*. In this work he borrows most of the categories of the classical English economists Smith and Ricardo but adapts them and introduces new concepts such as that of surplus value. One of the distinguishing marks of *Das Kapital* is that in it Marx studies the economy as a whole and not in one or another of its aspects. His analysis is based on the idea that man is a productive being and that all economic value comes from human labour. The system he analyzes is principally that of mid-19th-century England. It is a system of private enterprise and competition that arose in the 16th century from the development of sea routes, international trade, and colonialism. Its rise had been facilitated by changes in the forces of production (the division of labour and the concentration of workshops), the adoption of mechanization, and technical progress. The wealth of the societies that brought this economy into play had been acquired through an "enormous accumulation of commodities." Marx therefore begins with the study of this accumulation, analyzing the unequal exchanges that take place in the market.

Theory of surplus value

According to Marx, if the capitalist advances funds to buy cotton yarn with which to produce fabrics and sells the product for a larger sum than he paid, he is able to invest the difference in additional production. "Not only is the value advance kept in circulation, but it changes in its magnitude, adds a plus to itself, makes itself worth more, and it is this movement that transforms it into capital." The transformation, to Marx, is possible only because the capitalist has appropriated the means of production, including the labour power of the worker. Now labour power produces more than it is worth. The value of labour power is determined by the amount of labour necessary for its reproduction or, in other words, by the amount needed for the worker to subsist and beget children. But in the hands of the capitalist the labour power employed in the course of a day produces more than the value of the sustenance required by the worker and his family. The difference between the two values is appropriated by the capitalist, and it corresponds exactly to the surplus value realized by capitalists in the market. Marx is not concerned with whether in capitalist society there are sources of surplus value other than the exploitation of human labour—a fact pointed out by Joseph Schumpeter (*Capitalism, Socialism, and Democracy*). He remains content with emphasizing this primary source:

> Surplus value is produced by the employment of labour power. Capital buys the labour power and pays the wages for it. By means of his work the labourer creates new value which does not belong to him, but to the capitalist. He must work a certain time merely in order to reproduce the equivalent value of his wages. But when this equivalent value has been returned, he does not cease work, but continues to do so for some further hours. The new value which he produces during this extra time, and which exceeds in consequence the amount of his wage, constitutes surplus value.

The crises of capitalism

Throughout his analysis, Marx argues that the development of capitalism is accompanied by increasing contradictions. For example, the introduction of machinery is profitable to the individual capitalist because it enables him to produce more goods at a lower cost, but new techniques are soon taken up by his competitors. The outlay for machinery grows faster than the outlay for wages. Since only labour can produce the surplus value from which profit is derived, this means that the capitalist's rate of profit on his total outlay tends to decline. Along with the declining rate of profit goes an increase in unemployment. Thus, the equilibrium of the system is precarious, subject as it is to the internal pressures resulting from its own development. Crises shake it at regular intervals, preludes to the general crisis that will sweep it away. This instability is increased by the formation of a reserve army of workers, both factory workers and peasants, whose pauperization keeps increasing. "Capitalist production develops the technique and the combination of the process of social production only by exhausting at the same time the two sources from which all wealth springs: the earth and the worker." According to the Marxist dialectic, these fundamental contradictions can only be resolved by a change from capitalism to a new system.

Class struggle. Marx inherited the ideas of class and class struggle from Utopian socialism and the theories of Saint-Simon. These had been given substance by the writings of French historians such as Adolphe Thiers and François Guizot on the French Revolution of 1789. But unlike the French historians, Marx made class struggle the central fact of social evolution. "The history of all hitherto existing human society is the history of class struggles."

In Marx's view, the dialectical nature of history is expressed in class struggle. With the development of capitalism, the class struggle takes an acute form. Two basic classes, around which other less important classes are grouped, oppose each other in the capitalist system: the owners of the means of production, or bourgeoisie, and the workers, or proletariat. "The bourgeoisie produces its own grave-diggers. The fall of the bourgeoisie and the victory of the proletariat are equally inevitable" (*The Communist Manifesto*) because

> the bourgeois relations of production are the last contradictory form of the process of social production, contradictory not in the sense of an individual contradiction, but of a contradiction that is born of the conditions of social existence of individuals; however, the forces of production which develop

in the midst of bourgeois society create at the same time the material conditions for resolving this contradiction. With this social development the prehistory of human society ends.

When man has become aware of his loss, of his alienation, as a universal nonhuman situation, it will be possible for him to proceed to a radical transformation of his situation by a revolution. This revolution will be the prelude to the establishment of communism and the reign of liberty reconquered. "In the place of the old bourgeois society with its classes and its class antagonisms, there will be an association in which the free development of each is the condition for the free development of all."

Two views of revolution

But for Marx there are two views of revolution. One is that of a final conflagration, "a *violent* suppression of the old conditions of production," which occurs when the opposition between bourgeoisie and proletariat has been carried to its extreme point. This conception is set forth in a manner inspired by the Hegelian dialectic of the master and the slave, in *The Holy Family.* The other conception is that of a permanent revolution involving a provisional coalition between the proletariat and the petty bourgeoisie rebelling against a capitalism that is only superficially united. Once a majority has been won to the coalition, an unofficial proletarian authority constitutes itself alongside the revolutionary bourgeois authority. Its mission is the political and revolutionary education of the proletariat, gradually assuring the transfer of legal power from the revolutionary bourgeoisie to the revolutionary proletariat.

If one reads *The Communist Manifesto* carefully one discovers inconsistencies that indicate that Marx had not reconciled the concepts of catastrophic and of permanent revolution. Moreover, Marx never analyzed classes as specific groups of men opposing other groups of men. Depending on the writings and the periods, the number of classes varies; and unfortunately the pen fell from Marx's hand at the moment when, in *Das Kapital* (vol. 3), he was about to take up the question. Reading *Das Kapital,* one is furthermore left with an ambiguous impression with regard to the destruction of capitalism: will it be the result of the "general crisis" that Marx expects, or of the action of the conscious proletariat, or of both at once?

The contributions of Engels. Engels became a communist in 1842 and discovered the proletariat of England when he took over the management of the Manchester factory belonging to his father's cotton firm. In 1844, the year he began his close association and friendship with Marx, Engels was finishing his "Umrisse zu einer Kritik der Nationalökonomie" ("Outline of a Critique of Political Economy")—a critique of Smith, Ricardo, Mill, and J.B. Say. This remarkable study contained in seminal form the critique that Marx was to make of bourgeois political economy in *Das Kapital.* During the first years of his stay in Manchester, Engels observed carefully the life of the workers of that great industrial centre and described it in *Die Lage der arbeitenden Klassen in England* (*The Condition of the Working Class in England*), published in 1845 in Leipzig. This work was an analysis of the evolution of industrial capitalism and its social consequences. He collaborated with Marx in the writing of *The Holy Family, The German Ideology,* and *The Communist Manifesto.* The correspondence between them is of fundamental importance for the student of *Das Kapital,* for it shows how Engels contributed by furnishing Marx with a great amount of technical and economic data and by criticizing the successive drafts. This collaboration lasted until Marx's death and was carried on posthumously with the publication of the manuscripts left by Marx, which Engels edited, forming volumes 2 and 3 of *Das Kapital.* He also wrote various articles on Marx's work.

In response to criticism of Marx's ideas by a socialist named Eugen Dühring, Engels published several articles that were collected under the title *Herr Eugen Dührings Umwälzung der Wissenschaft,* which appeared in 1878 (*Herr Eugen Dühring's Revolution in Science* [*Anti-Dühring*]), and an unfinished work, *Dialektik und Natur* (1927; *Dialectics of Nature*), which he had begun around 1875–76. The importance of these writings to the subsequent development of Marxism can be seen from Lenin's observation that Engels "developed, in a clear and often polemical style, the most general scientific questions and the different phenomena of the past and present according to the materialist understanding of history and the economic theory of Karl Marx." But Engels was driven to simplify problems with a view to being pedagogical; he tended to schematize and systematize things as if the fundamental questions were settled. The connections that he thus established between some of Marx's governing ideas and some of the scientific ideas of his age gave rise to the notion that there is a complete Marxist philosophy. The idea was to play a significant role in the transition of Marxism from a "critique of daily life" to an integrated doctrine in which philosophy, history, and the sciences are fused.

Engels' critique of Dühring

Anti-Dühring is of fundamental importance for it constitutes the link between Marx and certain forms of modern Marxism. It contains three parts: Philosophy, Political Economy, and Socialism. In the first, Engels attempts to establish that the natural sciences and even mathematics are dialectical, in the sense that observable reality is dialectical: the dialectical method of analysis and thought is imposed on men by the material forces with which they deal. It is thus rightly applied to the study of history and human society. "Motion, in effect, is the mode of existence of matter," Engels writes. In using materialistic dialectic to make a critique of Dühring's thesis, according to which political forces prevail over all the rest in the molding of history, Engels provides a good illustration of the materialistic idea of history, which puts the stress on the prime role of economic factors as driving forces in history. The other chapters of the section Political Economy form a very readable introduction to the principal economic ideas of Marx: value (simple and complex), labour, capital, and surplus value. The section Socialism starts by formulating anew the critique of the capitalist system as it was made in *Das Kapital.* At the end of the chapters devoted to production, distribution, the state, the family, and education, Engels outlines what the socialist society will be like, a society in which the notion of value has no longer anything to do with the distribution of the goods produced because all labour "becomes at once and directly social labour," and the amount of social labour that every product contains no longer needs to be ascertained by "a detour." A production plan will coordinate the economy. The division of labour and the separation of town and country will disappear with the "suppression of the capitalist character of modern industry." Thanks to the plan, industry will be located throughout the country in the collective interest, and thus the opposition between town and country will disappear—to the profit of both industry and agriculture. Finally, after the liberation of man from the condition of servitude in which the capitalist mode of production holds him, the state will also be abolished and religion will disappear by "natural death."

One of the most remarkable features of *Anti-Dühring* is the insistence with which Engels refuses to base socialism on absolute values. He admits only relative values, linked to historical, economic, and social conditions. Socialism cannot possibly be based on ethical principles: each epoch can only successfully carry out that of which it is capable. Marx had written this in his preface of 1859.

GERMAN MARXISM AFTER ENGELS

The work of Kautsky and Bernstein. The theoretical leadership after Engels was taken by Karl Kautsky, editor of the official organ of the German Social Democratic Party, *Die Neue Zeit.* He wrote *Karl Marx' ökonomische Lehren* (1887; *The Economic Doctrines of Karl Marx*), in which the work of Marx is presented as essentially an economic theory. Kautsky reduced the ideas of Marx and Marxist historical dialectic to a kind of evolutionism. He laid stress on the increasing pauperization of the working class and on the increasing degree of capitalist concentration. While opposing all compromise with the bourgeois state, he accepted the contention that the socialist movement should support laws benefiting the workers provided that they did not reinforce the power of the state. Rejecting the idea of an alliance between the working class and the peasantry, he believed that the overthrow of the cap-

italist state and the acquisition of political power by the working class could be realized in a peaceful way, without upsetting the existing structures. As an internationalist he supported peace, rejecting war and violence. For him, war was a product of capitalism. Such were the main features of "orthodox" German Marxism at the time when the "revisionist" theories of Eduard Bernstein appeared.

Doctrinal disputes

Bernstein created a great controversy with articles that he wrote in 1896 for *Die Neue Zeit,* arguing that Marxism needed to be revised. His divergence widened with the publication in 1899 of *Die Voraussetzungen des Sozialismus und die Aufgaben der Sozialdemokratie* (*Evolutionary Socialism*), to which rejoinders were made by Kautsky in *Bernstein und das Sozialdemokratische Programm: Eine Antikritik* (1899; "Bernstein and the Social Democratic Program") and the Polish-born Marxist Rosa Luxemburg in *Sozialreform oder Revolution* (*Reform or Revolution*), both in 1899. Bernstein focused first of all upon the labour theory of value. Along with the economists of his time he considered it outdated, both in the form expounded by British classical economists and as set forth in *Das Kapital.* He argued, moreover, that class struggle was becoming less rather than more intense, for concentration was not accelerating in industry as Marx had forecast, and in agriculture it was not increasing at all. Bernstein demonstrated this on the basis of German, Dutch, and English statistical data. He also argued that cartels and business syndicates were smoothing the evolution of capitalism, a fact that cast doubt on the validity of Marx's theory of capitalistic crises. Arguing that quite a few of Marx's theories were not scientifically based, Bernstein blamed the Hegelian and Ricardian structure of Marx's work for his failure to take sufficient account of observable reality.

To this, Kautsky replied that, with the development of capitalism, agriculture was becoming a sector more and more dependent on industry, and that in addition an industrialization of agriculture was taking place. Luxemburg took the position that the contradictions of capitalism did not cease to grow with the progress of finance capitalism and the exploitation of the colonies, and that these contradictions were leading to a war that would give the proletariat its opportunity to assume power by revolutionary means.

The radicals. One of the most divisive questions was that of war and peace. This was brought to the fore at the outbreak of World War I, when Social Democratic deputies in the German Reichstag voted for the financing of the war. Among German Marxists who opposed the war were Karl Liebknecht and Luxemburg. Liebknecht was imprisoned in 1916 for agitating against the war. On his release in 1918 he took the leadership of the Spartacist movement, which was later to become the Communist Party of Germany. Luxemburg had also been arrested for her antimilitary activities. In addition to her articles, signed Junius, in which she debated with Lenin on the subject of the First World War and the attitude of the Marxists toward it (published in 1916 as *Die Krise der Sozialdemokratie* [*The Crisis in the German Social-Democracy*]), she is known for her book *Die Akkumulation des Kapitals* (1913; *The Accumulation of Capital*). In this work she returned to Marx's economic analysis of capitalism, in particular the accumulation of capital as expounded in volume 2 of *Das Kapital.* There she found a contradiction that had until then been unnoticed: Marx's scheme seems to imply that the development of capitalism can be indefinite, though elsewhere he sees the contradictions of the system as bringing about increasingly violent economic crises that will inevitably sweep capitalism away. Luxemburg concluded that Marx's scheme is oversimplified and assumes a universe made up entirely of capitalists and workers. If increases in productivity are taken into account, she asserted, balance between the two sectors becomes impossible; in order to keep expanding, capitalists must find new markets in noncapitalist spheres, either among peasants and artisans or in colonies and underdeveloped countries. Capitalism will collapse only when exploitation of the world outside it (the peasantry, colonies, etc.) has reached a limit. This conclusion has been the subject of passionate controversies.

Karl Liebknecht and Rosa Luxemburg

The Austrians. The Austrian school came into being when Austrian socialists started publishing their works independently of the Germans; it can be dated from either 1904 (beginning of the *Marx-Studien* collection) or 1907 (publication of the magazine *Der Kampf*). The most important members of the school were Max Adler, Karl Renner, Rudolf Hilferding, Gustav Eckstein, Friedrich Adler, and Otto Bauer. The most eminent was Bauer, a brilliant theoretician whose *Die Nationalitätenfrage und die Sozialdemokratie* (1906; "The Nationalities Question and the Social Democracy") was critically reviewed by Lenin. In this work he dealt with the problem of nationalities in the light of the experience of the Austro-Hungarian Empire. He favoured the self-determination of peoples and emphasized the cultural elements in the concept of nationhood. Hilferding was finance minister of the German Republic after World War I in the Cabinets of the Social Democrats Gustav Stresemann (1923) and Hermann Müller (1928). He is known especially for his work *Das Finanzkapital* (1910), in which he maintained that capitalism had come under the control of banks and industrial monopolies. The growth of national competition and tariff barriers, he believed, had led to economic warfare abroad. Hilferding's ideas strongly influenced Lenin, who analyzed them in *Imperialism, the Highest Stage of Capitalism* (1916).

RUSSIAN AND SOVIET MARXISM

Lenin's adaptation of Marxism

Das Kapital was translated into Russian in 1872. Marx kept up more or less steady relations with the Russian socialists and took an interest in the economic and social conditions of the tsarist empire. The man who actually introduced Marxism into Russia was Georgi Plekhanov, but the man who adapted Marxism to Russian conditions was Lenin.

Lenin. Vladimir Ilich Ulyanov, or Lenin, was born in 1870 at Simbirsk (Ulyanovsk today). He entered the University of Kazan to study law but was expelled the same year for participating in student agitation. In 1893 he settled in St. Petersburg and became actively involved with the revolutionary workers. With his pamphlet *What Is To Be Done?* (1902), he specified the theoretical principles and organization of a Marxist party as he thought it should be constituted. He took part in the second Congress of the Russian Social-Democratic Workers' Party, which was held in Brussels and London (1903), and induced the majority of the Congress members to adopt his views. Two factions formed at the Congress: the Bolshevik (from the Russian word for "larger") with Lenin as the leader and the Menshevik (from the Russian word for "smaller") with Julius Martov at the head. The former wanted a restricted party of militants and advocated the dictatorship of the proletariat. The latter wanted a wide-open proletarian party, collaboration with the liberals, and a democratic constitution for Russia. In his pamphlet *One Step Forward, Two Steps Back* (1904), Lenin compared the organizational principles of the Bolsheviks to those of the Mensheviks. After the failure of the 1905 Russian revolution, he drew positive lessons for the future in *Two Tactics of Social Democracy in the Democratic Revolution.* He fiercely attacked the influence of Kantian philosophy on German and Russian Marxism in *Materialism and Empirio-criticism* (1908). In 1912 at the Prague Conference the Bolsheviks constituted themselves as an independent party. During World War I Lenin resided in Switzerland, where he studied Hegel's *Science of Logic* and the development of capitalism and carried on debates with Marxists like Rosa Luxemburg on the meaning of the war and the right of nations to self-determination. In 1915 at Zimmerwald, and in 1916 at Kiental, he organized two international socialist conferences to fight against the war. Immediately after the February 1917 revolution he returned to Russia, and in October the Bolshevik coup brought him to power.

The situation of Russia and the Russian revolutionary movement at the end of the 19th century and the beginning of the 20th led Lenin to diverge, in the course of his development and his analyses, from the positions both of "orthodox Marxism" and of "revisionism." He rediscov-

ered the original thought of Marx by a careful study of his works, in particular *Das Kapital* and *The Holy Family*. He saw Marxism as a practical affair and tried to go beyond the accepted formulas to plan political action that would come to grips with the surrounding world.

A plan of political action

As early as 1894, in his populist study *The Friends of the People,* Lenin took up Marx's distinction between the "material social relations" of men and their "ideological social relations." In Lenin's eyes the importance of *Das Kapital* was that "while *explaining* the structure and the development of the social formation seen *exclusively* in terms of its relations of production, (Marx) has nevertheless everywhere and always analyzed the superstructure which corresponds to these relations of production." In *The Development of Russian Capitalism* (1897–99) Lenin sought to apply Marx's analysis by showing the growing role of capital, in particular commercial capital—the exploitation of the workers in the factories, and the large-scale expropriation of the peasants. It was thus possible to apply to Russia the models developed by Marx for western Europe. At the same time Lenin did not lose sight of the importance of the peasant in Russian society. Although a disciple of Marx, he did not believe that he had only to repeat Marx's conclusions. He wrote:

> We do not consider the theory of Marx to be a complete, immutable whole. We think on the contrary that this theory has only laid the cornerstone of the science, a science which socialists must further develop in all directions if they do not want to let themselves be overtaken by life. We think that, for the Russian socialists, an independent elaboration of the theory is particularly necessary.

Lenin laid great stress upon the dialectical method. In his early writings he defined the dialectic as "nothing more nor less than the method of sociology, which sees society as a living organism, in perpetual development (and not as something mechanically assembled and thus allowing all sorts of arbitrary combinations of the various social elements)..." (*The Friends of the People,* 1894). After having studied Hegel toward the end of 1914, he took a more activist view. Dialectic is not only evolution; it is praxis, leading from activity to reflection and from reflection to action.

The dictatorship of the proletariat. Lenin also put much emphasis on the leading role of the party. As early as 1902 he was concerned with the need for a cohesive party with a correct doctrine, adapted to the exigencies of the period, which would be a motive force among the masses, helping to bring them to an awareness of their real situation. In *What Is To Be Done?* he called for a party of professional revolutionaries, disciplined and directed, capable of defeating the police; its aim should be to establish the dictatorship of the proletariat. In order to do this, he wrote in *Two Tactics of Social-Democracy in the Democratic Revolution,* it was necessary "to subject the insurrection of the proletarian and *non-proletarian* masses to our influence, to our direction, to use it in our best interests." But this was not possible without a doctrine: "Without revolutionary theory, no revolutionary movement." On the eve of the revolution of October 1917, in *The State and Revolution* he set forth the conditions for the dictatorship of the proletariat and the suppression of the capitalist state.

Role of the peasantry

Lenin assigned major importance to the peasantry in formulating his program. It would be a serious error, he held, for the Russian revolutionary workers' movement to neglect the peasants. Even though it was clear that the industrial proletariat constituted the vanguard of the revolution, the discontent of the peasantry could be oriented in a direction favourable to the revolution by placing among the goals of the party the seizure of privately owned land. As early as 1903, at the third congress of the party, he secured a resolution to this effect. Thereafter, the dictatorship of the proletariat became the dictatorship of the proletariat and the peasantry. In 1917 he encouraged the peasants to seize land long before the approval of agrarian reform by the Constituent Assembly.

Among Lenin's legacies to Soviet Marxism was one that proved to be injurious to the party. This was the decision taken at his behest by the 10th congress of the party in the spring of 1921, while the sailors were rebelling at Kronstadt and the peasants were growing restless in the countryside, to forbid all factions, all factional activity, and all opposition political platforms within the party. This decision had grave consequences in later years when Stalin used it against his opponents.

Stalin. It is Joseph Stalin who codified the body of ideas that, under the name of Marxism-Leninism, has constituted the official doctrine of the Soviet and eastern European communist parties. Stalin was a man of action in a slightly different sense than was Lenin. Gradually taking over power after Lenin's death in 1924, he pursued the development of the Soviet Union with great vigour. By practicing Marxism, he assimilated it, at the same time simplifying it. Stalin's Marxism-Leninism rests on the dialectic of Hegel, as set forth in *A Short History of the Communist Party of the Soviet Union* (1938), and on a materialism that can be considered roughly identical to that of Feuerbach. His work *Problems of Leninism,* which appeared in 11 editions during his lifetime, sets forth an ideology of power and activism that rides roughshod over the more nuanced approach of Lenin.

Stalin and dialectical materialism

Soviet dialectical materialism can be reduced to four laws: (1) History is a dialectical development. It proceeds by successive phases which supersede one another. These phases are not separate, any more than birth, growth, and death are separate. Though it is true that phase B necessarily negates phase A, it remains that phase B was already contained in phase A and was initiated by it. The dialectic does not regard nature as an accidental accumulation of objects, of isolated and independent phenomena, but as a unified, coherent whole. Furthermore, nature is perpetually in movement, in a state of unceasing renewal and development, in which there is always something being born and developing and something disintegrating and disappearing. (2) Evolution takes place in leaps, not gradually. (3) Contradictions must be made manifest. All phenomena contain in themselves contradictory elements. "Dialectic starts from the point of view that objects and natural phenomena imply internal contradictions, because they all have a positive and a negative side." These contradictory elements are in perpetual struggle: it is this struggle that is the "internal content of the process of development," according to Stalin. (4) The law of this development is economic. All other contradictions are rooted in the basic economic relationship. A given epoch is entirely determined by the relations of production existing among men. They are social relations; relations of collaboration or mutual aid, relations of domination or submission; and finally, transitory relations that characterize a period of passage from one system to another. "The history of the development of society is, above all, the history of the development of production, the history of the modes of production which succeed one another through the centuries."

From these principles may be drawn the following inferences, essential for penetrating the workings of Marxist-Leninist thought and its application. No natural phenomenon, no historical or social situation, no political fact, can be considered independently of the other facts or phenomena that surround it; it is set within a whole. Since movement is the essential fact, one must distinguish between what is beginning to decay and what is being born and developing. Since the process of development takes place by leaps, one passes suddenly from a succession of slow quantitative changes to a radical qualitative change. In the social or political realm, these sudden qualitative changes are revolutions, carried out by the oppressed classes. One must follow a frankly proletarian-class policy that exposes the contradictions of the capitalist system. A reformist policy makes no sense. Consequently (1) nothing can be judged from the point of view of "eternal justice" or any other preconceived notion and (2) no social system is immutable. To be effective, one must not base one's action on social strata that are no longer developing, even if they represent for the moment the dominant force, but on those that are developing.

Differences between Marx's ideas and Stalin's

Stalin's materialist and historical dialectic differs sharply from the perspective of Karl Marx. In *The Communist*

Manifesto Marx applied the materialist dialectic to the social and political life of his time. In the chapter entitled "Bourgeois and Proletarians," he studied the process of the growth of the revolutionary bourgeoisie within feudal society, then the genesis and the growth of the proletariat within capitalism, placing the emphasis on the struggle between antagonistic classes. To be sure, he connected social evolution with the development of the forces of production. What counted for him, however, was not only the struggle but also the birth of consciousness among the proletariat. "As to the final victory of the propositions put forth in the *Manifesto,* Marx expected it to come primarily from the intellectual development of the working class, necessarily the result of common action and discussion" (Engels, preface to the re-publication of *The Communist Manifesto,* May 1, 1890).

The result of Stalin's dialectic, however, was what he called revolution from above, a dictatorial policy to increase industrialization and collectivize agriculture based upon ruthless repression and a strong centralization of power. For Stalin what counted was the immediate goal, the practical result. The move was from a dialectic that emphasized both the objective and the subjective to one purely objective, or more exactly, objectivist. Human actions are to be judged not by taking account of the intentions of the actor and their place in a given historical web but only in terms of what they signify objectively at the end of the period considered.

Trotskyism. Alongside Marxism-Leninism as expounded in the Soviet Union, there arose another point of view expressed by Stalin's opponent Leon Trotsky and his followers. Trotsky played a leading role in both the Russian Revolution of 1905 and that of 1917. After Lenin's death he fell out with Stalin. Their conflict turned largely upon questions of policy, both domestic and foreign. In the realm of ideas, Trotsky held that a revolution in a backward, rural country could only be carried out by the proletariat. Once in power the proletariat must carry out agrarian reform and undertake the accelerated development of the economy. The revolution must be a socialist one, involving the abolition of the private ownership of the means of production, or else it will fail. But the revolution cannot be carried out in isolation, as Stalin maintained it could. The capitalist countries will try to destroy it; moreover, to succeed the revolution must be able to draw upon the industrial techniques of the developed countries. For these reasons the revolution must be worldwide and permanent, directed against the liberal and nationalist bourgeoisie of all countries and using local victories to advance the international struggle.

Trotsky's theory of permanent revolution

Tactically, Trotsky emphasized the necessity of finding or creating a revolutionary situation, of educating the working class in order to revolutionize it, of seeing that the party remained open to the various revolutionary tendencies and avoided becoming bureaucratized, and finally, when the time for insurrection comes, of organizing it according to a detailed plan.

VARIANTS OF SOVIET MARXISM

Maoism. When the Chinese Communists took power in 1948, they brought with them a new kind of Marxism that came to be called Maoism after their leader Mao Zedong. The thought of Mao must always be seen against the changing revolutionary reality of China from 1930 onward. His thought is complex, a Marxist type of analysis combined with the permanent fundamentals of Chinese thought and culture.

One of its central elements has to do with the nature and role of contradictions in socialist society. For Mao, every society, including socialist (communist) society, contains "two different types of contradictions": (1) antagonistic contradictions—contradictions between us (the people) and our enemies (the Chinese bourgeoisie faithful), between the imperialist camp and the socialist camp, and so forth—which are resolved by revolution, and (2) nonantagonistic contradictions—between the government and the people under a socialist regime, between two groups within the Communist Party, between one section of the people and another under a communist regime, and so forth—which are resolved by vigorous fraternal criticism and self-criticism.

The notion of contradiction is specific to Mao's thought in that it differs from the conceptions of Marx or Lenin. For Mao, in effect, contradictions are at the same time universal and particular. In their universality, one must seek and discover what constitutes their particularity: every contradiction displays a particular character, depending on the nature of things and phenomena. Contradictions have alternating aspects—sometimes strongly marked, sometimes blurred. Some of these aspects are primary, others secondary. It is important to define them well for if one fails to do so, the analysis of the social reality and the actions that follow from it will be mistaken. This is quite far from Stalinism and dogmatic Marxism-Leninism.

Another essential element of Mao's thought, which must be seen in the context of revolutionary China, is the notion of permanent revolution. It is an old idea advocated in different contexts by Marx, Lenin, and Trotsky but lacking, in Mao's formulation, the international dimension espoused by his predecessors. For Mao it follows from his ideas about the struggle of man against nature (held since 1938, at least); the campaigns for the rectification of thought (1942, 1951, 1952); and the necessity of struggling against bureaucracy, wastage, and corruption in a country of 600,000,000 to 700,000,000 inhabitants, where very old civilizations and cultures still permeate both the bourgeois classes and the peasantry, where bureaucracy is thoroughly entrenched, and where the previous society was extremely corrupt. It arises from Mao's conviction that the rhythm of the revolution must be accelerated. This conviction appeared in 1957 in his speeches and became manifest in 1958 in the "Great Leap Forward," followed in 1966 by the Cultural Revolution.

Mao's theory of permanent revolution

Mao's concept of permanent revolution rests upon the existence of nonantagonistic contradictions in the China of today and of tomorrow. Men must be mobilized into a permanent movement in order to carry forward the revolution and to prevent the ruling group from turning bourgeois (as he asserts it has in the Soviet Union). It is necessary to shape among the masses a new vision of the world by tearing them from their passivity and their century-old habits. This is the background of the Cultural Revolution that began in 1966, following upon previous campaigns but differing from them in its magnitude and, it would seem, in the mobilization of youth against the cadres of the party. In these campaigns Mao drew upon his past as a revolutionary Marxist peasant leader, from his life in the red military and peasant bases and among the Red Guards of Yen-an, seeking in his past experience ways to mobilize the whole Chinese population against the dangers—internal and external—that confronted it in the present.

The distinguishing characteristic of Maoism is that it represents a peasant type of Marxism, with a principally rural and military outlook. While basing himself on Marxism-Leninism, adapted to Chinese requirements, Mao is rooted in the peasant life from which he himself came, in the revolts against the warlords and the bureaucrats that have filled the history of China. By integrating this experience into a universal vision of history, Mao gives it a significance that overflows the provincial limits of China.

In his effort to remain close to the Chinese peasant masses, Mao draws upon an idea of nature and a symbolism found in popular Chinese Taoism, though transformed by his Marxism. It can be seen in his many poems, which are written in the classical Chinese style. This idea of nature is accompanied in his written political works by the other, Promethean idea of man struggling in a war against nature, a conception in his thought that goes back at least to 1938 and became more important after 1955 as the rhythm of the revolution accelerated.

Marxism in Yugoslavia and Cuba. Among the eastern European countries, Yugoslavia is the one whose policy is least dependent on Soviet Marxism. The Titoist variant of Marxism-Leninism developed after the Yugoslav Communists, led by Josip Broz Tito, fell out with the Soviet Union in June 1948. It is a form of Marxist socialism that in some respects is opposed to Stalinist Marxism-Lenin-

ism. Thus it emphasizes the struggle against bureaucracy and national chauvinism, along with the need for workers' control of the factories. In other respects it remains faithful to Stalinism in the unity of its leadership and in the monolithic nature of the League of Communists of Yugoslavia.

The Yugoslav party claims that it has started on the road to the withering away of the state. It does not operationally direct economic and social life, as do other communist parties that hold power. In foreign policy, Titoism refuses allegiance to either of the great power blocs; it favours self-determination and independence for the ex-colonial countries of the Third World.

The Marxism of Fidel Castro expresses itself as a rejection of injustice in any form—political, economic, or social. In this sense it is related to the liberal democracy and Pan-Americanism of Simón Bolívar in Latin America during the 19th century. In its liberalism, Castro's early socialism resembled the various French socialisms of the first half of the 19th century. Only gradually did Castroism come to identify itself with Marxism-Leninism, although from the very beginning of the Cuban revolution Castro revealed his attachment to certain of Marx's ideas. Castro's Marxism rejects some of the tenets and practices of official Marxism-Leninism: it is outspoken against dogmatism, bureaucracy, and sectarianism. In one sense, Castroism is a Marxist-Leninist "heresy." It exalts the ethos of guerrilla revolution over party politics. At the same time it aims to apply a purer Marxism to the conditions of Cuba: alleged American imperialism, a single-crop economy, a low initial level of political and economic development. One may call it an attempt to realize a synthesis of Marxist ideas and the ideas of Bolívar.

In the ideological and political conflicts that divide the communist world, Castroism takes a more or less unengaged position. Castro is above all a nationalist and only after that a Marxist, much as Tito was in Yugoslavia.

Marxism in the Third World. The development of Marxist variants in the Third World has been primarily influenced by the undeveloped industrial state and the former colonial status of the nations in question. In the traditional Marxist view the growth of capitalism is seen as a step necessary for the breakup of precapitalist peasant society and for the rise of the revolutionary proletariat class. Some theorists believe, however, that capitalism introduced by imperialist rather than indigenous powers sustains rather than destroys the feudal structure of peasant society and promotes underdevelopment because resources and surplus are usurped by the colonial powers. Furthermore, the revolutionary socialist movement becomes subordinate to that of national liberation, which violates Marx's theory of class struggle by uniting all indigenous classes in the common cause of anti-imperialism. For these reasons, many Third World countries have chosen to follow the Maoist model, with its emphasis on agrarian revolution against feudalism and imperialism, rather than the Soviet one. Another alternative, one specific to the Third World, also exists. This policy bypasses capitalism and depends upon the established strength of the communist bloc for support against imperialism. In effect, the socialist state is achieved through alliance with the Soviet Union rather than through class struggle and true revolution.

Problems raised by imperialism

Marxism in the West. There are two main forms of Marxism in the West: that of the traditional communist parties and the more diffuse "New Left" form, which has come to be known as "Western Marxism." In general, the success of western European communist parties has been hindered by their perceived allegiance to Soviet authority rather than their own countries; the secretive, bureaucratic form of organization they inherited from Lenin; the ease with which they become integrated into capitalist society; and their consequent fear of compromising their principles by sharing power with bourgeois parties. The Western parties basically adhered to the policies of Soviet Marxism until the 1970s, when they began to advocate Eurocommunism, a moderate version of communism that they felt would broaden their base of appeal beyond the working class and thus improve their chances for political success. As described by Enrico Berlinguer, Georges Marchais, and Santiago Carrillo, the leaders of the Italian, French, and Spanish communist parties, respectively, Eurocommunism favoured a peaceful, democratic approach to achieving socialism, encouraged making alliances with other political parties, guaranteed civil liberties, and renounced the central authority of the Soviet party. By the 1980s Eurocommunism had largely been abandoned as unsuccessful, and communist parties in advanced capitalist nations returned to orthodox Marxism-Leninism despite the concomitant problems.

Western Marxism, by contrast, can be seen as a repudiation of Marxism-Leninism, although, when it was first formulated in the 1920s, its proponents believed they were loyal to the dominant Soviet Communist Party. Prominent figures in the evolution of Western Marxism include the central Europeans György Lukács, Karl Korsch, and Lucien Goldmann; Antonio Gramsci of Italy; the German theorists who constituted the Frankfurt school, especially Max Horkheimer, Theodor Adorno, Herbert Marcuse, and Jürgen Habermas; and Henri Lefebvre, Jean-Paul Sartre, and Maurice Merleau-Ponty of France.

Western Marxism

Western Marxism has been shaped primarily by the failure of the socialist revolution in the Western world. Western Marxists were concerned less with the actual political or economic practice of Marxism than with its philosophical interpretation, especially in relation to cultural and historical studies. In order to explain the inarguable success of capitalist society, they felt they needed to explore and understand non-Marxist approaches and all aspects of bourgeois culture. Eventually, they came to believe that traditional Marxism was not relevant to the reality of modern Western society.

Marx had predicted that revolution would succeed in Europe first, but, in fact, the Third World has proved more responsive. Orthodox Marxism also championed the technological achievements associated with capitalism, viewing them as essential to the progress of socialism. Experience showed the Western Marxists, however, that technology did not necessarily produce the crises Marx described and did not lead inevitably to revolution. In particular they disagreed with the idea, originally emphasized by Engels, that Marxism is an integrated, scientific doctrine that can be applied universally to nature; they viewed it is a critique of human life, not an objective, general science. Disillusioned by the terrorism of the Stalin era and the bureaucracy of the Communist Party system, they advocated the idea of government by workers' councils, which they believed would eliminate professional politicians and would more truly represent the interests of the working class. Later, when the working class appeared to them to be too well integrated into the capitalist system, the Western Marxists supported more anarchistic tactics. In general, their views are more in accord with those found in Marx's early, humanist writings than with later, dogmatic interpretations.

Western Marxism has found support primarily among intellectuals rather than the working class, and orthodox Marxists have judged it impractical. Nevertheless, the Western Marxists' emphasis on Marx's social theory and their critical assessment of Marxist methodology and ideas have coloured the way even non-Marxists view the world.

(H.C./D.T.McL.)

MAJOR WORKS

Misère de la philosophie (1847; *The Poverty of Philosophy,* 1900); *Manifest der Kommunistischen Partei* (1848; *Manifesto of the Communists,* 1883); *Die Klassenkämpfe in Frankreich 1848 bis 1850* (1850; *The Class Struggles in France, 1848 to 1850,* 1924); *Der 18te Brumaire des Louis Napoleon* (1852; *The Eighteenth Brumaire of Louis Bonaparte,* 1898); *Zur Kritik der politischen Ökonomie* (1859; *A Contribution to the Critique of Political Economy,* 1904); and *Das Kapital* (vol. 1, 1867; vol. 2–3 published by Engels in 1885 and 1894; *Capital: A Critical Analysis of Capitalist Production,* vol. 1 trans. by Samuel Moore and Edward Aveling, 1886; vol. 2–3 trans. by Ernest Untermann, 1907 and 1909).

Recommended later translations of these works include *Manifesto of the Communist Party,* trans. by Samuel Moore (1888, reprinted 1952); *The Communist Manifesto of Karl Marx and Friedrich Engels,* with an introduction and explanatory notes

by D. Ryazanoff, trans. by Eden Paul and Cedar Paul (1930); and *The Eighteenth Brumaire of Louis Bonaparte,* trans. by Eden Paul and Cedar Paul (1926). Selections from Marx's writings are available in David McLellan (ed.), *Selected Writings* (1977); Robert C. Tucker (ed.), *The Marx-Engels Reader,* 2nd ed. (1978); and Jon Elster (ed.), *Karl Marx: A Reader* (1986).

A major English-language edition, *Karl Marx, Frederick Engels: Collected Works,* trans. by Richard Dixon *et al.* (1975–), is in progress. Planned to consist of 50 volumes and to include the correspondence, it is being prepared by an international editorial committee in which scholars from Great Britain, the Soviet Union, and the United States are taking part. Thirty-two volumes had been published by 1988.

BIBLIOGRAPHY

Marx. The most comprehensive biography of Marx is DAVID MCLELLAN, *Karl Marx: His Life and Thought* (1973, reissued 1987). The classic biography of Marx, somewhat too partisan in his favour, is FRANZ MEHRING, *Karl Marx: The Story of His Life* (1935, reissued 1981; originally published in German, 1918). Marx's personal life is discussed in SAUL K. PADOVER, *Karl Marx, an Intimate Biography* (1978). JERROLD SEIGEL, *Marx's Fate: The Shape of a Life* (1978), is a psychological biography. YVONNE KAPP, *Eleanor Marx,* 2 vol. (1972–76), contains informative material on Marx's family life. Two good short biographies are ISAIAH BERLIN, *Karl Marx: His Life and Environment,* 4th ed. (1978, reprinted with corrections 1982); and WERNER BLUMENBERG, *Portrait of Marx: An Illustrated Biography* (1972; originally published in German, 1962).

For introductory analysis and commentaries of Marx's works, see DAVID MCLELLAN, *The Thought of Karl Marx,* 2nd ed. (1980); BRUCE MAZLISH, *The Meaning of Karl Marx* (1984); W.A. SUCHTING, *Marx, an Introduction* (1983), and *Marx and Philosophy: Three Studies* (1986); and RICHARD SCHMITT, *Introduction to Marx and Engels: A Critical Reconstruction* (1987). TERRELL CARVER, *A Marx Dictionary* (1987), provides brief definitions of Marxian concepts without interpretative deviations from the original. For more detailed studies, see ROMAN ROSDOLSKY, *The Making of Marx's 'Capital'* (1977, reissued 1980; originally published in German, 1968); DEREK SAYER, *Marx's Method: Ideology, Science and Critique in Capital,* 2nd ed. (1983); ROBERT PAUL WOLFF, *Understanding Marx: A Reconstruction and Critique of Capital* (1984); HAL DRAPER, *Karl Marx's Theory of Revolution,* 3 vol. (1977–86); D. ROSS GANDY, *Marx and History: From Primitive Society to the Communist Future* (1979); MURRAY WOLFSON, *Marx: Economist, Philosopher, Jew: Steps in the Development of a Doctrine* (1982); DANIEL LITTLE, *The Scientific Marx* (1986); THOMAS SOWELL, *Marxism: Philosophy and Economics* (1985); and JOHN CUNNINGHAM WOOD (ed.), *Karl Marx's Economics: Critical Assessments,* 4 vol. (1987). A particularly acute summary of the difficulties in Marx's work is JON ELSTER, *Making Sense of Marx* (1985). Many monographs explore Marx's political and ideological development: CAROL C. GOULD, *Marx's Social Ontology: Individuality and Community in Marx's Theory of Social Reality* (1978, reprinted 1980), a study based on Marx's *Grundrisse;* RICHARD E. OLSEN, *Karl Marx* (1978); S.S. PRAWER, *Karl Marx and World Literature* (1976); PAUL THOMAS, *Karl Marx and the Anarchists* (1980); and ALLEN W. WOOD, *Karl Marx* (1981).

Marxism. Good introductions to the study of Marxism include LESZEK KOŁAKOWSKI, *Main Currents of Marxism: Its Rise, Growth, and Dissolution,* 3 vol. (1978, reprinted 1981; originally published in Polish, 1976–78); GEORGE LICHTHEIM, *Marxism: An Historical and Critical Study,* 2nd ed. (1964, reprinted 1982); and DAVID MCLELLAN, *Marxism After Marx* (1979, reissued 1981), which contains an extensive bibliography. Some important analyses are assembled in DAVID MCLELLAN (ed.), *Marxism: Essential Writings* (1988). Studies of Marxism as a sociological doctrine may be found in KARL KORSCH, *Karl Marx* ((1938, reissued 1963); HENRI LEFEBVRE, *The Sociology of Marx* (1968, reprinted 1982; originally published in French, 1966); and SIDNEY HOOK, *Towards the Understanding of Karl Marx* (1933). Developments in Marxism as a political theory are discussed in ALFRED SCHMIDT, *History and Structure: An Essay on Hegelian-Marxist and Structuralist Theories of History* (1981; originally published in German, 1971); DAVID RUBINSTEIN, *Marx and Wittgenstein: Social Praxis and Social Explanation* (1981); TOM ROCKMORE, *Fichte, Marx, and the German Philosophical Tradition* (1980); S.H. RIGBY, *Marxism and History: A Critical Introduction* (1987); and PAUL PHILLIPS, *Marx and Engels on Law and Laws* (1980). Specialized studies include STANLEY MOORE, *Marx on the Choice Between Socialism and Communism* (1980); JOSÉ PORFIRIO MIRANDA, *Marx Against the Marxists: The Christian Humanism of Karl Marx* (1980; originally published in Spanish, 1978); RALPH MILIBAND, *Marxism and Politics* (1977), including a discussion of the applicability of Marxism to contemporary politics in the Third World and communist countries; ROBERT L. HEILBRONER, *Marxism, For and Against* (1980); ANTHONY GIDDENS, *A Contemporary Critique of Historical Materialism,* 2 vol. (1981–85), an alternative, based on anthropological research, to the Marxist idea that all history has been the history of class struggle; MAURICE GODELIER, *Perspectives in Marxist Anthropology* (1977; originally published in French, 1973), presenting the contrasting view that classical Marxism may provide a methodology for analysis of empirical data in history and anthropology; and IAN CUMMINS, *Marx, Engels, and National Movements* (1980).

A. JAMES GREGOR, *A Survey of Marxism: Problems in Philosophy and the Theory of History* (1965), emphasizes philosophical problems in lieu of political or economic ones. The outstanding work on Marxist ethics is EUGENE KAMENKA, *The Ethical Foundations of Marxism,* 2nd ed. (1972). See also HUGO MEYNELL, *Freud, Marx, and Morals* (1981); and GARY NELSON and LAWRENCE GROSSBERG (eds.), *Marxism and Interpretation of Culture* (1988).

DAVID HOROWITZ (ed.), *Marx and Modern Economics* (1968), is an excellent collection of essays by leading economic theorists. Other treatments of Marxist economics worth consulting are PAUL M. SWEEZY, *The Theory of Capitalist Development: Principles of Marxian Political Economy* (1942, reissued 1970); and JOHN STRACHEY, *The Nature of Capitalist Crisis* (1935). The place of Marxist thought in the intellectual history of the 20th century is assessed in JACK LINDSAY, *The Crisis in Marxism* (1981); ANTHONY BREWER, *Marxist Theories of Imperialism: A Critical Survey* (1980); PERRY ANDERSON, *Considerations on Western Marxism* (1976); and WALTER L. ADAMSON, *Marx and the Disillusionment of Marxism* (1985).

An account of the historical development of Marxism can be found in HENRI CHAMBRE, *From Karl Marx to Mao Tse-Tung: A Systematic Survey of Marxism-Leninism* (1963; originally published in French, 1959). GEORGE D.H. COLE, *A History of Socialist Thought,* 5 vol. in 7 (1953–65), presents a detailed study of the Marxist movement rather than the ideas; see especially vol. 2, *Socialist Thought: Marxism and Anarchism, 1850–1890.* TOM BOTTOMORE (ed.), *Interpretations of Marx* (1988), is an authoritative collection of essays.

The development and influence of Russian, Soviet, and eastern European Marxist theories is covered in a number of works by both Marxist and non-Marxist authors: HERBERT MARCUSE, *Soviet Marxism: A Critical Analysis* (1958, reprinted 1985); BERTRAM D. WOLFE, *Revolution and Reality: Essays on the Origin and Fate of the Soviet System* (1981); BARUCH KNEI-PAZ, *The Social and Political Thought of Leon Trotsky* (1978); UMBERTO MELOTTI, *Marx and the Third World* (1977, reprinted 1982; originally published in Italian, 1971); ADAM WESTOBY, *Communism Since World War II* (1981); and ERNEST MANDEL, *Revolutionary Marxism Today* (1979). Specialized studies include DONALD C. HODGES, *The Bureaucratization of Socialism* (1981); ROBERT J. ALEXANDER, *The Right Opposition: The Lovestoneites and the International Communist Opposition of the 1930's* (1981); ESTHER KINGSTON-MANN, *Lenin and the Problem of Marxist Peasant Revolution* (1983); BOGDAN SZAJKOWSKI (ed.), *Marxist Governments: A World Survey,* 3 vol. (1981); V. KUBÁLKOVÁ and A.A. CRUICKSHANK, *Marxism-Leninism and Theory of International Relations* (1980); HORACE B. DAVIS, *Toward a Marxist Theory of Nationalism* (1978); ISAAC DEUTSCHER, *Marxism in Our Time* (1971); JOHN P. BURKE, LAWRENCE CROCKER, and LYMAN H. LEGTERS (eds.), *Marxism and the Good Society* (1981), on Russia, Yugoslavia, and China; GERSON S. SHER (ed. and trans.), *Marxist Humanism and Praxis* (1978), a collection of 11 essays by Yugoslav Marxists of the dissident Praxis group; JOHN G. GURLEY, *Challengers to Capitalism: Marx, Lenin, Stalin, and Mao,* 3rd ed. (1988); and NICHOLAS ABERCROMBIE, STEPHEN HILL, and BRYAN S. TURNER, *The Dominant Ideology Thesis* (1980), a critique of current Marxist thought. Two important critical studies are DAVID LANE, *The Socialist Industrial State: Towards a Political Sociology of State Socialism* (1976); and DONALD WILHELM, *Creative Alternatives to Communism: Guidelines for Tomorrow's World* (1977, reprinted 1981).

Of special reference interest are JOHN LACHS, *Marxist Philosophy: A Bibliographical Guide* (1967); HARRY G. SHAFFER, *Periodicals on the Socialist Countries and on Marxism: A New Annotated Index of English-Language Publications* (1977); J. WILCZYNSKI, *An Encyclopedic Dictionary of Marxism, Socialism and Communism* (1981); and ROBERT A. GORMAN (ed.), *Biographical Dictionary of Marxism* (1986), and *Biographical Dictionary of Neo-Marxism* (1985), a compendium providing information on practitioners of Marxism in more than 50 countries.

(D.T.McL.)

Masks

Simply defined, a mask is a form of disguise. It is an object that is frequently worn over or in front of the face to hide the identity of a person and by its own features to establish another being. This essential characteristic of hiding and revealing personalities or moods is common to all masks. As cultural objects they have been used throughout the world in all periods since the Stone Age and have been as varied in appearance as in their use and symbolism.

This article deals with the general characteristics, functions, and forms of masks. It is divided into the following sections:

General characteristics 586
The making of masks 586
The wearing of masks 587
The role of the spectator 587
Meaning and aesthetic response 587
Preservation and collecting 587
The functions and forms of masks 587
Social and religious uses 587
Funerary and commemorative uses 589
Therapeutic uses 591
Festive uses 591
Theatrical uses 592

GENERAL CHARACTERISTICS

Masks have been designed in innumerable varieties, from the simplest of crude "false faces" held by a handle to complete head coverings with ingenious movable parts and hidden faces. Mask makers have shown great resourcefulness in selecting and combining available materials. Among the substances utilized are woods, metals, shells, fibres, ivory, clay, horn, stone, feathers, leather, furs, paper, cloth, and corn husks. Surface treatments have ranged from rugged simplicity to intricate carving and from polished woods and mosaics to gaudy adornments.

Carl Frank



Mask worn with costume: *makishi* dancer, a masked ancestral spirit who assists at initiation rites of the tribes of the northwestern region of Zambia.

Masks generally are worn with a costume, often so complete that it entirely covers the body of the wearer. Fundamentally the costume completes the new identity represented by the mask, and usually tradition prescribes its appearance and construction to the same extent as the mask itself. Costumes, like the masks, are made of a great variety of materials, all of which have a symbolic connection with the mask's total imagery. Ideally the costume should be seen with the mask while the wearer is in action.

The morphological elements of the mask are with few exceptions derived from natural forms. Masks with human features are classified as anthropomorphic and those with animal characteristics as theriomorphic. In some instances, the mask form is a replication of natural features or closely follows the lineaments of reality, and in other instances it is an abstraction. Masks usually represent supernatural beings, ancestors, and fanciful or imagined figures and can also be portraits. The localization of a particular spirit in a specific mask must be considered a highly significant reason for its existence. The change in identity of the wearer for that of the mask is vital, for if the spirit represented does not reside in the image of the mask, the ritual petitions, supplications, and offerings made to it would be ineffectual and meaningless. The mask, therefore, most often functions as a means of contact with various spirit powers, thereby protecting against the unknown forces of the universe by prevailing upon their potential beneficence in all matters relative to life.

Anthropomorphic and theriomorphic masks

The making of masks. With few exceptions, masks have been made by professionals who were either expert in this particular craft or were noted sculptors or artisans. In societies in which masks of supernatural beings have played a significant ceremonial role, it is presumed that the spirit power of the created image usually is strongly felt by the artist. A primary belief involved in both the conception and the rendering of these objects was that spirit power dwelled in all organic and inorganic matter, and therefore the mask will contain the spirit power of whatever material was used to make it. This power is considered a volatile, active force that is surrounded by various taboos and restrictions for the protection of those handling it. Certain prescribed rituals frequently have to be followed in the process of the mask's creation. A spirit power is also often believed to inhabit the artist's tools so that even these have to be handled in a prescribed manner. As the form of the mask develops it is usually believed to acquire power increasingly in its own right, and again various procedures are prescribed to protect the craftsman and to ensure the potency of the object. If all the conventions have been adhered to, the completed mask, when worn or displayed, is regarded as an object suffused with great supernatural or spirit power. In some cultures it is believed that because of the close association between the mask maker and the spirit of the mask, the artist absorbs some of its magic power. A few West African tribal groups in Mali believe, in fact, that the creators of masks are even potentially capable of using the object's supernatural powers to cause harm to others.

The mask maker

Aesthetically, the mask maker has usually been restricted in the forms he can use since masks generally have a traditional imagery with formal conventions. If they are not followed, the artist can bring upon himself the severe censure of his social group and the displeasure or even wrath of the spirit power inherent in the mask. This requirement for accuracy, however, does not restrict artistic expressiveness. The mask maker can and does give his own creative interpretation to the traditionally prescribed general forms, attributes, and devices. The artist, in fact, is usually sought out as a maker of masks because of his

known ability to give a vitally expressive or an aesthetically pleasing presentation of the required image.

The mask wearer

The wearing of masks. The wearer is also considered to be in direct association with the spirit force of the mask and is consequently exposed to like personal danger of being affected by it. For his protection, the wearer, like the mask maker, is required to follow certain sanctioned procedures in his use of the mask. In some respects he plays the role of an actor in cooperation or collaboration with the mask. Without his performing dance and posturing routines, which are often accompanied with certain sounds of music, the mask would remain a representation without a full life-force. The real drama and power of its form is the important contribution of the wearer. When he is attired in the mask, there is a loss of his previous identity and the assuming of a new one. Upon donning the mask, the wearer sometimes undergoes a psychic change and as in a trance assumes the spirit character depicted by the mask. Usually, however, the wearer skillfully becomes a "partner" of the character he is impersonating, giving to the mask not only an important spark of vitality by the light flashing from his own eyes but also bringing it alive by his movements and poses. But it would seem that the wearer often becomes psychologically completely attached to the character he is helping to create. He loses his own identity and becomes like an automaton, without his own will, which has become subservient to that of the personage of the mask. It appears, however, that at all times there remains some important, even if *sub rosa,* association between the mask and its wearer.

Supernatural and spirit powers

The role of the spectator. It is as consecrated objects imbued with supernatural power that masks are viewed by the spectators or participants at ceremonials where their presence is required. Whatever their specific identity may be, the masks usually refer back to early times, when their initial appearance occurred. This basic aspect of the mask is understood at least in essence by everyone. A paramount role of the mask is to give a sense of continuity between the present and the beginnings of time, a sense that is of vital importance for the integration of a culture with no written history. Psychologically the spectators become associated with the past through the spirit power of the mask, and this often leads the participants to a state of complete absorption or near-frenzy. This is not, however, a consistent reaction to masked ceremonials. That depends on the character whose presence the mask represents. In some cases, the spirit or supernatural being depicted is viewed with rejoicing and almost a familiarity, which leads to gaiety that has a cathartic aspect. Even so, the mask has a spirit content that is respected and revered, even if it is not showing a being with malignant potential. All of these forms have spirit and magical qualities and are thus esteemed as agents for the accomplishing of suprahuman acts.

Some masks, however, do represent malignant, evil, or potentially harmful spirits. These are often used to keep a required balance of power or a traditional social and political relationship of inherited positions within a culture. The characters depicted are also prescribed by tradition and enact roles to achieve the desired ends. The drama involving these masks is often associated with secret societies, especially in Africa, where the greatest range of mask forms and functions can be observed. These forms are often used in very restricted performances, where only select persons can view them. This is also true in other areas where masks are used, such as in Oceania, the Americas, and even in some of the folk mask rites still performed in Europe.

Meaning and aesthetic response. On the basis of present knowledge, it would appear that there is not or has not been any set response or reaction by any one of the three groups involved with the mask: the artist, the wearer, the spectator. There is, however, a reaction of a very particular kind common to every culture, a response such as awe, delight and pleasure, fear and even terror: these are as traditionally determined as the forms and costumes of the masks themselves. This is a learned and inherent pattern of conduct for each culture. Masks, therefore, that have a closely comparable appearance in several unrelated groups in quite different parts of the world often have totally dissimilar meanings and functions. It is thus practically impossible to determine either the meaning or use of a mask by its appearance alone. For example, some masks in Africa, as well as in Oceania and East Asia, have such a grotesque or frightening appearance as to lead one to suspect that they represent evil spirits with an intent to terrorize the spectators; actually they may have the opposite character and function. The significance of masks can be determined only by reference to accounts or personal observations of the masks in the setting of their own culture.

Art objects

The aesthetic effects of masks, on the other hand, since they derive from the forms and their disposition within the design, can readily be evaluated as art objects. But this evaluation is based on elements very different from those appraised within the mask's own culture. This is partly because the total artistic qualities of a mask derive both from its exterior forms and from its meaning and function within its cultural context. There exist, however, in all cultures criteria for determining the quality of objects as art. These criteria differ from one culture to another, and they may be known only from investigations carried out within the varying cultures.

Preservation and collecting. The preservation or disposal of masks is often decreed by tradition. Many masks and often their form and function are passed down through clans, families, special societies, or from individual to individual. They are usually spiritually reactivated or aesthetically restored by repainting and redecorating, without destroying the basic form and symbolism. In many instances, however, the mask is used only for one ceremony or occasion and then is discarded or destroyed, sometimes by burning.

The collecting of masks has largely been of recent origin. Not until the late 19th and early 20th century were they seriously appreciated as art objects or studied as cultural artifacts. Most masks have been obtained through archaeological excavations or in field expeditions, that is, in their place of origin.

THE FUNCTIONS AND FORMS OF MASKS

Masks are as extraordinarily varied in appearance as they are in function or fundamental meaning. Many masks are primarily associated with ceremonies that have religious and social significance or are concerned with funerary customs, fertility rites, or curing sickness. Other masks are used on festive occasions or to portray characters in a dramatic performance and in re-enactments of mythological events. Masks are also used for warfare and as protective devices in certain sports, as well as frequently being employed as architectural ornament.

Social and religious uses. Masks representing potentially harmful spirits were often used to keep a required balance of power or a traditional relationship of inherited positions within a culture. The forms of these masks invariably were prescribed by tradition, as were their uses. This type of mask was often associated with secret societies, especially in Africa, where the greatest range of types and functions can be found. They were also widely used among Oceanic peoples of the South Pacific and the American Indians and are even used in some of the folk rites still performed in Europe.

Admonitory and disciplinary masks

Masks have served an important role as a means of discipline and have been used to admonish women, children, and criminals. Common in China, Africa, Oceania, and North America, admonitory masks usually completely cover the features of the wearer. It is believed among some of the African Negro tribes that the first mask was an admonitory one. A child, repeatedly told not to, persisted in following its mother to fetch water. To frighten and discipline the child, the mother painted a hideous face on the bottom of her water gourd. Others say the mask was invented by a secret African society to escape recognition while punishing marauders. In New Britain, members of a secret terroristic society called the Dukduk appear in monstrous five-foot masks to police, to judge, and to execute offenders. Aggressive supernatural spirits of an almost demonic nature are represented by these masks, which are



Admonitory and ancestor masks.
(Left) Female *tumbuan* mask of the Dukduk secret society, New Britain. The head is of sacking with tuft of feathers and skirt of leaves. Height 1.5 m. (Right) Ancestor mask from the middle Sepik River area of New Guinea, used in clan initiation rites. Upper portion is basketwork with faces modelled of red earth mixed with coconut oil, trimmed with shells, feathers, and a skirt of bark fibres. Height 1.79 m.
By courtesy of the Museum fur Volkerkunde, Basel, Switzerland

constructed from a variety of materials, usually including tapa, or bark cloth, and the pith of certain reeds. These materials are painted in brilliant colours, with brick red and acid green predominating.

In many cultures throughout the world, a judge wears a mask to protect him from future recriminations. In this instance, the mask represents a traditionally sanctioned spirit from the past who assumes responsibility for the decision levied on the culprit.

Ancestor masks

Rituals, often nocturnal, by members of secret societies wearing ancestor masks are reminders of the ancient sanction of their conduct. In many cultures, these masked ceremonials are intended to prevent miscreant acts and to maintain the circumscribed activities of the tribe. Along the Guinea coast of West Africa, for instance, many highly realistic masks represent ancestors who enjoyed specific cultural roles; the masks symbolize sanction and control when donned by the wearer. Among some of the Dan and Ngere tribes of Liberia and Ivory Coast, ancestor masks with generic features act as intermediaries for the transmission of petitions or offerings of respect to the gods. These traditional ancestral emissaries exert by their spirit power a social control for the community.

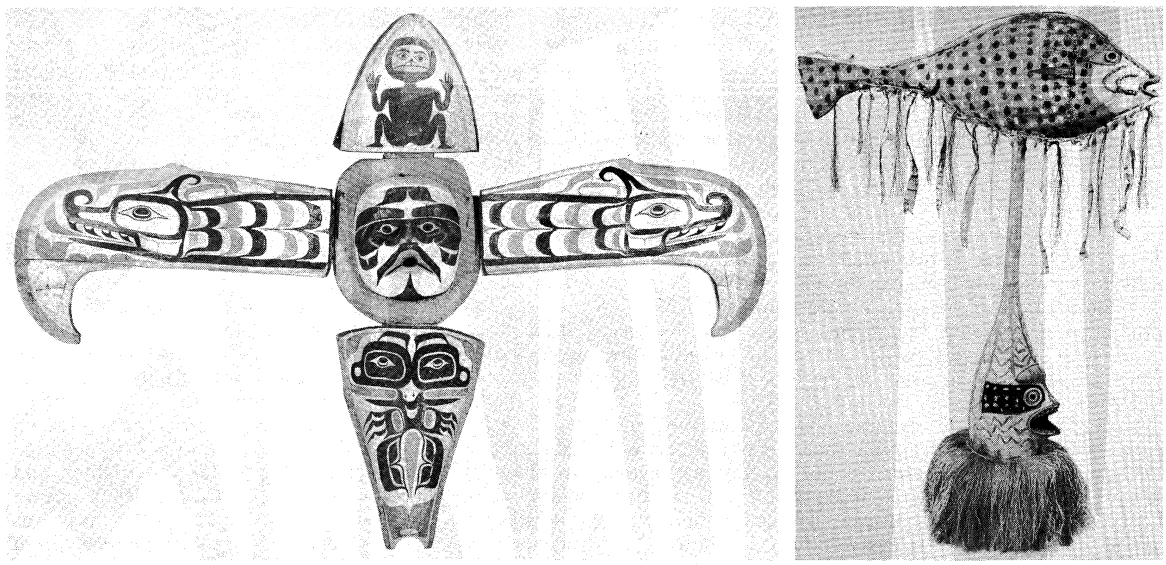
Particularly among Oceanic peoples, American Indians, and Negro tribes of Africa, certain times of the year are set aside to honour spirits or ancestors. Among nonliterate peoples who cannot record their own histories, masked rituals act as an important link between past and present, giving a sense of historic continuity that strengthens their social bond. On these occasions, masks usually recognizable as dead chieftains, relatives, friends, or even foes are worn or exhibited. Gifts are made to the spirits incarnated in the masks, while in other instances dancers wearing stylized mourning masks perform the prescribed ceremony.

In western Melanesia, the ancestral ceremonial mask occurs in a great variety of forms and materials. The Sepik River area in north central New Guinea is the source of an extremely rich array of these mask forms mostly carved in wood, ranging from small faces to large fantastic forms with a variety of appendages affixed to the wood, including shell, fibre, animal skins, seed, flowers, and feathers. These masks are highly polychromed with earth colours of red and yellow, lime white, and charcoal black. They often represent supernatural spirits as well as ancestors and therefore have both a religious and a social significance.

Initiation masks

Members of secret societies usually conduct the rituals of initiation, when a young man is instructed in his future role as an adult and is acquainted with the rules controlling the social stability of the tribe. Totem and spiritualistic masks are donned by the elders at these ceremonies. Sometimes the masks used are reserved only for initiations. Among the most impressive of the initiation masks are the exquisitely carved human faces of west coast African Negro tribes. In western and central Zaire, in Africa, large, colourful helmetlike masks are used as a masquerading device when the youth emerges from the initiation area and is introduced to the villagers as an adult of the tribe. After a lengthy ordeal of teaching and initiation rites, for instance, a youth of the Pende tribe appears in a distinctive colourful mask indicative of his new role as an adult. The mask is later cast aside and replaced by a

By courtesy of (left) the Brooklyn Museum, New York, (right) the Pitt Rivers Museum, Oxford, England



Totem masks.
(Left) Thunderbird mask of the Kwakiutl Indians of the Northwest Coast of North America. The bird's beak and wings open to reveal a human face (mask shown open). Painted wood. Width (open) 1.82 m. (Right) Fish mask from the Orokolo Bay area of New Guinea. Painted bark cloth over rattan frame with fringe of dried grass. Height 1.63 m.

small ivory duplicate, worn as a charm against misfortune and as a symbol of his manhood.

Totem masks

Believing everything in nature to possess a spirit, man found authority for himself and his family by identifying with a specific nonhuman spirit. He adopted an object of nature; then he mythologically traced his ancestry back to the chosen object; he preempted the animal as the emblem of himself and his clan. This is the practice of totem, which consolidates family pride and distinguishes social lines. Masks are made to house the totem spirit. The totem ancestor is believed actually to materialize in its mask; thus masks are of the utmost importance in securing protection and bringing comfort to the totem clan.

The Papuans of New Guinea build mammoth masks called *hevehe,* attaining 20 feet in height. They are constructed of a palm wood armature covered in bark cloth; geometric designs are stitched on with painted cane strips. These fantastic man–animal masks are given a frightening aspect. When they emerge from the men's secret clubhouse, they serve to protect the members of the clan. The so-called "totem" pole of the Alaskan and British Columbian Indian fulfills the same function. The African totem mask is often carved from ebony or other hard woods, designed with graceful lines and showing a highly polished surface. Animal masks, their features elongated and beautifully formalized, are common in western Africa. Dried grass, woven palm fibres, coconuts, and shells, as well as wood are employed in the masks of New Guinea, New Ireland, and New Caledonia. Represented are fanciful birds, fishes, and animals with distorted or exaggerated features.

The high priest and medicine man, or the shaman, frequently had his own very powerful totem, in whose mask he could exorcise evil spirits, punish enemies, locate game or fish, predict the weather, and, most importantly, cure disease.

The Northwest Coast Indians of North America in particular devised mechanical masks with movable parts to reveal a second face—generally a human image. Believing that the human spirit could take animal form and vice versa, the makers of these masks fused man and bird or man and animal into one mask. Some of these articulating masks acted out entire legends as their parts moved.

Funerary and commemorative uses. In cultures in which burial customs are important, anthropomorphic masks have often been used in ceremonies associated with the dead and departing spirits. Funerary masks were frequently used to cover the face of the deceased. Generally their purpose was to represent the features of the deceased, both to honour them and to establish a relationship through the mask with the spirit world. Sometimes they were used to force the spirit of the newly dead to depart for the spirit world. Masks were also made to protect the deceased by frightening away malevolent spirits.

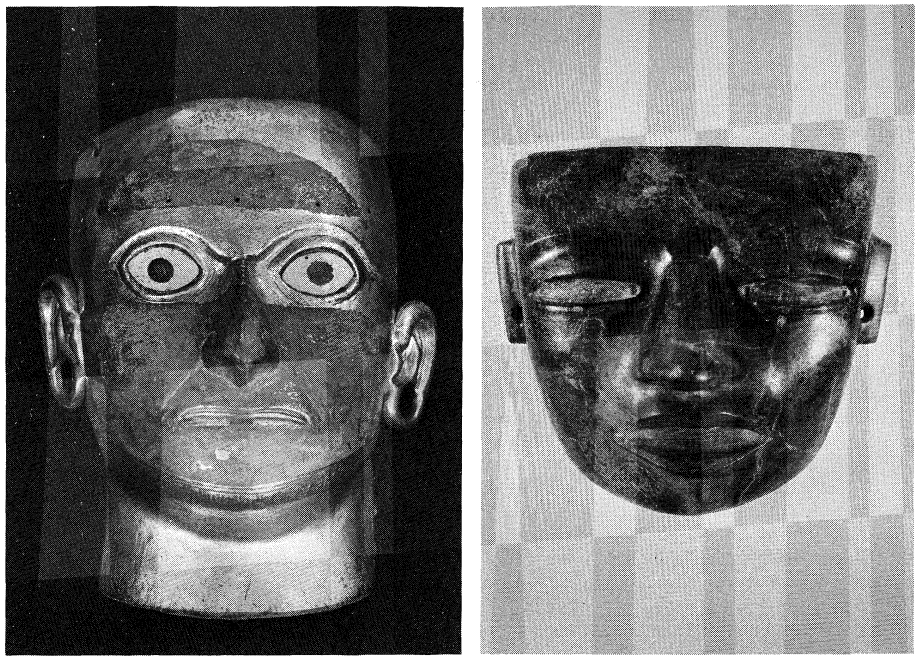
Burial masks

From the Middle Kingdom (*c.* 2040–1786 BC) to the 1st century AD, the ancient Egyptians placed stylized masks with generalized features on the faces of their dead. The funerary mask served to guide the spirit of the deceased back to its final resting place in the body. They were commonly made of cloth covered with stucco or plaster, which was then painted. For more important personages, silver and gold were used. Among the most splendid examples of the burial portrait mask is the one created *c.* 1350 BC for the pharaoh Tutankhamen. In Mycenaean tombs of

By courtesy of (top left) Museo Egizio, Turin, Italy, (top right) the National Archaeological Museum of Athens, TAP Service, (bottom left) the trustees of the British Museum; photographs, (top left) Foto Rosso, (bottom right) Nicholas Servian/Woodmansterne Limited



Funerary masks.
(Top left) Egyptian burial mask of the Ptolemaic period. Painted stucco over cloth. (Top right) Gold funeral mask placed over the face of an unknown Mycenaean ruler. Greece, 13th century BC. (Bottom left) Aztec skull mask from Mexico inlaid with turquoise and lignite with pyrite in the eye sockets. (Bottom right) Effigy mask of Edmund Sheffield, duke of Buckingham, 1735. Painted death mask with human hair, dressed in his peer's robes.



(Left) Metallic death mask with movable ears from the Moon Pyramid in the Moche Valley, Peru, 3rd–8th century AD. (Right) Ancient Mexican mask made of porphyry found north of Texcoco, Mexico, Teotihuacán civilization, 3rd–4th century AD.
By courtesy of (left) the Linden-Museum, Stuttgart, West Germany, (right) the American Museum of Natural History, New York City

c. 1400 BC, beaten gold portrait masks were found. Gold masks also were placed on the faces of the dead kings of Cambodia and Siam.

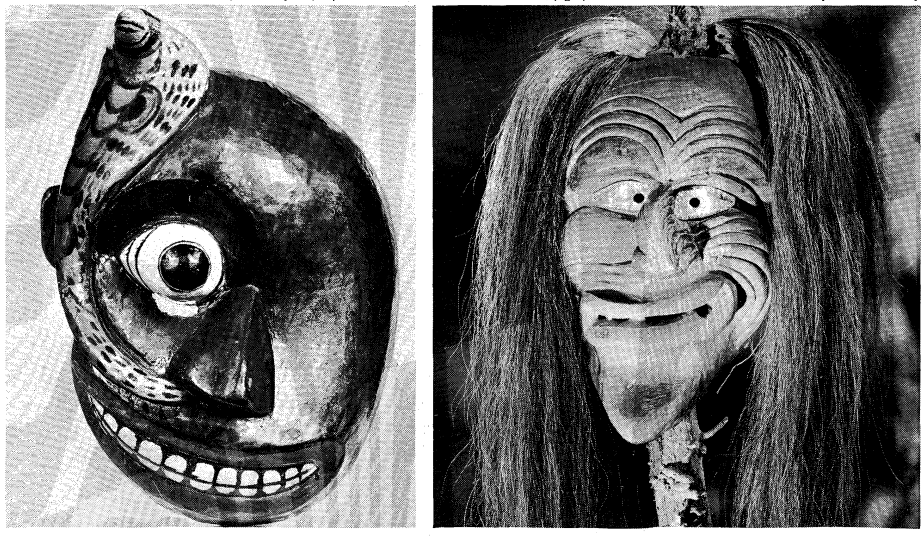
The mummies of Inca royalty wore golden masks. The mummies of lesser personages often had masks that were made of wood or clay. Some of these ancient Andean masks had movable parts, such as the metallic death mask with movable ears that was found in the Moon Pyramid at Moche, Peru. The ancient Mexicans made burial masks that seem to be generic representations rather than portraits of individuals.

In ancient Roman burials, a mask resembling the deceased was often placed over his face or was worn by an actor hired to accompany the funerary cortege to the burial site. In patrician families these masks or *imagines* were sometimes preserved as ancestor portraits and were displayed on ceremonial occasions. Such masks were usually modelled over the features of the dead and cast in wax. This technique was revived in the making of effigy masks for the royalty and nobility of Europe from the late Middle Ages through the 18th century. Painted and with human hair, these masks were attached to a dummy dressed in state regalia and were used for display, processionals, or commemorative ceremonials. From the 17th century to the 20th, death masks of famous persons became widespread among European peoples. With wax or liquid plaster of paris a negative cast of the human face could be produced that in turn acted as a mold for the positive image, frequently cast in bronze. In the 19th century, life masks made in the same manner became popular. Another type of life mask had been produced in the Fayyum region of Egypt during the 1st and 2nd centuries AD. These were realistic portraits painted in encaustic on wood during a person's lifetime; when the person died, they were attached directly to the facial area on the mummy shroud.

Death and life masks

The skull mask is another form usually associated with funerary rites. The skull masks of the Aztecs, like their wooden masks, were inlaid with mosaics of turquoise and lignite, and the eye sockets were filled with pyrites.

By courtesy of (left) the trustees of the British Museum, (right) the American Museum of Natural History, New York City



Therapeutic masks.
(Left) Disease devil mask *(rakasa)* from Sri Lanka worn to cure patients of deafness. (Right) False Face Society mask of the Iroquois Indians of North America.

Holes were customarily drilled in the back so the mask might be hung or possibly worn. In Melanesia, the skull of the deceased is often modelled over with clay, or resin and wax, and then elaborately painted with designs that had been used ceremonially by the deceased during his own lifetime.

Therapeutic uses. Masks have played an important part in magico-religious rites to prevent and to cure disease. In some cultures, the masked members of secret societies could drive disease demons from entire villages and tribes. Among the best known of these groups was the False Face Society of the North American Iroquois Indians. These professional healers performed violent pantomimes to exorcise the dreaded *Gahadogoka gogosa* (demons who plagued the Iroquois). They wore grimacing, twisted masks, often with long wigs of horsehair. Metallic inserts often were used around the eyes to catch the light of the campfire and the moon, emphasizing the grotesqueness of the mask.

Masks for protection from disease include the measle masks worn by Chinese children and the cholera masks worn during epidemics by the Chinese and Burmese. The disease mask is most developed among the Sinhalese in Sri Lanka (Ceylon), where 19 distinct *rākasa,* or disease devil masks, have been devised. These masks are of ferocious aspect, fanged, and with startling eyes. Gaudily coloured and sometimes having articulating jaws, they present a dragon-like appearance.

Uses in warfare and sport

Masks have long been used in military connections. A war mask will have a malevolent expression or hideously fantastic features to instill fear in the enemy. The ancient Greeks and Romans used battle shields with grotesque masks or attached terrifying masks to their armour, as did the Chinese warrior. Grimacing *menpo,* or mask helmets, were used by Japanese samurai.

Many sports require the use of masks. Some of these are merely functional, protective devices such as the masks worn by fencers, baseball catchers, or even skiers. To protect their faces in sports events and tournaments of arms, horsemen of the Roman army attached highly decorative and symbolic masks to their helmets.

Perhaps the earliest use of masks was in connection with hunting. Disguise masks were seemingly used in the early Stone Age in stalking prey and later to house the slain animal's spirit in the hope of placating it. The traditional animal masks worn by the Altaic and Tungusic shamans in Siberia are strictly close to such prehistoric examples as the image of the so-called sorcerer in the Cave of Les Trois Frères in Ariège, France.

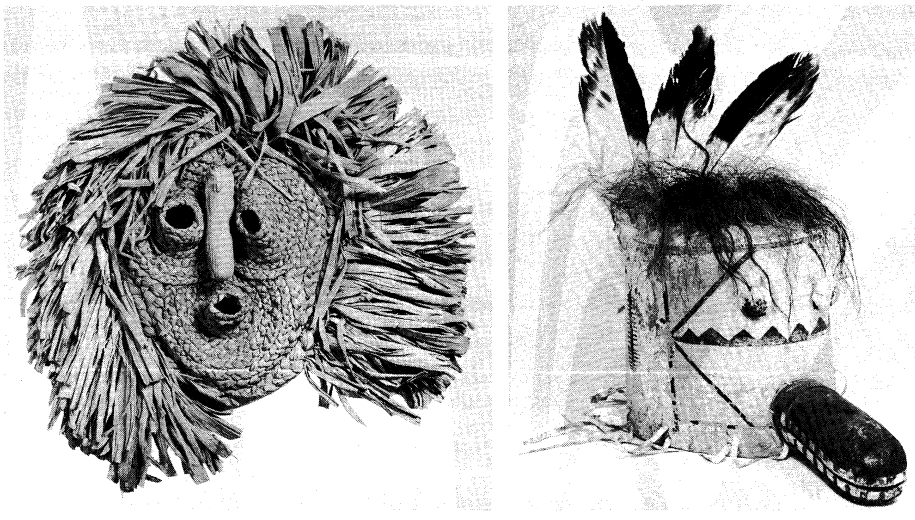
Uses in fertility ceremonies

Since agricultural societies first appeared in prehistory, the mask has been widely used for fertility rituals. The Iroquois, for instance, used corn husk masks at harvest rituals to give thanks for and to achieve future abundance of crops. Perhaps the most renowned of the masked fertility rites held by American Indians are those still performed by the Hopi and Zuni Indians of the Southwest U.S. Together with masked dancers representing clouds, rain spirits, stars, Earth Mother, sky god, and others, the shaman takes part in elaborate ceremonies designed to assure crop fertility. Spirits called *kachinas*, who first brought rain to the Pueblo tribes, are said to have left their masks behind when sent to dwell in the bottom of a desert lake. Their return to help bring the rain is incarnated by the masked dancer. Cylindrical masks, covering the entire head and resting on the shoulders, are of a primal type. They are made of leather and humanized by the addition of hair and a variety of adjuncts. Eyes are represented by incisions or by buckskin balls filled with deer hair and affixed to the mask. The nose is often of rolled buckskin or corncob. Frequently the mask has a projecting wooden cylinder for a bill or a gourd stem cut with teeth for a snout. Horns are attached to some masks. Many colours are used in their painting; plumes and beads are attached, and the sex of the mask is distinguished by its shape: round head indicates male and square indicates female. In the western Sudan area of Africa many tribes have masked fertility ceremonials. The *segoni-kun* masks that are fashioned by the Bambara tribes in Mali are aesthetically among the most interesting. Antelopes, characterized by their elegant simplicity, are carved in wood and affixed to woven fibre caps that are hung with raffia and cover the wearer. The antelope is believed to have introduced agriculture, and so when crops are sown, members of Tjiwara society cavort in the fields in pairs to symbolize fertility and abundance.

Festive uses. Masks for festive occasions are still commonly used in the 20th century. Ludicrous, grotesque, or superficially horrible, festival masks are usually conducive to good-natured license, release from inhibitions, and ribaldry. These include the Halloween, Mardi Gras, or "masked ball" variety. The disguise is assumed to create a momentary, amusing character, often resulting in humorous confusions, or to achieve anonymity for the prankster or ribald reveller.

Throughout contemporary Europe and Latin America, masks are associated with folk festivals, especially those generated by seasonal changes or marking the beginning and end of the year. Among the most famous of the folk masks are the masks worn to symbolize the driving away of winter in parts of Austria and Switzerland. In Mexico and Guatemala, annual folk festivals employ masks for storytelling and caricature, such as for the Dance of the Old Men and the Dance of the Moors and the Christians. The Eskimo make masks with comic or satiric features

By courtesy of the Museum of the American Indian, New York City



Agriculture fertility masks.
(Left) Corn husk mask of the Seneca Indians of the Iroquois nation of New York. Height 43 cm. (Right) Sekya, a *kachina* mask of the Zuni Indians of New Mexico. Painted leather, trimmed with feathers and hair. Height 43 cm.



Detail from "The Ridotto" by Pietro Longhi (1702–85), showing the masks popularly worn at fashionable Venetian functions of the 18th century. In the Galleria dell'Accademia Carrara, Bergamo, Italy.
SCALA—Art Resource

that are worn at festivals of merrymaking, as do the Ibos of Nigeria.

Theatrical uses. Masks have been used almost universally to represent characters in theatrical performances. Theatrical performances are a visual literature of a transient, momentary kind. It is most impressive because it can be seen as a reality; it expends itself by its very revelation. The mask participates as a more enduring element, since its form is physical.

Ancient Greek and Roman

The mask as a device for theatre first emerged in Western civilization from the religious practices of ancient Greece. In the worship of Dionysus, god of fecundity and the harvest, the communicants' attempt to impersonate the deity by donning goatskins and by imbibing wine eventually developed into the sophistication of masking. When a literature of worship appeared, a disguise, which consisted of a white linen mask hung over the face (a device supposedly initiated by Thespis, a 6th-century-BC poet who is credited with originating tragedy), enabled the leaders of the ceremony to make the god manifest. Thus symbolically identified, the communicant was inspired to speak in the first person, thereby giving birth to the art of drama. In Greece the progress from ritual to ritual-drama was continued in highly formalized theatrical representations. Masks used in these productions became elaborate headpieces made of leather or painted canvas and depicted an extensive variety of personalities, ages, ranks, and occupations. Heavily coiffured and of a size to enlarge the actor's presence, the Greek mask seems to have been designed to throw the voice by means of a built-in megaphone device and, by exaggeration of the features, to make clear at a distance the precise nature of the character. Moreover, their use made it possible for the Greek actors—who were limited by convention to three speakers for each tragedy—to impersonate a number of different characters during the play simply by changing masks and costumes. Details from frescoes, mosaics, vase paintings, and fragments of stone sculpture that have survived to the present day provide most of what is known of the appearance of these ancient theatrical masks. The tendency of the early Greek and Roman artists to idealize their subjects throws doubt, however, upon the accuracy of these reproductions. In fact, some authorities maintain that the masks of the ancient theatre were crude affairs with little aesthetic appeal.

Alinari—Art Resource



Masked Roman actors in a comedy scene, 1st-century relief from Pompeii.

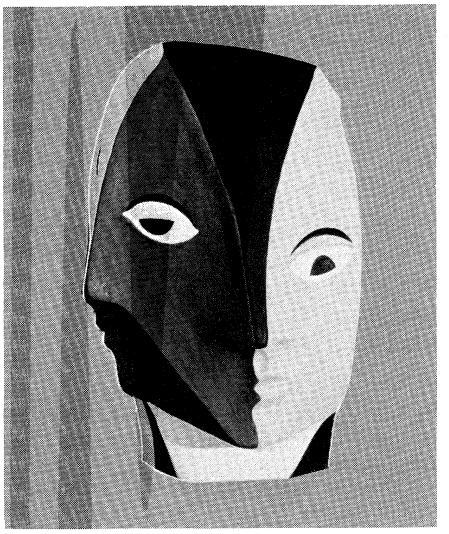
Medieval and Renaissance

In the Middle Ages, masks were used in the mystery plays of the 12th to the 16th century. In plays dramatizing portions of the Old and New Testaments, grotesques of all sorts, such as devils, demons, dragons, and personifications of the seven deadly sins, were brought to stage life by the use of masks. Constructed of papier-mâché, the masks of the mystery plays were evidently marvels of ingenuity and craftsmanship, being made to articulate and to belch fire and smoke from hidden contrivances. But again, no reliable pictorial record has survived. Masks used in connection with present-day carnivals and Mardi Gras and those of folk demons and characters still used by central European peasants, such as the *Perchten* masks of Alpine Austria, are most likely the inheritors of the tradition of medieval masks.

The 15th-century Renaissance in Italy witnessed the rise of a theatrical phenomenon that spread rapidly to France, to Germany, and to England, where it maintained its popularity into the 18th century. Comedies improvised from scenarios based upon the domestic dramas of the ancient Roman comic playwrights Plautus (254?–184 BC) and Terence (186/185–159 BC) and upon situations drawn from anonymous ancient Roman mimes flourished under the title of commedia dell'arte. Adopting the Roman stock figures and situations to their own usages, the players of the commedia were usually masked. Sometimes the masking was grotesque and fanciful, but generally a heavy leather mask, full or half face, disguised the commedia player. Excellent pictorial records of both commedia costumes and masks exist; some sketches show the characters of Arlecchino and Colombina wearing black masks covering merely the eyes, from which the later masquerade mask is certainly a development.

Modern

Except for vestiges of the commedia in the form of puppet and marionette shows, the drama of masks all but disappeared in Western theatre during the 18th, 19th, and first half of the 20th centuries. In modern revivals of ancient Greek plays, masks have occasionally been employed, and such highly symbolic plays as *Die versunkene Glocke* (*The Sunken Bell*; 1897) by the German Gerhart Hauptmann (1862–1946) and dramatizations of *Alice in Wonderland* have required masks for the performers of grotesque or animal figures. The Irish poet-playwright W.B. Yeats (1865–1939) revived the convention in his *Dreaming of the Bones* and in other plays patterned upon the Japanese Nō drama. In 1926 theatregoers in the United States witnessed a memorable use of masks in *The Great God Brown* by the American dramatist Eugene O'Neill (1888–1953), wherein actors wore masks of their own faces to indicate changes in the internal and external lives of their characters. Oskar Schlemmer (1888–1943), a German artist associated with the Bauhaus, became interested in the late 1920s and '30s in semantic phenomenology as applied to the design of masks for theatrical productions. Modern art movements are often reflected in the design of contemporary theatrical masks. The stylistic concepts of Cubism and Surrealism, for example, are apparent in the masks executed for a 1957 production of *La favola*



Mask by Amleto Sartori for the character of Vanna Scoma in Pirandello's *La favola del figlio cambiato,* 1957. Painted leather.
Antonello Perissinotto

del figlio cambiato (*The Fable of the Transformed Son*) by the Italian dramatist Luigi Pirandello (1867–1936). A well-known mid-century play using masks was *Les Nègres* (1958; *The Blacks,* 1960) by the French writer Jean Genet. The mask, however, has unquestionably lost its importance as a theatrical convention in the 20th century, and its appearance in modern plays is exceptional rather than general.

East Asian

In many ways akin to Greek drama in origin and theme, the Nō drama of Japan has remained a significant part of national life since its beginnings in the 14th century. Nō masks, of which there are about 125 named varieties, are rigidly traditional and are classified into five general types: old persons (male and female), gods, goddesses, devils, and goblins. The material of the Nō mask is wood with a coating of plaster, which is lacquered and gilded. Colours are traditional. White is used to characterize a corrupt ruler; red signifies a righteous man; a black mask is worn by the villain, who epitomizes violence and brutality. Nō masks are highly stylized and generally characterized. They are exquisitely carved by highly respected artists known as *tenka-ichi,* "the first under heaven." Shades of feeling are portrayed with beautifully sublimated realism. When the masks are subtly moved by the player's hand or body motion, their expression appears to change.

In Tibet, sacred dramas are performed by masked lay actors. A play for exorcising demons called the "Dance of the Red Tiger Devil" is performed at fixed seasons of the year exclusively by the priests or lamas wearing awe-inspiring masks of deities and demons. Masks employed in this mystery play are made of papier-mâché, cloth, and occasionally gilt copper. In the Indian state of Sikkim and in Bhutan, where wood is abundant and the damp climate is destructive to paper, they are carved of durable wood. All masks of the Himalayan peoples are fantastically painted and are usually provided with wigs of yak tail in various colours. Formally they often emphasize the hideous.

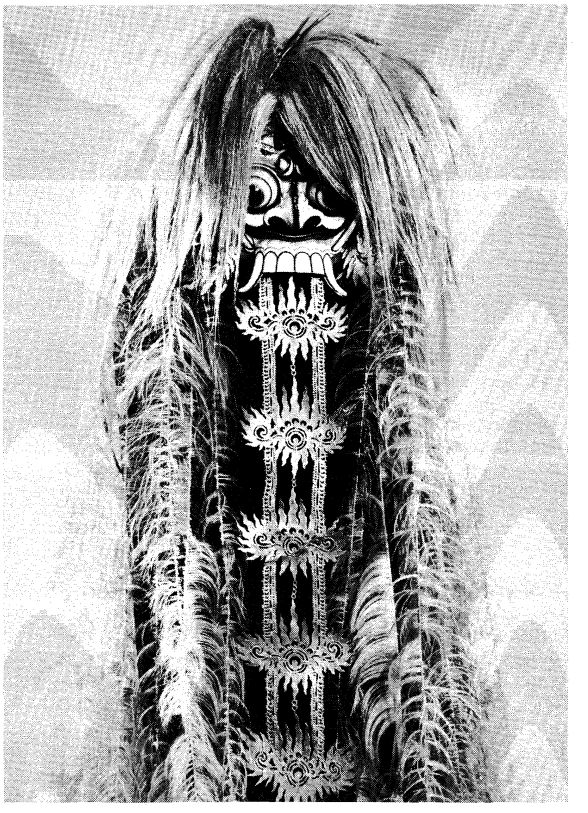
Masks, usually made of papier-mâché, are employed in the religious or admonitory drama of China; but for the greater part the actors in popular or secular drama make up their faces with cosmetics and paint to resemble masks, as do the Kabuki actors in Japan. The makeup mask both identifies the particular character and conveys his personality. The highly didactic sacred drama of China is performed with the actors wearing fanciful and grotesque masks. Akin to this "morality" drama are the congratulatory playlets, pageants, processions, and dances of China. Masks employed in these ceremonies are highly ornamented, with jewelled and elaborately filigreed headgears. In the lion and dragon dances of both China and Japan, a stylized mask of the beast is carried on a pole by itinerant players, whose bodies are concealed by a dependent cloth. The mask and cloth are manipulated violently, as if the animal were in pursuit, to the taps of a small drum. The mask's lower jaw is movable and made to emit a loud continuous clacking by means of a string.

Indonesian

On Java and Bali, wooden masks, *tupeng,* are used in certain theatrical performances, called *wayang wong.* These dance dramas developed from the shadow puppet plays of the 18th century and are performed not only as amusement but as a safeguard against calamities. The stories are in part derived from ancient Sanskrit literature, especially the Hindu epics, although the Javanese later became Muslims. The brightly painted masks are made of wood and leather and are often fitted with horsehair and metallic or gilded paper accoutrements. They are ordinarily held in the teeth by means of a strap of leather or rattan that has been fastened across the inside. Occasionally an actor interrupts the unseen narrator, the Dalang, who is speaking the play. The mask is then held in front of the face while the player says his line. The use of theatrical masks in Java is exceptional, since masks, being forbidden under the prohibition of images, are practically unknown in the Islāmic world.

In the 20th century, with the breaking down of primitive and folk cultures, the mask has increasingly become a decorative object, although it has long been used in art as an ornamental device. In Haiti, India, Indonesia, Japan, Kenya, and Mexico, masks are largely produced for tourists. The collecting of old masks has been a part of the current interest in so-called primitive and folk arts. Masks also have exerted a decided influence on modern art movements, especially in the first decades of the 20th century, when painters in France and Germany found a source of inspiration in the tribal masks of Africa and western Oceania.

By courtesy of the Tropenmuseum, Amsterdam



Javanese *tupeng* mask representing the witch Rangda. Lacquered wood, cloth, metal, and horsehair.

BIBLIOGRAPHY. "Masks" in the *Encyclopedia of World Art,* vol. 9, col. 520–570 (1964), a good historical survey; WILLIAM N. FENTON, "Masked Medicine Societies of the Iroquois," *Smithsonian Institution, Annual Report for 1940,* pp. 397–429 (1941), a very good general discussion of Iroquois masks, with illustrations; MARCEL GRIAULE, *Masques Dogons,* 2nd ed. (1963), a profusely illustrated classic study of the masks of the

Dogon, a people of Mali, within their cultural setting; GEORGE W. HARLEY, *Masks as Agents of Social Control in Northeast Liberia,* Peabody Museum Papers 32, no. 2 (1950, reprinted 1975), a useful, illustrated article on this topic; EDWARD A. KENNARD, *Hopi Kachinas,* 2nd ed. (1971), an important study; DOROTHY J. RAY, *Eskimo Masks: Art and Ceremony* (1967), one of the best studies of Eskimo masks, with many fine plates and a bibliography; F.E. WILLIAMS, *Drama of Orokolo* (1940, reprinted 1969), a classic study of masks of the Gulf of Papua, New Guinea, with fine illustrations and a bibliography; MALCOLM KIRK, *Man as Art: New Guinea* (1981), with especially good photographs; DONALD B. CORDRY, *Mexican Masks* (1980), a study of how Mexican masks are related to both the European and the Indian traditions; SIMON OTTENBERG, *Masked Rituals of Afikpo* (1975), a survey of a Nigerian masquerade tradition; LEON UNDERWOOD, *Masks of West Africa* (1948), a small book but important for the subject, with good plates.

(P.S.W.)

The Foundations of Mathematics

The study of the foundations of mathematics has dealt with the concepts and the assumptions about those concepts with which mathematics starts. Especially since 1900, foundational investigations have come to include also an inquiry into the nature of mathematical theories and the scope of mathematical methods.

This article is divided into the following sections:

The axiomatic method 594
  Geometry according to Euclid 594
  Non-Euclidean geometry 594
  Formal axiomatic method 594
The genetic method 595
  Arithmetic and analysis 595
  Cardinal number and the theory of sets 595
The crisis in foundations after 1900 596
  The paradoxes 596
  Intuitionism 596
    Historical background of Intuitionism
    Basic features of Intuitionism
    The criticisms of Realism and Formalism
    Redevelopment of mathematics
    Intuitionist logic
  Logicism, Formalism, and the metamathematical method 598
    Logicism
    Formalism
    Turing's computing machine and formal systems
    Church's theorem: logical systems without decision procedures
    The effect of Gödel's theorems on Formalism
    Degrees of decidability in arithmetic
Current directions 602
  Intuitionistic foundations of mathematics 602
  Non-Intuitionistic studies on the foundations of mathematics 602
Bibliography 602

THE AXIOMATIC METHOD

Geometry according to Euclid. Euclid's *Elements,* a treatise written about 330 BC on the foundations of mathematics as well as on geometry, is said to have had a greater circulation than any book in history except the Bible. Euclid employs the deductive method, which had come down from Pythagoras two centuries earlier.

Euclid's use of the deductive method

First, Euclid introduces fundamental concepts such as point, line, plane, and angle. He relates these to physical space by such definitions as "A line is length without breadth"; but, since their definitions play no further role in his system, they constitute primitive or undefined terms.

Second, Euclid asserts certain primitive propositions or postulates about these primitive terms. These are propositions that the reader is asked to accept as true immediately on the basis of the definitions of the primitive terms relative to the physical world.

Third, Euclid deduces from the postulates further propositions, called theorems, such as the theorem of Pythagoras, "The sum of the squares constructed on the arms of a right triangle is equal to the square constructed on the hypotenuse." That is, he shows that their truth follows logically from the truth of the primitive propositions. While doing so, he also introduces defined terms, such as triangle, right triangle, and hypotenuse, which are single words that can take the place of phrases using besides ordinary language only the undefined terms.

The net result is that the reader, starting with postulates of which he is supposed to be immediately convinced, is led, by a long series of simple steps each intended to be convincing, to accept the truth of other propositions, of which he would be far from convinced had they been asserted at the outset.

Non-Euclidean geometry. Euclid's parallel postulate, on which depends the theorem "Through a point P not on a line L there can be drawn exactly one line parallel to L," had long seemed less evident than his other postulates. Attempts to prove it as a theorem from the other postulates culminated in the 19th century in the independent discoveries by the Russian and Hungarian mathematicians Nikolay Ivanovich Lobachevsky and János Bolyai that a different parallel postulate can be substituted to give a non-Euclidean geometry in which instead: "For some plane, some line L in the plane, and some point P in the plane and not on L, there exist at least two distinct lines in the plane passing through point P and not intersecting line L." Bolyai was aware of the noncontradictory character of the new geometry in 1825 and Lobachevsky possibly in 1826, though the publications of both men were delayed.

Insofar as Euclidean geometry fits physical space, it is impossible to say that the geometry of Lobachevsky and Bolyai does not also fit. It is only necessary that the discrepancies between the two be too small to show up in any measurements that can be made. So, while Euclid thought of his geometry as a true description of physical space, it is necessary to regard it as one of a number of different possible mathematical models of that space.

That a theorem of Euclidean geometry be exactly true can only be a property of the geometry as a logical structure. A person who accepts Euclidean geometry as a valid logical structure is equally obliged to accept the geometry of Lobachevsky and Bolyai. For example, as the leader in German mathematics Felix Klein observed in 1871, the postulates of Lobachevsky and Bolyai's plane geometry are true in Euclidean plane geometry when in those postulates the plane is interpreted to be the interior of a circle, the lines to be the chords of that circle, and congruence, distance, and angle are defined by a method given in England by Arthur Cayley in 1859.

Formal axiomatic method. Here the meaning of the undefined terms of a postulational theory has been varied while holding the deductive structure fixed, a procedure that is characteristically modern. One can deduce a theory from its postulates without saying at all what the undefined terms are to mean. The theory applies to any system of objects that when taken as supplying the meanings of the undefined terms make the postulates true.

Manipulation of undefined terms

Indeed, in a postulational theory, one should be able to perform the deductions treating the undefined terms as meaningless. For, if their meanings had to be used, whatever had to be taken from their meanings should instead be provided by additional postulates. In Euclid's *Elements* some properties of the undefined terms did in fact enter from their meanings via the figures. These hidden assumptions were brought to light and stated as postulates by the German mathematicians Moritz Pasch in 1882 and David

Hilbert, the latter in his *Grundlagen der Geometrie* (1899; *The Foundations of Geometry,* 1902). Thus, Hilbert states as a postulate, "If a line L cuts the side AB of a triangle ABC and does not pass through A, B or C, then it must also cut AC or BC."

The postulational method when the undefined terms are treated as meaningless is called the formal postulational method, or formal axiomatic method. In this connection, "axiom" and "postulate" are often used synonymously, in contrast with the material axiomatic method of Euclid.

It might be thought that meaninglessness of the undefined terms would rob a theory of its interest, but the opposite can be the case. Thus, for some theories many different interpretations of their undefined terms each make the postulates true (*e.g.,* Euclidean geometry without the parallel postulate or in algebra abstract groups). The deduction of such a theory independently of any interpretation makes it a mathematical tool prepared in advance for diverse applications. If in essential respects only one such interpretation exists, the postulates are categorical; that is, admitting of no further qualifications. If there is no such interpretation, the theory is without value, and one of the problems for foundations is to exclude such theories.

THE GENETIC METHOD

Arithmetic and analysis. Systems of numbers are more often introduced by the genetic method, so called by Hilbert. In this the objects are generated or constructed in a certain orderly manner, and the theorems express properties of the system of the objects assured by this mode of generation, the term genetic being taken from biology. Of course, deductive reasoning is used, just as in the axiomatic method, to obtain unobvious properties from obvious or immediate ones, and the immediate properties can be listed as postulates. Thus, the natural numbers 0, 1, 2, $\cdots$, used in counting discrete objects, are describable genetically as the objects that are generated by starting with a first object, 0, and by proceeding from any object, n, already generated to the next object, $n+1$, objects differently generated being distinct. A suitable selection of immediate properties constitutes the five postulates of Giuseppe Peano, an Italian mathematician and logician, expressed in 1889, which are categorical for the arithmetic of the natural numbers. By starting from these postulates, rather than from the generation, that arithmetic can then be considered as a formal axiomatic theory. But the genetic method will be used here in describing some other mathematical systems of objects.

The word arithmetic has several senses; but one of the most common of them is the theory of the natural numbers and of similar systems of numbers, such as the integers (an extension of the natural numbers, by way of including negatives of natural numbers) $\cdots$, -2, -1, 0, 1, 2, $\cdots$, or the rational numbers such as $\frac{1}{2}$, $\frac{2}{3}$, $\frac{4}{5}$, which are the numbers expressible as fractions p/q in which p and q are integers with q not 0.

Analysis deals with such systems of objects as the real numbers, used in the measurement of continuously variable magnitudes. A theory of real numbers not resting on vague geometrical imagery was only achieved in the latter part of the 19th century, in Germany and France, by Karl Weierstrass, Richard Dedekind, Charles Méray, and Georg Cantor. A version using infinite decimals will suffice here. For example, $\frac{9}{4} = 2.25 = 2.24999\cdots$, $-\frac{9}{4} = -2.25 = -3+0.75 = -3+0.74999\cdots$, $1 = 0.99999\cdots$, $\sqrt{2} = 1.41421\cdots$, $\pi = 3.14159\cdots$. In each of these examples, a real number x is expressed as a sum $X + 0.x_0x_1x_2x_3x_4\cdots$ of an integer X and a nonterminating decimal fraction $0.x_0x_1x_2x_3x_4\cdots$ (*i.e.,* one that does not have all digits 0 after some place). The real numbers are given genetically, each number uniquely, as all possible such sums.

Cardinal number and the theory of sets. If, in a tribe of aborigines unable to count above 20, whose chief is by custom the member with the most numerous herd, there were two contenders for the chieftaincy, the tribe could decide the winner (or that the two contenders are tied) by passing their herds, no matter how large, through a gate, a pair of animals, one from each herd, at a time.

Cantor's contribution

Cantor was the first who applied this idea systematically to the comparison of infinite collections. A one-to-one correspondence between two collections, or sets, of objects S and T is a collection of ordered pairs in which the first member of each pair belongs to S, the second member to T, and each member of S occurs as first member, each member of T as second member, of just one pair. Two sets have the same cardinal number if there is a one-to-one correspondence between them.

The natural numbers 0, 1, 2, $\cdots$, the positive integers 1, 2, 3, $\cdots$, the squares of the positive integers 1, 4, 9, $\cdots$ (an example used by Galileo in 1638), and the integers 0, 1, -1, 2, -2, $\cdots$ have the same cardinal number. Infinite sets with this cardinal are countable (or countably infinite, to distinguish them from finite sets, which are also countable), because they can be counted off or listed by using the natural numbers. A fifth example of a countably infinite set is the rational numbers; for each rational can be written as a positive or negative ratio of integers, $(-1)^n p/q$, with the least possible natural numbers n, p, q, and then the rationals can be enumerated by various techniques in order to make it explicit that they are countable. One such technique is to enumerate $(-1)^n p/q$ in the order of magnitude of the positive integers $2^n 3^p 5^q$.

A part, or subset, of a set S is a set each member of which is a member of S. As Galileo had noted, a part not the whole of an infinite set can be in one-to-one correspondence with the whole. (Dedekind in 1888 used this property to distinguish infinite sets from finite sets.) A set T is of greater cardinal number than another set S if there is a one-to-one correspondence between S and a subset of T but not vice versa.

Cantor showed in 1874 by a method now called the diagonal method that there are sets of greater cardinal number than the natural numbers (see Box, 1–4).

Thus, infinity is not simple, but indeed there are a whole host of infinite or transfinite cardinal numbers. This result is only one of many in Cantor's theory of sets, which greatly extended the mathematical horizon, besides seeming to provide the ideas for a general foundation of mathematics.

(1) A (single-valued) number-theoretic function f of one variable x is a correspondence by which, to each natural number x, there corresponds a unique natural number $f(x)$, called the value of f for x as argument. The set of these functions is uncountable, as can be shown by the following reasoning. For any infinite list $f_0, f_1, f_2, \cdots$ of such functions, the function f defined by $f(x) = f_x(x) + 1$ (so the values of f come by altering every value in the diagonal, if the values of $f_0, f_1, f_2, \cdots$ have been written in successive rows) is not in the list; for if f were f_q, the contradictory equation $f_q(q) = f_q(q) + 1$ would be obtained upon taking the value for q as argument. The set of these functions is of greater cardinal number than the natural numbers; for on the other hand, the natural numbers n are in one-to-one correspondence with the constant functions $f(x) = n$.

(2) In the generation of the real numbers using infinite decimals, the nonterminating decimals can be construed as number-theoretic functions taking values ≤ 9; so the reals are uncountable, by reasoning the same as above except taking $f(x) = 8$ when $f_x(x) = 9$.

(3) A given set of natural numbers is represented by a function which takes 0 or 1 as value according as the argument is in the set or not; so the sets of natural numbers are uncountable, by reasoning the same as above except taking $f(x) = 0$ when $f_x(x) = 1$.

(4) It is not hard to show that these three uncountable sets have the same cardinal. The third illustrates Cantor's theorem, "The cardinal number of the set of the subsets of a set S is greater than the cardinal number of S."

THE CRISIS IN FOUNDATIONS AFTER 1900

The paradoxes. In the last decade of his scientific life Cantor became aware that his theory of sets pushed too far leads to contradictions; two may be described.

If S be the set of all sets and T the set of the subsets of S, then, since T corresponds one-to-one to itself as a subset of S, it cannot have a greater cardinal than S. Yet by Cantor's theorem it must; this is Cantor's paradox, stated in 1899.

The phrases in the English language are countable. After listing all the phrases, strike out those that do not define number-theoretic functions. Application of Cantor's diagonal method to the list remaining enables one to define in the English language a function that differs from every function definable in the English language. This paradox was found by J.A. Richard in 1905.

Richard's discovery

These modern paradoxes have a counterpart in the ancient paradox of the liar, which in one form appears in a paradoxical statement attributed to Epimenides (6th century BC) and in another is attributed to Eubulides (4th century BC), who has a person say, "What I am now saying is a lie." The quoted sentence if true must by what it says be false, and if false be true.

No one has succeeded in disposing of the paradoxes by simply pointing out a fallacy upon which all can agree.

Cantor's paradox is evidently connected with the use of too-large sets. It is difficult, however, to determine which sets are too large and to replace Cantor's definition "any collection of definite well-distinguished objects of our perception or our thought into a whole."

In fact, it is possible to reformulate set theory to exclude too-large sets. In doing so, axioms are listed to govern just what sets shall be allowed; and set theory becomes a formal axiomatic theory instead of a genetic one. The first system of axiomatic set theory was given in Italy by Ernst Zermelo in 1908 and improved in Germany and Scandinavia by Abraham Adolf Fraenkel and Thoralf Skolem in 1922–23. Axiomatic set theory provides probably the simplest system in which most existing mathematics, but not the paradoxes as far as is known, can be deduced. The Austrian logician Kurt Gödel in 1938 showed that Cantor's continuum hypothesis, "The cardinal number of the real numbers is the least cardinal number greater than that of the natural numbers," can be added as a new axiom to set theory without introducing contradiction, and Paul J. Cohen of the United States in 1963 showed the like for its negation.

That the paradoxes arose shows there is some defect in the methods, which hitherto had been trusted, of constructing and reasoning about mathematical objects. Transforming the theory of sets from a genetic into a formal axiomatic theory leaves over the problem of whether there is an interpretation of the undefined terms that makes the axioms in some sense true. To ignore these issues is to adopt a mathematical nihilism. Formally axiomatized mathematics cannot be all of mathematics. To draw upon another formal axiomatic theory for such an interpretation is a regress. At some level there must be a meaningful mathematics. At the very least it must be believed that the theorems of a formal axiomatic theory really do follow from the axioms. As a further example, when mathematicians say that by a certain method it can be determined in a finite number of steps whether an equation $ax+by+c=0$ with given integral coefficients, a, b, c has a solution for x, y in integers, they certainly do not mean merely that this proposition can be deduced from certain axioms.

The great French mathematician and philosopher of science Henri Poincaré and the outstanding English mathematician and philosopher Bertrand Russell about 1905 concluded that the explanation of the paradoxes lies in a certain circularity of definition. When a set M and an object m are so defined that m is a member of M but is defined only by reference to M, the definition of M or of m is called impredicative. An impredicative definition can be found in each of the paradoxes. Thus, the sickness of mathematics could be cured by banning impredicative definitions, if it did not inconveniently turn out that the cure would kill the patient; for, as was especially emphasized by the German-born U.S. mathematician and physicist Hermann Weyl in his book *Das Kontinuum* (1918; "The Continuum"), there are impredicative definitions also in analysis. (S.C.K.)

Explanation of the paradoxes

Intuitionism. In the first quarter of the 20th century three main schools of thought, called Logicism, Formalism, and Intuitionism, arose to account for and resolve the crisis in the foundations of mathematics. Each of these is a version of a more or less traditional view of the nature of mathematics. The first is a form of Platonic Realism, according to which mathematics investigates a realm of abstract objects (such as points, numbers, sets, and the like) that "exist" external to and independent of the investigator. In Logicism it was argued that all mathematical notions are reducible to the idea of abstract property and that mathematics is derived from basic logical principles concerning properties; this is discussed in more detail below. The attempted reduction in Logicism to evident principles was not completely successful, but the work on this led to an axiomatization of Cantor's notion of set (or class) as a basis for current mathematics. Formalism holds that mathematics consists simply of the manipulation of finite configurations of symbols according to prescribed rules. Some of these configurations may have concrete meaning; others are considered as meaningless. In the Formalist view, the choice of rules is dictated by pragmatic considerations or, even better, by theoretical justification of their use in deriving concretely meaningful statements; this is also discussed below. Finally, Intuitionism is a version of the conception of mathematics as a self-sustained intellectual activity that deals with mental constructions governed by self-evident laws. These three schools of thought were principally led, respectively, by Russell, Hilbert, and the Dutch mathematician and philosopher L.E.J. Brouwer.

Mathematical logic was considerably developed in order to make these positions precise and study them more deeply. Set theory was allied with the increasingly general and nonconstructive tendency in mathematics. On the other hand, Intuitionism led to a considerable redevelopment of portions of mathematics in constructive terms, to eliminate what were regarded as philosophically untenable features. It and Formalism together bordered on the development of the important notion of mechanically computable function.

Historical background of Intuitionism. Immanuel Kant, an 18th-century philosopher who was the founder of critical philosophy, is often cited as a progenitor of Intuitionism in mathematics, inasmuch as he viewed space and time as forms that the mind imposes upon human sensibility or perception and, thus, as a priori and Idealistic. Leopold Kronecker, a 19th-century mathematician specializing in number theory and algebra, is also viewed as a progenitor because of his attitude to existence proofs. These typically give information about possible solutions to an equation or condition, such as the existence of at least one function that solves a differential equation or the existence of some point at which a continuous function achieves its maximum value in an interval. Kronecker doubted the significance of proofs, of a sort increasingly used in mathematics, that established the existence of such solutions by indirect arguments without necessarily being able to construct any one of them explicitly or even provide a finite procedure for finding one. In the beginning of the 20th century there was a French semi-Intuitionistic school of mathematics led by Émile Borel, which concentrated on explicit constructions in mathematical analysis. During the same period, Poincaré was very influential through his informal writings, where he stressed the basic intuition of the sequence of integers and questioned the use of certain definitions selecting members from assumed completed infinite totalities (the so-called impredicative definitions mentioned above).

Kronecker's views

Basic features of Intuitionism. The basic concepts and results of Intuitionism in its full modern development were laid out, however, by Brouwer, who is, therefore, now considered to be the founder of the school. He regarded the starting point for mathematics to be the intuitive understanding of the generation of the sequence of (positive) integers, that is, of the repeated duplication, $|$, $||$, $|||$, $\cdots$ of a single entity considered abstractly.

Intuition of the integers

Brouwer believed that this was derived from a more basic inner awareness of time, whose unity is separated into a "two-ity" by the present moment; this is related to Kant's views as to the a priori status of time. Brouwer also took this idea of repeated division to be the intuitive source of the generation of sequences of rationals which make up the real number system. Thus, in contrast to the Platonic Realist position, the infinite collections considered were conceived of as potential rather than completed infinite totalities. On the Platonist view it is true to say, *e.g.*, that either the sequence 0123456789 occurs at some position in the decimal expansion $3.14159 \cdot \cdot \cdot$ of π or there exists no such position; one can further proceed to define a number n as the least position where the first case holds, if indeed it holds, and as 0 otherwise. In opposition to the Platonist conception, Brouwer restricted mathematical knowledge to that which can be known about mental constructions and proofs. For him, to know a proposition P one must have a proof of it; to know its negation not-P one must have a disproof of it; that is, a proof that P leads to a contradiction. For the proposition P that expresses that 0123456789 occurs at some position in the decimal expansion of π, there is at present neither a proof of P nor a disproof of P; for this reason there is not an explicit construction of the number n above.

More generally, on the Platonist position, one accepts for any proposition P (even ones which refer to infinite totalities) the law of the excluded middle, P or not-P; in other words, that P is true or P is false. This cannot be accepted in general by the Intuitionist in his interpretation since he cannot be sure that for every P one shall be able to find either a proof of it or a disproof of it. From this point of view, some of the most fundamental arguments and definitions in mathematical analysis and other branches of mathematics become unsupported, and these subjects require a complete re-examination.

Brouwer's work on Intuitionism falls into several distinct periods: the first, beginning with his thesis in 1907, was primarily negative, with vigorous criticisms of the Realist (Cantorian and Logicist) as well as Formalist viewpoints; and the second, in the 1920s, was positive, being occupied with the reconstruction of mathematics according to the Intuitionistic tenets. After a long period of silence, Brouwer returned to the scene in the late 1940s with his "theory of the creative subject," which was even more radically subjectivist than his previous ideas.

The criticisms of Realism and Formalism. Brouwer's criticism of the Realist position concentrated on the application therein of the so-called classical logic of propositions and predicates referring to infinite totalities. In particular, by the law of the excluded middle in classical logic, if n is used to symbolize any member of the infinite set of integers $1, 2, 3, \cdot \cdot \cdot$ and $P(n)$ denotes a property or predicate P of n, one must admit:

(A) "Either there exists an n such that $P(n)$ holds, or, for all n, $P(n)$ fails."

One example of this has already been considered in the preceding section.

For another example, each real number that is between 0 and 1 (possibly equalling 0 or 1) has some representation $r = 0.a_1a_2a_3 \cdot \cdot \cdot$ where each $a_n = 0, 1, \cdot \cdot \cdot, 9$. If $P(n)$ is the property "a_n is less than 9," then because $r = 1$ just in case each a_n is equal to 9 ($1 = 0.999 \cdot \cdot \cdot$), the general logical law (A) reduces to the statement in this case that $r \neq 1$ or $r = 1$. Similarly, one may use (A) to justify "r is less than s or r is greater than or equal to s" for any real numbers r and s, which is basic to the classical treatment of mathematical analysis. One may then go on to define discontinuous functions f of real numbers r, such as the so-called step function f that has the value 0 if r is less than 1 and 1 if r is greater than or equal to 1.

Of course the law (A) can be checked directly when n ranges over a given finite set of integers and $P(n)$ is a decidable property; *i.e.*, there is available a method that will tell for each n whether or not $P(n)$ holds. Brouwer argued that the assumption of (*A*) is based on an unjustified extrapolation of the experience with finite sets to the infinite.

Brouwer's criticism of Formalism was principally directed at Hilbert's formulation of that position, described in more detail below. Briefly, Hilbert's idea was to justify all of mathematics on the basis of the most elementary and concrete methods of finite combinatorial reasoning. The vehicle for this was to be the representation of mathematical statements as strings of symbols in formal logical systems.

The justification was to consist in giving Finitist proofs of consistency of such formal systems and, thence, of their correctness for deriving finitistically meaningful statements. Brouwer did not object to the use of Finitist reasoning, because this is just a part of what is Intuitionistically acceptable. The difference is that Finitism uses constructions only on concretely conceived objects such as the integers, whereas Intuitionism uses them also on abstractly conceived objects, such as proofs and constructions. Brouwer saw no special reason for the proposed restriction to Finitist methods. But his main objection to Hilbert's program was that in his view mere consistency of formal systems of mathematics could not justify them; only a prior understanding of the symbols of these systems and direct intuitive verification of the axioms and rules could do that. Moreover, one could not expect to pin down, once and for all, in any formal system the nature of the ongoing creative process of concept formation and discovery in mathematics. Brouwer's criticisms of Hilbert's program were in a way subsequently borne out by the famous work of Kurt Gödel in 1931, showing the inherent limitations of that program (see below *Logicism, Formalism, and the metamathematical method*).

Brouwer's objection to Hilbert's program

Redevelopment of mathematics. The system of real numbers forms the basis of the calculus and more generally of mathematical analysis; this, in turn, is essential for the solution of the equations employed in scientific applications such as the differential equations of mathematical physics. It seemed to many that the Intuitionistic abandonment of classical logic would undermine a vast body of knowledge. Nevertheless, Brouwer set out to re-create mathematics in accordance with his tenets. His positive contributions to this reconstruction included a new theory of real numbers and of real functions (real-valued operations on real numbers), a theory of sets (called species), and a theory of ordinals (finite and transfinite construction stages). The groundwork was thus laid for a redevelopment of substantial portions of classical algebra, analysis, geometry, and topology. Even so, Brouwer attracted few converts to the Intuitionistic position thereby. Some of the reasons for this lie in the character of his expositions, which in the first period were polemical and in the second (technical) period were rather frequently obscure, involving unfamiliar or even "suspect" notions and familiar terminology with new meanings. This may be illustrated with Brouwer's assertion:

(B) Every real function is continuous.

Continuity of a real function

This assertion is patently contradictory with the classical interpretation since, for example, the step function described above is not continuous. To understand the Intuitionistic interpretation of (B), one must first explain how real numbers are to be considered. In the classical interpretation, every real number r is the limit of a convergent sequence s of rational numbers $s_1, s_2, s_3, \cdot \cdot \cdot$; this is indicated by the symbolism $s = \{s_n\}$. There are many possible such s for each real r; one, for example, is given by the sequence $s_n = a_0 + a_1/10^1 + \cdot \cdot \cdot + a_n/10^n$ when $r = a_0.a_1a_2a_3 \cdot \cdot \cdot$ is written in decimal form. Intuitionistically, one can operate directly only with the sequences $s = \{s_n\}$, which are called real-number generators. Two such sequences $s = \{s_n\}$ and $s' = \{s'_n\}$ are said to be equivalent, and one writes $s = s'$ if the difference $s_n - s'_n$ approaches 0 as n increases without bound. (This relation is to be distinguished from that of strict identity, where, for each n, $s_n = s'_n$.) A real function f can only be given by a construction F that associates with each generator s another generator $t = F(s)$ such that if $s = s'$ (in the sense just explained) then $F(s) = F(s')$.

Now each specific real-number generator $s = \{s_n\}$ must also be given by a construction that determines s_n given n. In practice, however, one does not make use of a rule for s when determining the terms of the generator $F(s)$, rather

only the successive values $s_1, s_2, s_3, \cdots$, so to speak, as they appear. Brouwer introduced the new concept of infinitely proceeding sequence (ips); roughly speaking, this is any sequence of which one can determine as many values as desired, although at any stage not more than an initial segment need be known. Every constructively defined s is an ips, but one may also think of s that are generated by some random processes or free acts of choice. Any construction that assigns to each ips s a rational number can only make use of some finite amount of information $s_1, s_2, \cdots, s_n$ about s. Brouwer's assertion (B) easily follows from this general continuity principle. Of course there is no constructive F that determines a step function f; for otherwise it would follow that one could determine for any ips s whether or not s is less than 1 simply on the basis of some finite initial segment of s. This, of course, is absurd, because the remaining terms $s_{n+1}, s_{n+2}, \cdots$ could radically alter the situation.

With the new meanings given to the basic notions of analysis in this way, each classical theorem must be reexamined to see whether it has a correct Intuitionistic version. For example, a classical result often used is that if f is a continuous real function defined on a closed interval, then $f(r_0)$ has a maximum possible value for some r_0 in that interval (and similarly for a minimum). This is not Intuitionistically demonstrable, because one can constructively define a function f such that there is no constructive way to determine an r_0 at which f attains its maximum value. If f satisfies stronger conditions (such as differentiability) that are frequently met in applications, however, it is possible to obtain the existence of such r_0 constructively as well. When a classical result has a correct Intuitionistic version it usually requires a sharpening of the hypotheses, but for existential results one then also gets a sharpening of the conclusion.

Intuitionistic logic. The assertions that are made in Intuitionistic mathematics are supposed to be reports of completed proofs that have been carried out by a series of steps that are evident for the objects conceived when combined by suitable (mental) constructions. It is of interest to compare in general terms the type of reasoning that is admitted under this conception with that which makes use of classical logic. For this one needs an explanation of the conditions under which one has a constructive proof (or grounds) for a compound assertion; the following is based on the explanation provided by Heyting. Here P, Q denote any propositions, S is any given set, the variable x ranges over the objects of S, and $P(x)$ is a predicate of x.

Conditions for Compound Assertions

FORM OF ASSERTION	GROUNDS FOR THE ASSERTION
(a) *P or Q*	(a) At least one of P, Q has already been proved.
(b) *P and Q*	(b) Both P and Q have been proved.
(c) *P implies Q*	(c) One has a construction C of which it has been proved that whenever C is applied to any possible proof of P then the result is a proof of Q.
(d) *not P*	(d) The same as for "P implies $1 \neq 1$"; that is, every possible proof of P is shown to be transformed into a contradiction.
(e) *there exists an x such that P(x)*	(e) One has constructed an s in S and proved $P(s)$.
(f) *for all x, P(x)*	(f) One has a proof which is shown to specialize to a proof of $P(s)$ for each s in S.

It should be understood that these do not constitute an explanation of the logical connectives in terms of prior, logic-free notions. They do show, however, the circumstances under which the connectives are used in practice, and they indicate the form of such an explanation in the strict sense of the word.

The conditions given do not justify the law of excluded middle P or not-P, because, as already discussed, there is no reason to expect that for any proposition P it is possible either to find a proof of it or to show how to derive a contradiction from it. If one accepts Brouwer's notion of ips, however, and the general continuity principle for these, one can actually disprove this classical law. This is true because, as has been pointed out, this implies the general law (A) above, of which the following is then a special case:

(C) For every ips $s = \{s_n\}$, either there exists an m such that $s_m = 0$ or for all m, $s_m \neq 0$.

A consequence is the following:

(D) For every ips $s = \{s_n\}$, there exists an m such that either $s_m = 0$ or for all p, $s_p \neq 0$.

By the general continuity principle, one should be able to associate with any s, two numbers n and m such that for any s' with $s'_1 = s_1, \cdots, s'_n = s_n$, either $s'_m = 0$ or for all p, $s'_p \neq 0$; one can also assume that m is less than or equal to n. If one starts with a sequence s of nonzero terms, however, one can always define s', which agrees with s on the initial segment $1, \cdots, n$ but is such that $s'_{n+1} = 0$. Thus (D), and hence also the law of excluded middle, leads to a contradiction. In other words, the following has been shown:

(E) It is not the case that (C) holds.

By a similar argument the following can be shown:

(F) Not for all ips s and t, either s is less than t or t is less than or equal to s. (S.Fe.)

Logicism, Formalism, and the metamathematical method. While the philosophical debate at the foundations of mathematics was identified for many years with the debate among the three schools of thought, Logicism, Formalism, and the Intuitionism considered above, a broadened basis for reasoning followed from the work of Gödel, Church, and Turing in the 1930s. Thus, a contemporary presentation of the problems at the foundations of mathematics is coloured by the automata-based reasoning that is identified with Turing's name. Whereas Intuitionism yielded somewhat to Formalism at the hands of Heyting, the techniques made available to Logicism and Formalism were enormously broadened with the introduction of Turing's machine (see below *Turing's computing machine and formal systems*). While the lines separating the traditional three schools of thought have become less distinct since mid-20th century, it now seems natural to present the points of view with Intuitionism on the one hand and Logicism and Formalism, as transformed by the metamathematical method (in which the object of a mathematical study is a mathematical system), on the other. Such a division should not be interpreted to mean that metamathematical methods have not been fruitfully applied in studies of systematizations arising from the Intuitionistic schools.

Impact of Turing's machine

Logicism. The Logicistic thesis is that mathematics is a branch of logic. A German logician and mathematician, Gottfried Wilhelm Leibniz, in 1666 had conceived of logic as a universal science embracing the principles underlying all others. Richard Dedekind in 1888 and, especially, Gottlob Frege in 1884 and 1893, both of Germany, derived arithmetic from logic. The thesis that all of mathematics is reducible to logic alone is due to Russell.

Russell, after rediscovering results of Frege, had to make provision against the paradoxes. This provision took the form of a ramified theory of types. Roughly described, the primary objects not being subjected to logical analysis he took to be of one type—say, type 0—properties of objects of type 0 to be of type 1, properties of properties of type 1 to be of type 2, etc.; and he excluded from use any properties not belonging to one of these types. Then, to exclude impredicative definitions within a type, he separated the types above type 0 into orders, the properties defined using the totality of properties of a given order being placed in the next higher order. Finally, to overcome the obstacle that exclusion of impredicative definitions opposes to analysis, he postulated his axiom of reducibility, which asserts that to any property belonging to an order above the lowest there is a property of order 0 possessed by exactly the same objects.

The theory of types

On this basis, mathematics was deduced from logic, in a symbolic language, in the monumental *Principia Mathematica,* by the English mathematician Alfred North Whitehead and Bertrand Russell, in 1910–13.

The use of the axiom of reducibility weakens the case

for the Logicistic thesis. For, if the axiom be true, should not the reductions to order 0 be constructed instead of being merely postulated as existing? Weyl hence remarked in 1946 that in *Principia Mathematica* "mathematics is no longer founded on logic, but on a sort of logician's paradise."

Leon Chwistek in 1921 and F.P. Ramsey in 1926 introduced a simple theory of types in which the separation into orders is suppressed. This theory still allows the deduction of mathematics from logic, with the paradoxes apparently excluded, but it does not improve the philosophical position.

A fundamental objection to the Logicistic thesis is posed by the Intuitionists, who find an essential mathematical kernel in the idea of iteration, which already has to be used in establishing the hierarchy of types and even in listing the propositions in a proof.

The logician Willard Van Orman Quine of the United States in 1937 weakened the type restriction to one of stratification. The resulting system is a form of axiomatic set theory rather than of type theory. Indeed, though the Logicistic thesis did not so develop historically, it could be based as well on a set theory as on a type theory.

Formalism. Hilbert accepted that the parts of classical mathematics employing the completed infinite go beyond intuitive evidence. But, instead of rejecting those parts, he proposed in 1905 and, more specifically, after 1918 roughly to formulate classical mathematics up to a suitable point as a formal axiomatic theory and to prove this theory consistent—*i.e.,* free from contradiction.

Previous to this proposal of Hilbert, proofs of consistency had been given for axiomatic theories by interpreting their undefined terms in another theory so that the axioms become true or briefly by giving a model. Thus, the plane Lobachevskyan geometry was interpreted in Euclidean geometry by Beltrami. The consistency of geometries generally is reduced to that of analysis by René Descartes's analytic geometry of 1637, which employs real numbers as coordinates of points, so that geometrical propositions become propositions in the algebra of real numbers. For proving the consistency of analysis itself, the method of models offers no hope, because no theory is available in which analysis could be interpreted that appears safer than analysis.

Hilbert's proposal called for proving consistency by a new method, based directly on the meaning of consistency as freedom from contradiction. Not only should all the mathematical assumptions of a theory be listed but also all of the principles of logic, both the primitive propositions and the rules for drawing inferences, to be used in deducing theorems from the mathematical postulates. This listing done, it is possible to undertake to show that no two proofs in the theory terminate in theorems one of which is the negation of the other. Hilbert would thus make the proofs of the formal theory the object of an intuitive theory, called proof theory, or metamathematics. The latter should use only intuitively convincing methods, called by him finitary methods. These methods in practice are Intuitionistic methods on the level of number theory. The formal theory provides a sort of stereotyped model of the previously existing informal theory, which is called the interpretation of the formal theory. This interpretation should play no role in the metamathematics, though it motivates the choice of the particular formal theory that is the object of the metamathematics.

To guarantee that all the principles of logic used in deducing theorems have been listed as axioms and rules of inference, the meanings of all the words should be left out of account in deducing theorems, just as in a formal axiomatic theory it should be possible to leave out of account the meanings of the undefined mathematical terms. Because of the irregularities and ambiguities of the existing word languages, it is feasible to do this only by reconstructing the informal theory as a symbolic language, which can be done by combining symbolic logic with the symbolization common in modern mathematics. The resulting formal theory is called a formal system, or logistic system, a statement in it a formula (or by some writers a well-formed formula, to emphasize that it must be formed from the symbols by listed formation rules) and the method the logistic method. The description of a formal system must be given from outside in an informal theory. The method was already available, but Hilbert added the proposal to make the formal system the object of mathematical investigations in this informal theory or metamathematics.

Brouwer–Hilbert controversy

There ensued a controversy in which Brouwer questioned how a metamathematical proof of consistency of classical mathematics would meet the objection that the propositions of classical mathematics beyond a certain elementary level lack a clear intuitive meaning. The upshot was that the Intuitionists would accept Hilbert's program as unobjectionable if the Formalists refrained from construing a consistency proof as in some manner vindicating classical mathematics intuitively. Hilbert drew an analogy with the stratagem in mathematics of adjoining ideal elements to a previously constituted system to achieve some theoretical end (*e.g.,* the irrational numbers were added to the rationals so that all convergent sequences should have limits, the imaginary numbers to the reals so that all algebraic equations should have solutions). He called the statements of classical mathematics with an intuitive meaning real, the others ideal. The statements involving the completed infinite are ideal. So, he said, in classical mathematics the ideal statements are adjoined to the real to give the theory as a whole a simpler logical structure, and it is no more reasonable to require that each separate statement should be interpretable taken by itself than that in theoretical physics each separate statement should be verifiable by experiment. The possibility appeared that classical mathematics reinforced by a proof of consistency might even serve as a tool for the Intuitionistic proof of real statements via shortcuts through the ideal statements.

Hilbert's program was undertaken after 1920 by him and his collaborators, the German Wilhelm Ackermann and the Frenchman Paul Bernays, and by others, including the Hungarian-born U.S. mathematician John von Neumann and the Frenchman Jacques Herbrand. The start was made by investigating a formal system of classical arithmetic (with several variants), obtained by adding Peano's postulates and definitions of sum and product to the first-order functional calculus as developed by Peano and as described in Hilbert and Bernays' *Grundlagen der Mathematik* ("The Foundation of Mathematics") in 1934 and 1939. A proof by Ackermann in 1924–25 of consistency of a subsystem did not lead at once to a proof of consistency of the whole system. Then in 1931 results of Gödel appeared that clarified the situation. These results are presented below from a standpoint that was reached after 1935.

The metamathematical method proved valuable for a variety of purposes in studying formal systems generally, as well as the ones Hilbert considered. Thus, a formal system of Intuitionistic logic and mathematics was described by Heyting in 1930.

Turing's computing machine and formal systems. For certain classes of questions in mathematics, a method is known that can be applied to answer by "yes" or "no" any given question of the class. For example, for any positive integers a and b, the question "Is a a multiple of b?" can be answered by dividing a by b and seeing whether the remainder is 0 or not. As a less obvious example, there is a method for deciding, for any given integers a, b, c, whether $ax + by + c = 0$ has solutions for x, y in integers. Such a method is called an algorithm, or decision procedure; and the problem of finding such a method for a class of questions is called the decision problem for that class. Similarly, for a class of questions that require for their answers not "yes" or "no" but the exhibiting of some object, one speaks of an algorithm, or computation procedure, and a computation problem.

In the logistic method, the description of a formal system must enable one to decide whether any given formula is an axiom and whether any given formula comes by a permitted inference from one or two others; so there is a decision procedure for whether a given finite list of formulas is a proof in the system. But a decision procedure is not thereby provided for whether a given formula

is provable; *i.e.*, for whether there is a proof of it. The problem of finding a decision procedure for the latter is often called the decision problem for the formal system and was recognized in Germany as a problem of logic by Ernst Schröder in 1895, Leopold Löwenheim in 1915 and Hilbert in 1918. In the United States Emil L. Post gave in 1921 a decision procedure for the classical propositional calculus.

The decision problem

A decision procedure for a formal system embracing a given portion of mathematics would make automatic the solution of any problem in that portion statable as a "yes" or "no" question. It is therefore implausible that there should be a decision procedure for a formal system embracing the usual arithmetic, in view of the experience of mathematicians in trying to solve the one particular problem whether Fermat's last theorem is true; and many other similarly refractory problems are known. (The French mathematician Pierre Fermat around 1637 claimed to have proved that the equation $x^n + y^n = z^n$ has no solution in positive integers x, y, z for any positive integer $n > 2$; but no one since then has been able to confirm or disprove this so-called "last theorem" of his.) For there to be a decision procedure would mean, so to speak, that the infinitely many arithmetical problems involve only a finite number of difficulties such that once those are overcome no further ingenuity will be required. But it might have seemed equally farfetched in 1918 to hope that mathematics could get a hold on the problem of finding a decision procedure for arithmetic, which would enable this problem to be solved by showing that there cannot be one. Exactly this latter was done in 1936 by Alonzo Church.

The consideration of the computation problems for number–theoretic functions can lead to the question of what functions can be computed or effectively calculated, in the sense that, for a given argument, the value can be found by use only of preassigned rules, applicable without ingenuity on the part of the computer. Instead of a human, it can be proposed to have a machine to do the computation, according to rules built into its structure.

As this description is somewhat vague, what is needed is an exact mathematical formulation that comes to the same end. Such a formulation was given briefly by Post in 1936 and independently in detail by Turing in 1936–37. What Turing did was to describe a kind of theoretical computing machine, not limited in use by a fixed maximum amount of storage of information nor liable to malfunctioning, as are actual computing machines. A Turing machine operates in the following manner. At discrete moments the machine assumes one of a finite list of states (fixed for a given machine), the first of which is called the passive state, the other active, of which there are k in number, in which k is 1 or a greater integer. A linear tape, potentially infinite to the right, is ruled in squares, each of which is either blank or has printed upon it one of a finite number j (one or more) of symbols (fixed for a given machine), including, say, the tally | ; but only a finite number of squares are printed. The tape passes through the machine so that one of its squares is scanned. If at the given moment the machine is in an active state, then it can alter the condition of the scanned square (by erasing, printing, or both), and/or move the tape so that the scanned square becomes the one next right or left, and/or change to another state between the given moment and the next or any combination of the preceding; this act is determined (for a given Turing machine) by the state of the machine and the condition of the scanned square at the given moment.

Computation of number-theoretic functions

Turing used his machines primarily to compute in succession the digits of decimal expansions of real numbers, but they can be adapted to compute number-theoretic functions. Let an argument x be supplied to the machine by placing a blank followed by $x + 1$ tallies on successive squares of the tape starting with the leftmost square, leaving the tape otherwise blank, and placing the tape in the machine with the rightmost tally scanned and the machine in its first active state. The machine is said to compute a value y for x as argument, if then it eventually reaches the passive state with the $x + 1$ tallies followed immediately to the right by a blank and $y + 1$ tallies, the rightmost being scanned, and the tape being otherwise blank. If, for each natural number x, the machine computes the correct value $y = f(x)$ of a function f, when supplied the argument x, it computes the function f, which is then said to be Turing-computable.

Turing advocated the thesis that any function that can be effectively calculated is Turing-computable. His arguments include showing that various operations that a human computer can perform can be analyzed into successions of the simple operations his machines perform. Church had already in 1936 proposed the thesis that every effectively calculable function is λ-definable in a sense due to himself and S.C. Kleene of the United States or, equivalently, general recursive in a sense due to Gödel, who built on a suggestion of Jacques Herbrand of France; and this thesis was subsequently proved equivalent to Turing's, so the λ-definable functions, the general recursive functions, and the Turing-computable functions are the same.

Church's theorem: logical systems without decision procedures. The pattern of behaviour of a Turing machine is completely described by a $k \times (j + 1)$ table showing the act to be performed for each of the k active states with each of the $j + 1$ conditions of the scanned square. A system for encoding such tables can be established that, to each table, gives a natural number t as code number so that a person who knows the system can from t find the table and thence the behaviour of a machine M_t with this table. If t is the code number of a machine, M_t, that computes a value for x as argument, let $f_t(x)$ be this value; and let $f_t(x)$ be undefined otherwise.

The function f may be defined by letting $f(x) = f_x(x) + 1$ if $f_x(x)$ is defined, and $f(x) = 0$ otherwise. This function f is not computable. For, if machine M_q with code number q computes a value for q as argument, that value is $f_q(q)$, not $f_q(q) + 1$, as it would have to be if M_q computes f. It follows that the property "$f_x(x)$ is defined" is undecidable. For otherwise $f(x)$ could be computed, given x, thus: first decide whether $f_x(x)$ is defined; then, if it is, imitate the behaviour of M_x to compute it and add 1; if it is not, write 0.

Thus, from Turing's thesis one has proved that there is a class of quite elementary arithmetical questions for which there is no decision procedure. This essentially is Church's theorem (1936), which he proved from his thesis using a corresponding class of questions in terms of λ-definability. For each x, the proposition "$f_x(x)$ is defined" is expressible in any formal system S that includes the usual vocabulary of arithmetic, by a formula A_x effectively determinable given x. If the quoted proposition is true, it is demonstrable intuitively by exhibiting the sequence of the acts of machine M_x in computing $f_x(x)$. If S includes ordinary arithmetical reasoning, then that demonstration can be translated into a proof in S of A_x; *i.e.*, A_x is provable in S whenever (the proposition expressed by it is) true. Finally, if it be supposed about S that A_x is provable in S only when true, then there can be no decision procedure for provability in S; for, if there were, by answering whether A_x is provable in S, one would have a decision procedure for "$f_x(x)$ is defined."

Moreover, there is no decision procedure for the important logical system F of first-order functional (or predicate) calculus, as Church in 1936 and Turing in 1936–37 showed. For a proof of A_x in S can be arranged as a deduction by F of A_x from a suitable finite set of arithmetical axioms independent of x, the conjunction of which can be written as one formula C. If the symbolism has been suitably chosen, then A_x is provable in S, if and only if the implication $C \supset A_x$ is provable in F (the symbols $C \supset A_x$ are read "if C, then A_x"). From these beginnings, a considerable theory has grown about "unsolvable decision problems," with further contributions in the United States, Russia, England, and Poland by Post, A.A. Markov (the younger), Turing, Alfred Tarski, Andrzej Mostowski, P.S. Novikov, and others.

The effect of Gödel's theorems on Formalism. A formal system such as S is consistent if for no formula B are both B and its negation ~B provable in S; complete if for each formula B without free variables (otherwise ~B may not express the opposite of B) either B or ~B is provable in S.

It may be further supposed that $\sim A_x$ is provable in S only when true. It is absurd that S be complete (and that $\sim A_x$ be provable in S whenever true). For A_x is provable in S (when and) only when true. So if, for each x, either A_x or $\sim A_x$ were provable, it could be decided whether $f_x(x)$ is defined or not, by searching through all the proofs in S until one is found of A_x or of $\sim A_x$. This is Gödel's theorem of 1931 in a negative form.

Gödel's theorem

To obtain the positive form, the process can be considered, for a given x, of searching through all the proofs in S for one of $\sim A_x$ and writing 0 if one is found. A Turing machine, say machine M_q with code q, executes this process when supplied with x. Now by the choice of the formulas A_x, $\sim A_q$ is true if and only if $f_q(q)$ is undefined, which by the choice of M_q is the case if and only if $\sim A_q$ is unprovable in S. In brief, $\sim A_q$ expresses its unprovability (compare the paradox of the liar). So, were $\sim A_q$ provable, it would be false, contradicting the supposition that $\sim A_x$ is provable only when true. So $\sim A_q$ is unprovable but true, and A_q is also unprovable (since false).

Hilbert's program called for embodying classical mathematics, including arithmetic, analysis, and set theory short of the paradoxes, in a formal system and proving that system consistent by finitary methods. The part of Gödel's theorem that $\sim A_q$ is unprovable but true shows that not even the first step can be carried out fully. Despite the value of the logistic method as a way of defining exactly the presuppositions of a given portion of mathematics, this process of definition restricts that portion to include less of arithmetic than free use of intuitive reasoning would give, at least if what is already incorporated in the formal system is intuitively correct. The part of the theorem that neither $\sim A_q$ nor A_q is provable in S shows the inadequacy of S to decide by proof or disproof, irrespective of the interpretation, all the statements expressible in it. Thus, A_q is said to be formally undecidable in S.

In showing $\sim A_q$ unprovable (and true), the assumption was used (for $x = q$) that $\sim A_x$ is provable in S only when true. This can be replaced by the metamathematical assumption (not ostensibly using the interpretation) that S is consistent. For, if $\sim A_q$ were provable, $f_q(q)$ would be defined, and hence, as noted above, A_q would be provable, which would contradict the consistency.

The assumption that A_x is provable only when true, used in showing A_q unprovable, can be replaced by the assumption that S is ω-consistent, in the sense that for no formula $B(x)$ are $B(0)$, $B(1)$, $B(2)$, ... and $\sim(x)\,B(x)$ all provable in S, when "(x)" expresses "for all x." In verifying this, $B(x)$ will express "at the xth moment M_q has not yet computed $f_q(q)$." Adjoining A_q to S as a new axiom produces a formal system that is consistent but not ω-consistent, if S is consistent. J. Barkley Rosser of the United States in 1936 modified Gödel's formally undecidable formula so that consistency sufficed; in fact, letting M_r carry out the process of searching through all the proofs in S for one of $\sim A_x$ and writing 0 if one is found before one of A_x is found, $\sim A_r$ is unprovable but true, and A_r is unprovable, if S is consistent.

As Gödel observed in proving his theorem, the objects employed in a formal system, say S, are countable; and after correlating natural numbers (called now Gödel numbers) effectively to them, by talking about those numbers instead of the formal objects, the metamathematics of S becomes a part of arithmetic and so expressible in S. Now the implication "If S is consistent, then $\sim A_q$ is true," was obtained above as part of Gödel's theorem. If Consis be the formula of S that expresses "S is consistent" in S via the Gödel numbering, then the quoted implication translated into S becomes $\text{Consis} \supset \sim A_q$. Gödel claimed that, by imitating formally in S the intuitive proof of his theorem, it should be possible to prove this formula in S. (This was done in detail by Hilbert and Bernays in 1939 in their formal system of arithmetic.) But then, if Consis were also provable, $\sim A_q$ would be, which is contrary to Gödel's (first) theorem if S is consistent. Thus Gödel obtained his second theorem: if S is consistent, then the formula Consis, which says in S via the Gödel numbering that S is consistent, is unprovable in S.

Gödel's second theorem

Gödel's second theorem raises an obstacle to the second part of Hilbert's program. One might have supposed that the finitary methods to be used in proving consistency would be part of the methods formalized in the system, which amounts to saying that a finitary proof of consistency of S should be formalizable as a proof in S of Consis. Only such methods appear to have been used in attempts before 1931. Gödel's second theorem shows that not even all the methods incorporated in S, including the nonfinitary ones, would suffice.

Some mathematicians felt this ended Hilbert's plan to make classical mathematics more secure by a consistency proof. Others thought it possible that methods might be found for proving consistency that, though not formalizable within S, could be construed as finitary and thus safer than some of the methods incorporated in S. Thus, Gerhard Gentzen in 1936 gave a proof of the consistency of a system S of classical arithmetic, in which the nonelementary method (not formalizable in S) is an extension of mathematical induction from the natural number sequence to a certain segment of the transfinite ordinal numbers that Cantor had introduced for extending the counting process beyond that provided by natural numbers.

Gödel's (first) theorem provided confirmation for the view advanced earlier by the Intuitionists on philosophical grounds that the possibilities for intuitive mathematical thinking cannot be circumscribed in advance. Taking Brouwer's interpretation of a statement "a y exists such that $P(y)$" as meaning such a y can be found and applying Church's thesis to the case the y depends on x, Kleene in 1943 proposed the further thesis that a statement "for each x, a y exists such that $P(x, y)$" is provable Intuitionistically only if such a y can be given as a computable function of x. Using this thesis, there are theorems of classical arithmetic, by Kleene in 1943, and of classical analysis, by Ernst Specker in 1949, that are unprovable Intuitionistically (not merely in some formal system but absolutely); so classical mathematics reinforced by a consistency proof cannot serve as a tool of Intuitionistic proof, except of a very restricted class of statements. Kleene in 1945 and David Nelson in 1946 together showed that the Intuitionistic formal arithmetic based on Heyting's Intuitionistic formal logic does conform to the thesis. Gödel in 1932–33 (partially anticipated by A.N. Kolmogorov in 1924–25) correlated to each arithmetical formula B a classically equivalent formula B′ that is provable in the Intuitionistic formal arithmetic if B is provable in the classical formal arithmetic. In particular, the correlation proves very simply that the classical formal arithmetic is consistent if the Intuitionistic is. However, the Intutitionistic arithmetic distinguishes differing degrees of indirectness among statements classically equivalent.

Using the classical law of excluded middle, Gödel in 1930 proved a completeness theorem for the first-order functional calculus F. In one form, this says that, if a formal system obtained by adjoining mathematical axioms to F is consistent, it has a model; *i.e.*, there is an interpretation that makes the axioms true. Indeed, the model can be constructed using the natural numbers as the objects. In another form, the theorem says that in such a system every formula is provable that is true under all the interpretations of the undefined terms that make the axioms true. In view of this, when S is based on first-order functional calculus, the unprovability in S of $\sim A_q$ implies that, while $\sim A_q$ is true under the intended interpretation of the undefined terms, there is some other interpretation that makes the axioms true (a nonstandard model) under which $\sim A_q$ is false. This illustrates the theorem of Skolem in 1933 that no list of axioms in the symbolism of the first-order functional calculus can characterize the natural numbers categorically. If higher functional calculus is used, the deductive apparatus will be incomplete. It thus appears that the logistic method is inadequate to characterize the natural numbers categorically. Peano's axioms characterize them categorically but only through an interpretation that cannot be rendered fully through the deductive possibilities. Also one sees the Gödel formal undecidability of A_q as similar to the undecidability of Euclid's parallel postulate from the other postulates of Euclidean geometry.

Skolem's theorem

Degrees of decidability in arithmetic. By Church's thesis, the decidable properties of a natural number are

countable; for each decidable property P is decided by a Turing machine that computes 0 or 1 for x as argument according as x has the property P or not, and each Turing machine's behaviour is described by a code number. Thence the existence of undecidable properties follows simply by Cantor's theorem, by which the sets of natural numbers are uncountable; for each property P corresponds to the set S of the natural numbers that have the property.

Church's theorem, however, goes beyond this: it exhibits an undecidable property of a very simple form (see 5). It can also be found that the properties in arithmetic are arranged in a hierarchy, with decidable properties at the bottom and above each level another level containing new properties defined by more uses of "exist" or "all" (see 5). This is a consequence of the (arithmetical) hierarchy theorem of Kleene in 1943, abstracted in 1940, found independently in another version by Andrzej Mostowski in 1947.

(5) Church's theorem exhibits an undecidable property of a very simple form. In this treatment the undecidable property is "$f_x(x)$ is defined." If "$T(t, x, y)$" abbreviates "t is the code number of a Turing machine M_t that, when supplied with x as argument, completes the computation of a value at moment y (but not earlier)," and if "(Ey)" abbreviates "there exists a y such that" or briefly "exists y" or "for some y," then $[f_x(x)$ is defined$] \equiv (Ey)T(x, x, y)$ (for all x), the symbol $\equiv$ representing "is identical to." But $T(t, x, y)$ and hence $T(x, x, y)$ is a decidable relation. So the undecidable property is of the form $(Ey)R(x, y)$ where $R(x, y)$ is a decidable relation. Using $\sim(Ey)T(x, x, y) \equiv (y)\sim T(x, x, y)$, there is also an undecidable property of the form $(y)R(x, y)$ with $R(x, y)$ decidable [where "(y)" means "for all y"]. Any decidable property $R(x)$ is expressible in both the forms $(Ey)R(x, y)$ and $(y)R(x, y)$, by taking $R(x, y) \equiv R(x)$ & $y = y$, but neither $(Ey)T(x, x, y)$ nor $(y)\sim T(x, x, y)$ is expressible in the form the other has. For, given any decidable relation $R(x, y)$, $(Ey)R(x, y) \equiv (Ey)T(p, x, y)$ where p is the code number of a Turing machine M_p that, when supplied with x as argument, searches for the least y such that $R(x, y)$. So, if $(y)\sim T(x, x, y) \equiv (Ey)R(x, y)$ held for some decidable R, we would have $(y)\sim T(p, p, y) \equiv (Ey)R(p, y) \equiv (Ey)T(p, p, y) \equiv \sim(y)\sim T(p, p, y)$, which is absurd. By similar reasoning, to each of the two forms $(y)(Ez)R(x, y, z)$ and $(Ey)(z)R(x, y, z)$ with decidable R, there is a property, $(y)(Ez)T(x, x, y, z)$ or $(Ey)(z)\sim T(x, x, y, z)$, respectively, of that form that is not expressible in the other form and hence not in any of the simpler forms $R(x)$, $(Ey)R(x, y)$, or $(y)R(x, y)$; and likewise with three, four, · · · alternating exist-operations and all-operations. This is the (arithmetical) hierarchy theorem of Kleene.

Further content is lent to this theory by the notion due to Turing (1939) of a machine that is like a computer except that it has access to an "oracle"; *e.g.*, such a device may be designed to answer questions whether numbers have property P with the help of an oracle that answers questions whether numbers have property Q. This means that, in the process of answering a given question "Does x have property P?" the machine may ask and receive from the oracle answers to questions "Does y have property Q?" If such a device exists, then it may be said that P is decidable from Q. In fact, $(Ey)\,T(x, x, y)$, though undecidable absolutely, is decidable from $(y)\,(Ez)\,T(x, x, y, z)$, but not vice versa; so it may be said that $(y)\,(Ez)\,T(x, x, y, z)$ is "more undecidable" than $(Ey)\,T(x, x, y)$. The hierarchy has been extended upward and applied to properties of sets.

(S.C.K.)

CURRENT DIRECTIONS

Intuitionistic foundations of mathematics. Brouwer's program for the constructive redevelopment of mathematics was continued by Heyting and others. In the 1960s, the U.S. mathematician Errett Bishop carried out a substantial extension of this program. At the same time his treatment differed from Brouwer's in that he avoided the use of notions, such as ips, that classical mathematicians had viewed with suspicion. This and other work showed that Intuitionism can provide a coherent and viable alternative to the other positions for the foundations of mathematics insofar as they apply to practice.

The logical study of formal systems has also revealed the potentialities of Intuitionistic reasoning. First, proceeding from the conditions for asserting compound statements described above, Heyting set up an Intuitionistically justified formal system of logic. This was formulated in such a way that the system of classical logic can be obtained simply by adjoining the laws of excluded middle. The same was done for a system of elementary number theory. In 1932, Gödel found a translation of the systems based on classical logic into the corresponding Intuitionistic ones. This was the first step in providing a theoretical explanation for the possibility of an Intuitionistic account of significant portions of mathematics.

Much work has been done since the midcentury in setting up formal systems for Intuitionistic theories of functions, ips, species, etc., and studying their relations with classical systems. This has also served to clarify the more distinctive Intuitionistic notions and principles. The theory of Turing computable functions has also turned out to be an important tool in the study of these systems. To the classical mathematician they provide a precise substitute for the informal idea of constructive function. To the Intuitionist, however, there is no reason to believe that the mechanically computable functions exhaust the conceivable (potential) totality of constructive functions. (S.Fe.)

Non-Intuitionistic studies on the foundations of mathematics. From around 1960, work on the foundations of mathematics outside the Intuitionistic school has been described as falling mainly into four areas: recursion theory, proof theory, model theory, and set theory. Recursion, proof, model, and set theory

Recursion theory is another name for the theory of computability and decidability. This name came into use from treatments based on the general recursive functions.

Proof theory is another name for metamathematics. One of the newer developments in proof theory is the singling out of portions of analysis formalized so as to use only properties of sets up to a given level in the recursion-theoretic hierarchy.

Model theory deals with the interpretations, or models, that satisfy the axioms of a given formal system. For example, by Gödel's completeness theorem of 1930, if a formal system based on the first-order functional calculus F is consistent, there is a model in which the objects are the natural numbers. But the axioms might be those of an axiomatic set theory, which purport to describe an uncountably infinite totality! The model is then a "nonstandard model" of set theory. The opposition of Gödel's completeness and incompleteness theorems reveals that there is a nonstandard model of arithmetic.

Set theory, as a branch of foundational investigations, received great impetus from the construction in 1963 by Cohen of models that falsify the continuum hypothesis of Cantor. Cohen's and related methods are being used to show that various conjectures in set theory, analysis, and topology cannot be settled on the basis of the usual axioms of Zermelo and Fraenkel, sometimes augmented by new axioms. (S.C.K.)

BIBLIOGRAPHY. Some general references on the foundations of mathematics are HERMANN WEYL, *Philosophy of Mathematics and Natural Science* (1949 and 1963); PAUL BENACERRAF and HILARY PUTNAM (eds.), *Philosophy of Mathematics: Selected Readings* (1964); and RAYMOND L. WILDER, *Introduction to the Foundations of Mathematics*, 2nd ed. (1965). Advanced books, generally without emphasis on Intuitionism, include S.C. KLEENE, *Introduction to Metamathematics* (1952, reprinted 1971) and *Mathematical Logic* (1967); EVERT W. BETH, *The Foundations of Mathematics*, 2nd ed. rev. (1965); JOHN B. ROSSER, *Simplified Independence Proofs* (1969); and ANDRZEJ MOSTOWSKI, *Constructible Sets with Applications* (1969). An excellent source book for all the schools of thought and developments in the foundations of mathematics in the period 1879–1931 is JEAN VAN HEIJENOORT (ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931* (1967). For Intuitionism, see AREND HEYTING, *Intuitionism: An Introduction* (1956); and A.S. TROELSTRA, *Principles of Intuitionism* (1969), two general expositions.

(S.C.K./S.Fe.)

The History of Mathematics

The following article is an account of how important parts of mathematics have developed historically. As a consequence of the exponential growth of science most of this mathematics has developed since the 15th century AD, and it is a historical fact that from the 15th century to the late 20th century new developments in mathematics have been largely concentrated in Europe and North America. For these reasons the bulk of this article is devoted to European developments since 1500.

This does not mean, however, that developments elsewhere have been unimportant. Indeed, to understand the history of mathematics in Europe it is necessary to know its history at least in Mesopotamia and Egypt, in ancient Greece, and in Islāmic civilization from the 9th to the 15th centuries. The way in which these civilizations influenced one another, and the important direct contributions Greece and Islām made to later developments, are discussed in the first parts of this article.

India's contributions to the development of contemporary mathematics were made through the considerable influence of Indian achievements on Islāmic mathematics during its formative years. In order to provide a portrait of the mathematical achievements of one major Asian civilization, the article contains an overview of some of the principal periods and achievements of mathematics in China.

It is important to be aware of the character of the sources for the study of the history of mathematics. The history of Mesopotamian and Egyptian mathematics is based on the many extant original documents written by scribes. Although in the case of Egypt these documents are few, they are all of a type and leave little doubt that Egyptian mathematics was, on the whole, elementary and profoundly practical in its orientation. For Mesopotamian mathematics, on the other hand, there are a large number of clay tablets, which reveal mathematical achievements of a much higher order than those of the Egyptians. The tablets indicate that the Mesopotamians had a great deal of remarkable mathematical knowledge, although they offer no evidence that this knowledge was organized into a deductive system. Future research may reveal more about the early development of mathematics in Mesopotamia or about its influence on Greek mathematics, but it seems likely that this picture of Mesopotamian mathematics will stand.

From the period before Alexander the Great no Greek mathematical documents have been preserved except for fragmentary paraphrases, and even for the subsequent period it is well to remember that the oldest copies of Euclid's *Elements* are in Byzantine manuscripts of the 10th century AD. This stands in complete contrast to the situation described above for Egyptian and Babylonian documents. Although in general outline the present account of Greek mathematics is secure, in such important matters as the origin of the axiomatic method, the pre-Euclidean theory of ratios, and the discovery of the conic sections, historians have given competing accounts based on fragmentary texts, quotations of early writings culled from nonmathematical sources, and a considerable amount of conjecture.

Many important treatises from the early period of Islāmic mathematics have not survived or have survived only in Latin translations, so that there are still many unanswered questions about the relationship between early Islāmic mathematics and the mathematics of Greece and India. In addition, the amount of surviving material from later centuries is so large in comparison with that which has been studied that it is not yet possible to offer any sure judgment of what medieval Islāmic mathematics did not contain, and this means that it is not yet possible to evaluate with any assurance what was original in European mathematics from the 11th to the 15th century.

In modern times the invention of printing has largely solved the problem of obtaining secure texts and has allowed historians of mathematics to concentrate their editorial efforts on the correspondence or the unpublished works of mathematicians. However, the exponential growth of mathematics means that, for the period from the 19th century on, historians are able to treat only the major figures in any detail. In addition there is, as the period gets nearer the present, the problem of perspective. Mathematics, like any other human activity, has its fashions, and the nearer one is to a given period, the more likely these fashions are to look like the wave of the future. For this reason, the present article makes no attempt to assess the most recent developments in the subject.

(J.L.B.)

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

The article is divided into the following sections:

Mathematics in ancient Mesopotamia 604
- The numeral system and arithmetic operations
- Geometric and algebraic problems
- Mathematical astronomy

Mathematics in ancient Egypt 605
- The numeral system and arithmetic operations
- Geometry
- Assessment of Egyptian mathematics

Greek mathematics 607
- The development of pure mathematics 607
  - The pre-Euclidean period
  - The *Elements*
  - The three classical problems
- Geometry in the 3rd century BC 609
  - Archimedes
  - Apollonius
  - Applied geometry
- Later trends in geometry and arithmetic 612
  - Greek trigonometry and mensuration
  - Number theory
  - Survival and influence of Greek mathematics

Mathematics in medieval Islām 613
- Origins
- Mathematics in the 9th century
- Mathematics in the 10th century
- Omar Khayyam
- Islāmic mathematics to the 15th century

European mathematics during the Middle Ages and Renaissance 615
- The transmission of Greek and Arabic learning
- The universities
- The Renaissance

Mathematics in the 17th and 18th centuries 616
- The 17th century 616
  - Institutional background
  - Numerical calculation
  - Analytic geometry
  - The calculus
- The 18th century 620
  - Institutional background
  - Analysis and mechanics
  - History of analysis
- Other developments 622
  - Theory of equations
  - Foundations of geometry

Mathematics in the 19th and 20th centuries 623
- Projective geometry
- Making the calculus rigorous
- Fourier series
- Elliptic functions
- The theory of numbers
- The theory of equations
- Gauss
- Non-Euclidean geometry

Riemann
Riemann's influence
Differential equations
Linear algebra
The foundations of geometry
The foundations of mathematics
Cantor
Mathematical physics
Algebraic topology
Developments in pure mathematics
Mathematical physics and the theory of groups
Mathematics in China and Japan 633B
Chinese mathematics to the 13th century 633B
Outline of the history
The *Nine Chapters*
The commentary of Liu Hui
The *Ten Classics of Mathematics*
Some major developments from the 11th century to the 13th century
The decline of the Sung-Yüan mathematics
Japan in the 17th century 633E
The introduction of Chinese books
The elaboration of Chinese methods
Bibliography 633E

Mathematics in ancient Mesopotamia

Until the 1920s it was commonly supposed that mathematics had its birth among the ancient Greeks. What was known of earlier traditions, such as the Egyptian as represented by the Rhind Papyrus (itself edited for the first time only in 1877), offered at best a meagre precedent. This impression gave way to a very different view as Orientalists succeeded in deciphering and interpreting the technical materials from ancient Mesopotamia.

Because the physical medium of the Mesopotamian scribes was durable (they wrote on clay tablets that were then fired or dried), the surviving evidence of this culture is substantial. Existing specimens of mathematics represent all the major eras—the Sumerian kingdoms of the 3rd millennium BC, the Akkadian and Babylonian regimes (2nd millennium), and the empires of the Assyrians (early 1st millennium), Persians (6th through 4th centuries BC), and Greeks (3rd century BC to 1st century AD). The level of competence was already high as early as the Old Babylonian dynasty, the time of the lawgiver king Hammurabi (*c.* 18th century BC), but after that there were few notable advances. In the application of mathematics to astronomy, however, a major flourishing took place in the Persian and Seleucid (Greek) periods.

The numeral system and arithmetic operations. Unlike the Egyptians, the mathematicians of the Old Babylonian period went far beyond the immediate challenges of their official accounting duties; for example, they introduced a versatile numeral system, which, like the modern system, exploited the notion of place value, and they developed computational methods that took advantage of this means of expressing numbers; they solved linear and quadratic problems by methods much like those now used in school algebra; their success with the study of what are now called Pythagorean number triples was a remarkable feat in number theory. The scribes who made such discoveries must have believed mathematics to be worthy of study in its own right, not just as a practical tool.

The older Sumerian system of numerals followed an additive decimal (base-10) principle similar to that of the Egyptians. But the Old Babylonian system converted this into a place-value system with the base of 60 (sexagesimal). The reasons for the choice of 60 are obscure, but one good mathematical reason might have been the existence of so many divisors (2, 3, 4, and 5, and some multiples) of the base, which would have greatly facilitated the operation of division. For numbers from 1 to 59, the symbols 𒁹 for 1 and 𒌋 for 10 were combined in the simple additive manner (*e.g.*, 𒌋𒌋𒌋𒁹𒁹 represented 32). But, to express larger values, the Babylonians applied the concept of place value: for example, 60 was written as 𒁹, 70 as 𒁹𒌋, 80 as 𒁹𒌋𒌋, and so on. In fact, 𒁹 could represent any power of 60. Only the context determined which power was intended, for the Babylonians employed no symbol for zero and had no mark to separate numbers into integral and fractional parts (as with the modern decimal point). Thus, the three-place numeral 3 7 30 could represent $3\frac{1}{8}$ (*i.e.*, $3 + 7/60 + 30/60^2$), $187\frac{1}{2}$ (*i.e.*, $3 \times 60 + 7 + 30/60$), 11,250 (*i.e.*, $3 \times 60^2 + 7 \times 60 + 30$), or a multiple of these numbers by any power of 60.

The four arithmetic operations were performed in the same way as in the modern decimal system, except that carrying occurred whenever a sum reached 60, rather than 10. Multiplication was facilitated by means of tables; one typical tablet lists the multiples of a number by 1, 2, 3, . . . , 19, 20, 30, 40, and 50. To multiply two numbers several places long, the scribe first broke the problem down into several multiplications, each by a one-place number, and then looked up the value of each product in the appropriate tables. He found the answer to the problem by adding up these intermediate results. These tables were also meant to assist in division, for the specified values that head them were all reciprocals of regular numbers. Multiplication

Regular numbers are those whose prime factors divide the base; the reciprocals of such numbers thus have only a finite number of places (by contrast, the reciprocals of nonregular numbers produce an infinitely repeating numeral). In base 10, for example, only numbers with factors of 2 and 5 (*e.g.*, 8 or 50) are regular, and the reciprocals ($1/8 = 0.125$, $1/50 = 0.02$) have finite expressions; but the reciprocals of other numbers (such as 3 and 7) repeat infinitely ($0.\overline{3}$ and $0.\overline{142857}$, respectively, where the bar indicates the digits that continually repeat). In base 60, only numbers with factors of 2, 3, and 5 are regular; for example, 6 and 54 are regular, so that their reciprocals (10 and 1 6 40) are finite. The entries in the multiplication table for 1 6 40 are thus simultaneously multiples of its reciprocal $1/54$. To divide a number by any regular number, then, one can consult the table of multiples for its reciprocal.

An interesting tablet in the collection of Yale University (see Figure 1) shows a square with its diagonals; on one side is written "30," under one diagonal "42 25 35," and right along the same diagonal "1 24 51 10" (*i.e.*, $1 + 24/60 + 51/60^2 + 10/60^3$). This third number is the correct value of $\sqrt{2}$ to four sexagesimal places (equivalent in the decimal system to 1.414213 . . . , which is too low by only 1 in the seventh place), while the second number is the product of the third number and the first and so gives the length of the diagonal when the side is 30. The scribe thus appears to have known an equivalent of the familiar long method of finding square roots. An additional element of sophistication is that, by choosing 30 (that is, $1/2$) for the side, the scribe obtained as the diagonal the reciprocal of

Yale Babylonian Collection



Figure 1: Babylonian mathematical tablet.

the value of $\sqrt{2}$ (since $\sqrt{2}/2 = 1/\sqrt{2}$), a result useful for purposes of division.

Geometric and algebraic problems. In a Babylonian tablet now in Berlin, the diagonal of a rectangle of sides 40 and 10 is solved as $40 + 10^2/(2 \times 40)$. Here, a very effective approximating rule is being used (that the square root of the sum of $a^2 + b^2$ can be estimated as $a + b^2/2a$), the same rule found frequently in later Greek geometric writings. Both these examples for roots illustrate the Babylonians' arithmetic approach in geometry. They also show that the Babylonians were aware of the relation between the hypotenuse and the two legs of a right triangle (now commonly known as the Pythagorean theorem) more than a thousand years before the Greeks used it.

A type of problem that occurs frequently in the Babylonian tablets seeks the base and height of a rectangle, where their product and sum have specified values. From the given information the scribe worked out the difference, since $(b - h)^2 = (b + h)^2 - 4bh$. In the same way, if the product and difference were given, the sum could be found. And, once both the sum and difference were known, each side could be determined, for $2b = (b + h) + (b - h)$ and $2h = (b + h) - (b - h)$. This procedure is equivalent to a solution of the general quadratic in one unknown. In some places, however, the Babylonian scribes solved quadratic problems in terms of a single unknown, just as would now be done by means of the quadratic formula.

Quadratic problems

Although these Babylonian quadratic procedures have often been described as the earliest appearance of algebra, there are important distinctions. The scribes lacked an algebraic symbolism; although they must certainly have understood that their solution procedures were general, they always presented them in terms of particular cases, rather than as the working through of general formulas and identities. They thus lacked the means for presenting general derivations and proofs of their solution procedures. Their use of sequential procedures rather than formulas, however, is less likely to detract from an evaluation of their effort now that algorithmic methods much like theirs have become commonplace through the development of computers.

As mentioned above, the Babylonian scribes knew that the base (b), height (h), and diagonal (d) of a rectangle satisfy the relation $b^2 + h^2 = d^2$. If one selects values at random for two of the terms, the third will usually be irrational, but it is possible to find cases in which all three terms are integers: for example, 3, 4, 5 and 5, 12, 13. (Such solutions are sometimes called Pythagorean triples.) A tablet in the Columbia University Collection presents a list of 15 such triples (decimal equivalents are shown in parentheses at the right; the gaps in the expressions for h, b, and d separate the place values in the sexagesimal numerals):

h	b	d			
2	1 59	2 49	(120	119	169)
57 36	56 7	1 20 45	(3,456	3,367	4,825)
1 20	1 16 41	1 50 49	(4,800	4,601	6,649)
3 45	3 31 49	5 9 1	(13,500	12,709	18,541)
1 12	1 5	1 37	(72	65	97)
...	...	...	...	...	...
1 30	56	1 46	(90	56	106)

(The entries in the column for h have to be computed from the values for b and d, for they do not appear on the tablet; but they must once have existed on a portion now missing.) The ordering of the lines becomes clear from another column, listing the values of d^2/h^2 (brackets indicate figures that are lost or illegible), which form a continually decreasing sequence: [1 59 0] 15, [1 56 56] 58 14 50 6 15, . . . , [1] 23 13 46 40. Accordingly, the angle formed between the diagonal and the base in this sequence increases continually from just over 45° to just under 60°. Other properties of the sequence suggest that the scribe knew the general procedure for finding all such number triples—that for any integers p and q, $2d/h = p/q + q/p$ and $2b/h = p/q - q/p$. (In the table, the implied values p and q turn out to be regular numbers falling in the standard set of reciprocals, as mentioned above in connection with the multiplication tables.) Scholars are still debating nuances of the construction and the intended use of this table, but no one questions the high level of expertise implied by it.

Mathematical astronomy. The sexagesimal method developed by the Babylonians has a far greater computational potential than what was actually needed for the older problem texts. With the development of mathematical astronomy in the Seleucid period, however, it became indispensable. The astronomers sought to predict future occurrences of important phenomena, such as lunar eclipses and critical points in planetary cycles (conjunctions, oppositions, stationary points, and first and last visibility). They devised a technique for computing these positions (expressed in terms of degrees of latitude and longitude, measured relative to the path of the Sun's apparent annual motion) by successively adding appropriate terms in arithmetic progression. The results were then organized into a table listing positions as far ahead as the scribe chose. (Although the method is purely arithmetic, one can interpret it graphically: the tabulated values form a linear "zigzag" approximation to what is actually a sinusoidal variation.) While observations extending over centuries are required for finding the necessary parameters (*e.g.*, periods, angular range between maximum and minimum values, and the like), only the computational apparatus at their disposal made the astronomers' forecasting effort possible.

Calculation of planetary positions

Within a relatively short time (perhaps only a century or less), the elements of this system came into the hands of the Greeks. Although Hipparchus (2nd century BC) favoured the geometric approach of his Greek predecessors, he took over parameters from the Mesopotamians and adopted their sexagesimal style of computation. Through the Greeks it passed to Arabic scientists in the Middle Ages and thence to Europe, where it remained prominent in mathematical astronomy during the Renaissance and the early modern period. To this day it persists in the use of minutes and seconds to measure time and angles.

Aspects of the Old Babylonian mathematics may have come to the Greeks before this, early in the 5th century BC, the formative period of Greek geometry. There are a number of parallels that scholars have noted: for example, the Greek technique of "application of area" (see below) corresponded to the Babylonian quadratic methods (although in a geometric, not arithmetic, form). Further, the Babylonian rule for estimating square roots was widely used in Greek geometric computations, and there may also have been some shared nuances of technical terminology. Although details of the timing and manner of such a transmission are obscure because of the absence of explicit documentation, it seems that Western mathematics, while stemming largely from the Greeks, is considerably indebted to the older Mesopotamians.

Mathematics in ancient Egypt

The introduction of writing in Egypt in the predynastic period (*c.* 3000 BC) brought with it the formation of a special class of literate professionals, the scribes. By virtue of their writing skills, the scribes took on all the duties of a civil service: record keeping, tax accounting, the management of public works (building projects and the like), even the prosecution of war through overseeing military supplies and payrolls. Young men enrolled in scribal schools to learn the essentials of the trade, which included not only reading and writing but also the basics of mathematics.

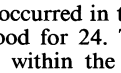
One of the texts popular as a copy exercise in the schools of the New Kingdom (13th century BC) was a satirical letter in which one scribe, Hori, taunts his rival, Amenem-opet, for his incompetence as an adviser and manager. "You are the clever scribe at the head of the troops," Hori chides at one point; "a ramp is to be built, 730 cubits long, 55 cubits wide, with 120 compartments—it is 60 cubits high, 30 cubits in the middle . . . and the generals and the scribes turn to you and say, 'You are a clever scribe, your name is famous. Is there anything you don't know? Answer us, how many bricks are needed?' Let each compartment be 30 cubits by 7 cubits."

This problem, and three others like it in the same letter, cannot be solved without further data. But the point of

the humour is clear, as Hori challenges his rival with these hard, but typical, tasks.

What is known of Egyptian mathematics tallies well with the tests posed by the scribe Hori. The information comes primarily from two long papyrus documents that once served as textbooks within scribal schools. The Rhind Papyrus (in the British Museum) is a copy made in the 17th century BC of a text two centuries older still. In it is found a long table of fractional parts to help with division, followed by the solutions of 84 specific problems in arithmetic and geometry. The Golenishchev Papyrus (in the Moscow Museum of Fine Arts), dating from the 19th century BC, presents 25 problems of a similar type. These problems reflect well the functions the scribes would perform, for they deal with how to distribute beer and bread as wages, for example, or how to measure the areas of fields or the volumes of pyramids and other solids.

Mathematical texts

The numeral system and arithmetic operations. The Egyptians, like the Romans after them, expressed numbers according to a decimal scheme, using separate symbols for 1, 10, 100, 1,000, and so on; each symbol appeared in the expression for a number as many times as the value it represented occurred in the number itself. For example,  stood for 24. This rather cumbersome notation was used within the hieroglyphic writing found in stone inscriptions and other formal texts, but in the papyrus documents the scribes employed a more convenient abbreviated script, called hieratic writing, where, for example, 24 was written  .

In such a system, addition and subtraction amount to counting how many symbols of each kind there are in the numerical expressions and then rewriting with the resulting number of symbols. The texts that survive do not reveal what, if any, special procedures the scribes used to assist in this. But for multiplication they introduced a method of successive doubling. For example, to multiply 28 by 11, one constructs a table of multiples of 28 like the following:

1	28
2	56
4	112
8	224
16	448
...	...

The several entries in the first column that together sum to 11 (*i.e.*, 8, 2, and 1) are checked off. The product is then found by adding up the multiples corresponding to these entries; thus, $224 + 56 + 28 = 308$, the desired product.

To divide 308 by 28, the Egyptians applied the same procedure in reverse. Using the same table as in the multiplication problem, one can see that 8 produces the largest multiple of 28 that is less then 308 (for the entry at 16 is already 448), and 8 is checked off. The process is then repeated, this time for the remainder (84) obtained by subtracting the entry at 8 (224) from the original number (308). This, however, is already smaller than the entry at 4, which consequently is ignored, but it is greater than the entry at 2 (56), which is then checked off. The process is repeated again for the remainder obtained by subtracting 56 from the previous remainder of 84, or 28, which also happens to exactly equal the entry at 1 and which is then checked off. The entries that have been checked off are added up, yielding the quotient: $8 + 2 + 1 = 11$. (In most cases, of course, there is a remainder that is less than the divisor.)

For larger numbers, this procedure can be improved by considering multiples of one of the factors by 10, 20, ... or even by higher orders of magnitude (100, 1,000, ...), as necessary (in the Egyptian decimal notation, these multiples are easy to work out). Thus, one can find the product of 28 by 27 by setting out the multiples of 28 by 1, 2, 4, 8, 10, and 20. Since the entries 1, 2, 4, and 20 add up to 27, one has only to add up the corresponding multiples to find the answer.

Computations involving fractions

Computations involving fractions are carried out under the restriction to unit parts (that is, fractions that in modern notation are written with 1 as the numerator). To express the result of dividing 4 by 7, for instance, which in modern notation is simply $4/7$, the scribe wrote $1/2 + 1/14$. The procedure for finding quotients in this form merely extends the usual method for the division of integers, where one now inspects the entries for $2/3$, $1/3$, $1/6$, etc., and $1/2$, $1/4$, $1/8$, etc., until the corresponding multiples of the divisor sum to the dividend. (The scribes included $2/3$, one may observe, even though it is not a unit fraction.) In practice, the procedure can sometimes become quite complicated (for example, the value for $2/29$ is given in the Rhind Papyrus as $1/24 + 1/58 + 1/174 + 1/232$) and can be worked out in different ways (for example, the same $2/29$ might be found as $1/15 + 1/435$ or as $1/16 + 1/232 + 1/464$, etc.). A considerable portion of the papyrus texts is devoted to tables to facilitate the finding of such unit-fraction values.

These elementary operations are all that one needs for solving the arithmetic problems in the papyri. For example, "to divide 6 loaves among 10 men" (Rhind Papyrus, problem 3), one merely divides to get the answer $1/2 + 1/10$. In one group of problems an interesting trick is used: "A quantity (*aha*) and its 7th together make 19—what is it?" (Rhind Papyrus, problem 24). Here one first supposes the quantity to be 7: since $1\,1/7$ of it becomes 8, not 19, one takes $1\,1/8$ (that is, $2 + 1/4 + 1/8$), and its multiple by 7 ($16 + 1/2 + 1/8$) becomes the required answer. This type of procedure (sometimes called the method of "false position" or "false assumption") is familiar in many other arithmetic traditions (*e.g.*, the Chinese, Hindu, Muslim, and Renaissance European), although they appear to have no direct link to the Egyptian.

Geometry. The geometric problems in the papyri seek measurements of figures, like rectangles and triangles of given base and height, by means of suitable arithmetic operations. In a more complicated problem, a rectangle is sought whose area is 12 and whose height is $1/2 + 1/4$ times its base (Golenishchev Papyrus, problem 6). To solve the problem, the ratio is inverted and multiplied by the area, yielding 16; the square root of the result (4) is the base of the rectangle, and $1/2 + 1/4$ times 4, or 3, is the height. The entire process is analogous to the process of solving the algebraic equation for the problem ($x \times 3/4\, x = 12$), though without the use of a letter for the unknown. An interesting procedure is used to find the area of the circle (Rhind Papyrus, problem 50): $1/9$ of the diameter is discarded and the result is squared. For example, if the diameter is 9, the area is set equal to 64. The scribe recognized that the area of a circle is proportional to the square of the diameter and assumed for the constant of proportionality (that is, $\pi/4$) the value $64/81$. This is a rather good estimate, being about $6/10$ of 1 percent too large. (It is not as close, however, as the now common estimate of $3\,1/7$, first proposed by Archimedes, which is only about $4/100$ of 1 percent too large.) But there is nothing in the papyrus indicating that the scribes were aware that this rule was only approximate rather than exact.

A remarkable result is the rule for the volume of the truncated pyramid (Golenishchev Papyrus, problem 14). The scribe assumes the height to be 6, the base to be a square of side 4, and the top a square of side 2. He multiplies one-third the height times 28, finding the volume to be 56; here, 28 is computed from $2 \times 2 + 2 \times 4 + 4 \times 4$. Since this is correct it can be assumed that the scribe also knew the general rule: $A = (h/3)(a^2 + ab + b^2)$. How the scribes actually derived the rule is a matter for debate, but it is reasonable to suppose that they were aware of related rules, such as that for the volume of a pyramid: one-third the height times the area of the base.

Measurement of distances

The Egyptians employed the equivalent of similar triangles to measure distances. For instance, the *seked* of a pyramid is stated as the number of palms in the horizontal corresponding to a rise of one cubit (seven palms). Thus, if the *seked* is $5\,1/4$ and the base is 140 cubits, the height becomes $93\,1/3$ (Rhind Papyrus, problem 57). The Greek sage Thales of Miletus (6th century BC) is said to have measured the height of pyramids by means of their shadows (the report derives from Hieronymus, a disciple of Aristotle in the 4th century BC). In light of the *seked* computations, however, this report must indicate an aspect of Egyptian surveying that extended back at least 1,000 years before the time of Thales.

Assessment of Egyptian mathematics. The papyri thus bear witness to a mathematical tradition closely tied to the practical accounting and surveying activities of the scribes. Occasionally, the scribes loosened up a bit: one problem (Rhind Papyrus, problem 79), for example, seeks the total from seven houses, seven cats per house, seven mice per cat, seven ears of wheat per mouse, and seven *hekat* of grain per ear (result: 19,607). The underlying scenario can only be guessed at, but it is probably of the playful "As I was going to St. Ives" type—certainly, the scribe's interest in progressions (for which he appears to have a rule) goes beyond practical considerations. Other than this, however, Egyptian mathematics falls firmly within the range of practice.

Even allowing for the scantiness of the documentation that survives, the Egyptian achievement in mathematics must be viewed as modest. Its most striking features are competence and continuity. The scribes managed to work out the basic arithmetic and geometry necessary for their official duties as civil managers, and their methods persisted with little evident change for at least a millennium, perhaps two. Indeed, when Egypt came under Greek domination in the Hellenistic period (from the 3rd century BC onward), the older school methods continued. Quite remarkably, for example, the older unit-fraction methods are still prominent in Egyptian school papyri, written in the demotic (Egyptian) and Greek languages as late as the 7th century AD.

Influence on Greek mathematics

To the extent that Egyptian mathematics left a legacy at all, it was through its impact on the emerging Greek mathematical tradition between the 6th and 4th centuries BC. Because the documentation from this period is limited, the manner and significance of the influence can only be conjectured. But the report about Thales is only one of several such accounts of Greek intellectuals learning from Egyptians; Herodotus and Plato describe with approval Egyptian practices in the teaching and application of mathematics. This literary evidence has historical support, since the Greeks maintained continuous trade and military operations in Egypt from the 7th century BC onward. It is thus plausible that basic precedents for the Greeks' earliest mathematical efforts—how they dealt with fractional parts or measured areas and volumes, or their use of ratios in connection with similar figures—came from the learning of the ancient Egyptian scribes.

Greek mathematics

THE DEVELOPMENT OF PURE MATHEMATICS

The pre-Euclidean period. The Greeks divided the field of mathematics into arithmetic (the study of "multitude," or discrete quantity) and geometry (that of "magnitude," or continuous quantity) and considered both to have originated in practical activities. Proclus, in his *Commentary on Euclid,* for example, observes that geometry, literally, "measurement of land," first arose in surveying practices among the ancient Egyptians, for the flooding of the Nile compelled them each year to redefine the boundaries of properties. Similarly, Proclus continues, arithmetic started with the commerce and trade of Phoenician merchants. Although Proclus wrote quite late in the ancient period (in the 5th century AD), his account drew upon views proposed much earlier, by Herodotus (mid-5th century BC), for example, and by Eudemus, a disciple of Aristotle (late 4th century BC).

However plausible, this view is difficult to check, for there is only meagre evidence of practical mathematics from the early Greek period (roughly, the 8th through the 4th century BC). Inscriptions on stone, for example, reveal use of a numeral system the same in principle as the familiar Roman numerals. Herodotus seems to have known of the abacus as an aid for computation by both Greeks and Egyptians, and about a dozen stone specimens of Greek abaci survive from the 5th and 4th centuries BC. In the surveying of new cities in the Greek colonies of the 6th and 5th centuries, there was regular use of a standard length of 70 *plethra* (one *plethron* equals 100 feet) as the diameter of a square of side 50 *plethra;* in fact, the actual diameter of the square is $50\sqrt{2}$, so this was equivalent to using $^7/_5$ (or 1.4) as an estimate for $\sqrt{2}$, which is now known to equal 1.414 .... In the 6th century BC the engineer Eupalinus of Megara directed an aqueduct through a mountain on the island of Samos, and historians still debate how he did it. In a further indication of the practical aspects of early Greek mathematics, Plato describes in his *Laws* how the Egyptians drilled their children in practical problems in arithmetic and geometry; he clearly considered this a model for the Greeks to imitate.

Such hints about the nature of early Greek practical mathematics are confirmed in later sources, for example, in the arithmetic problems in papyrus texts from Ptolemaic Egypt (from the 3rd century BC onward) and the geometric manuals by Hero of Alexandria (1st century AD). In its basic manner this Greek tradition was much like the earlier traditions in Egypt and Mesopotamia. Indeed, it is likely that the Greeks borrowed from such older sources to some extent.

What was distinctive of the Greeks' contribution to mathematics—and what in effect made them the creators of "mathematics," as the term is usually understood—was its development as a theoretical discipline. This means two things: mathematical statements are general, and they are confirmed by proof. For example, the Mesopotamians had procedures for finding whole numbers *a, b,* and *c* for which $a^2 + b^2 = c^2$ (*e.g.,* 3, 4, 5; 5, 12, 13; or 119, 120, 169). From the Greeks came a proof of a general rule for finding all such sets of numbers (now called Pythagorean triples): if one takes any whole numbers p and q, one being odd, the other even, then $a = (p^2 - q^2)/2$, $b = 2pq$, and $c = (p^2 + q^2)/2$. As Euclid proves in Book X of the *Elements,* numbers of this form (and only of this form) satisfy the relation for Pythagorean triples. Further, the Mesopotamians appear to have understood that sets of such numbers *a, b,* and *c* form the sides of right triangles, but the Greeks proved this result (Euclid, in fact, proves it twice, in *Elements,* Book I, proposition 47, and in a more general form in *Elements,* Book VI, proposition 31), and these proofs occur in the context of a systematic presentation of the properties of plane geometric figures.

Pythagorean triples

The *Elements,* composed by Euclid of Alexandria, around 300 BC, was the pivotal contribution to theoretical geometry, but the transition from practical to theoretical mathematics had occurred much earlier, sometime in the 5th century BC. Initiated by men like Pythagoras of Samos (late 6th century) and Hippocrates of Chios (late 5th century), the theoretical form of geometry was advanced by others, most prominently the Pythagorean Archytas of Tarentum, Theaetetus of Athens, and Eudoxus of Cnidus (4th century). Because the actual writings of these men do not survive, knowledge about their work depends on remarks made by later writers. While even this limited evidence reveals how heavily Euclid depended on them, it does not set out clearly the motives behind their studies.

It is thus a matter of debate how and why this theoretical transition took place. A frequently cited factor is the discovery of the irrational. The early Pythagoreans held that "all things are number." This might be taken to mean that any geometric measure can be associated with some number (that is, some whole number or fraction; in modern terminology, rational number), for in Greek usage the term for number, *arithmos,* refers exclusively to whole numbers or, in some contexts, to ordinary fractions. This assumption is common enough in practice, as when the length of a given line is said to be so many feet plus a fractional part. However, it breaks down for the lines that form the side and diameter of the square. (For example, if it is supposed that the ratio between the side and diameter may be expressed as the ratio of two whole numbers, it can be shown that both of these numbers must be simultaneously even and odd; since this is impossible, no such whole numbers can exist.) Geometrically, this means that there is no length that could serve as a unit of measure of both the side and diameter; that is, the side and diameter cannot each equal the same length multiplied by (different) whole numbers. Accordingly, the Greeks called such pairs of lengths "incommensurable." (In modern terminology, unlike that of the Greeks, the term "number" is applied to such quantities as $\sqrt{2}$, but they are called irrational.)

This result was already well known at the time of Plato and may well have been discovered within the school of Pythagoras in the 5th century BC, as some late authorities like Pappus of Alexandria (4th century AD) maintain. In any case, by 400 BC it was known that lines corresponding to $\sqrt{3}$, $\sqrt{5}$, and other square roots are incommensurable with a fixed unit length. The more general result, the geometric equivalent of the theorem that $\sqrt{p}$ is irrational whenever p is not a rational square number, is associated with Plato's friend Theaetetus. Both Theaetetus and Eudoxus contributed to the further study of irrationals, and their followers collected the results into a substantial theory, as represented by the 115 propositions of Book X of the *Elements*.

The theory of irrationals

The discovery of irrationals must have affected the very nature of early mathematical research, for it made clear that arithmetic was insufficient for the purposes of geometry, despite the assumptions made in practical work. Further, once such seemingly obvious assumptions as the commensurability of all lines turned out to be, in fact, false, then in principle all mathematical assumptions were rendered suspect. At the least, it became necessary to justify carefully all claims made about mathematics. Even more basically, it became necessary to establish what a reasoning has to be like to qualify as a proof. Apparently, Hippocrates of Chios, in the 5th century BC, and others soon after him had already begun the work of organizing geometric results into a systematic form in textbooks called "elements" (meaning "fundamental results" of geometry). These were to serve as sources for Euclid in his comprehensive textbook a century later.

The early mathematicians were not an isolated group but part of a larger, intensely competitive intellectual environment of pre-Socratic thinkers in Ionia and Italy, as well as Sophists at Athens. By insisting that only permanent things could have real existence, the philosopher Parmenides (5th century BC) called into question the most basic claims about knowledge itself. In contrast, Heracleitus (*c.* 500 BC) maintained that all permanence is an illusion, for the things that are perceived arise through a subtle balance of opposing tensions. What is meant by "knowledge" and "proof" thus came into debate.

Mathematical issues were often drawn into these debates. For some, like the Pythagoreans (and, later, Plato), the certainty of mathematics was held as a model for reasoning in other areas, like politics and ethics. But for others, mathematics seemed prone to contradiction. Zeno of Elea (5th century BC) posed paradoxes about quantity and motion. In one such paradox, it is assumed that a line can be bisected again and again without limit; if the division ultimately results in a set of points of zero length, then even infinitely many of them sum up only to zero, but, if it results in tiny line segments, then their sum will be infinite. In effect, the length of the given line must be both zero and infinite. In the 5th century BC a solution of such paradoxes was attempted by Democritus and the "atomists," philosophers who held that all material bodies are ultimately made up of invisibly small "atoms" (the Greek word *atomon* means "indivisible"). But in geometry such a view came into conflict with the existence of incommensurable lines, since the atoms would become the measuring units of all lines, even incommensurable ones. Protagoras and Democritus puzzled over whether the tangent to a circle meets it at a point or a line. The Sophists Antiphon and Bryson (both 5th century BC) considered how to compare the circle to polygons inscribed in it.

Influence of the pre-Socratics

The pre-Socratics thus revealed difficulties in specific assumptions about the infinitely many and the infinitely small, about the relation of geometry to physical reality, as well as in more general conceptions like "existence" and "proof." Philosophical questions such as these need not have affected the technical researches of mathematicians, but they did make them aware of difficulties that could bear on fundamental matters and so made them the more cautious in defining their subject matter.

Any such review of the possible effects of factors such as these is purely conjectural, since the sources are fragmentary and never make explicit how the mathematicians responded to the issues that were raised. But it is the particular concern over fundamental assumptions and proofs that distinguishes Greek mathematics from the earlier traditions. Plausible factors behind this concern can be identified in the special circumstances of the early Greek tradition—its technical discoveries and its cultural environment—even if it is not possible to describe in detail how these changes took place.

The Elements. The principal source for reconstructing pre-Euclidean mathematics is Euclid's *Elements,* for the major part of its contents can be traced back to research from the 4th century BC and in some cases even earlier. The first four books present constructions and proofs of plane geometric figures: Book I deals with the congruence of triangles, the properties of parallel lines, and the area relations of triangles and parallelograms; Book II establishes equalities relating to squares, rectangles, and triangles; Book III covers basic properties of circles; and Book IV sets out constructions of polygons in circles. Much of the content of Books I–III was already familiar to Hippocrates, and the material of Book IV can be associated with the Pythagoreans, so that this portion of the *Elements* has roots in 5th-century research. It is known, however, that questions about parallels were debated in Aristotle's school (around 350 BC), and so it may be assumed that efforts to prove results—such as the theorem stating that, for any given line and given point, there always exists a unique line through that point and parallel to the line—were tried and failed. Thus, the decision to found the theory of parallels on a postulate, as in Book I of the *Elements,* must have been a relatively recent development in Euclid's time. (The postulate would later become the subject of much study, and in modern times it led to the discovery of the so-called non-Euclidean geometries.)

Book V sets out a general theory of proportion, that is, a theory that does not require any restriction to commensurable magnitudes. This general theory derives from Eudoxus. On the basis of the theory, Book VI describes the properties of similar plane rectilinear figures and so generalizes the congruence theory of Book I. It appears that the technique of similar figures was already known in the 5th century BC, even though a fully valid justification could not have been given before Eudoxus worked out his theory of proportion.

The theory of numbers

Books VII–IX deal with what the Greeks called "arithmetic," the theory of whole numbers. It includes the properties of numerical proportions, greatest common divisors, least common multiples, and relative primes (Book VII); propositions on numerical progressions and square and cube numbers (Book VIII); and special results, like unique factorization into primes, the existence of an unlimited number of primes, and the formation of "perfect" numbers, that is, those numbers that equal the sum of their proper divisors (Book IX). In some form, Book VII stems from Theaetetus and Book VIII from Archytas.

Book X presents a theory of irrational lines and derives from the work of Theaetetus and Eudoxus. The remaining books treat the geometry of solids. Book XI sets out results on solid figures analogous to those for planes in Books I and VI; Book XII proves theorems on the ratios of circles, the ratios of spheres, and the volumes of pyramids and cones; Book XIII shows how to inscribe the five regular solids in a given sphere (compare the constructions of plane figures in Book IV). The measurement of curved figures in Book XII is inferred from that of rectilinear figures; for a particular curved figure a sequence of rectilinear figures is considered in which succeeding figures in the sequence become continually closer to the curved figure; the particular method used by Euclid derives from Eudoxus. The solid constructions in Book XIII derive from Theaetetus.

In sum the *Elements* gathered together the whole field of elementary geometry and arithmetic that had developed in the two centuries before Euclid. Doubtless, Euclid must be credited with particular aspects of this work, certainly with its editing as a comprehensive whole. But it is not possible to identify for certain even a single one of its results as having been his discovery. Other more advanced fields, though not touched on in the *Elements,* were already being vigorously studied in Euclid's time, in some

cases by Euclid himself. For these fields his textbook, true to its name, provides the appropriate "elementary" introduction.

One such field is the study of geometric constructions. Euclid, like geometers in the generation before him, divided mathematical propositions into two kinds: "theorems" and "problems." A theorem makes the claim that all terms of a certain description have a specified property; a problem seeks the construction of a term that is to have a specified property. In the *Elements* all the problems are constructible on the basis of three stated postulates: that a line can be constructed by joining two given points, that a given line segment can be extended in a line indefinitely, and that a circle can be constructed with a given point as centre and a given line segment as radius. These postulates in effect restricted the constructions to the use of the so-called Euclidean tools—*i.e.,* a compass and a straightedge or unmarked ruler.

The Euclidean tools

The three classical problems. Although Euclid solves more than 100 construction problems in the *Elements,* many more were posed whose solutions required more than just compass and straightedge. Three such problems stimulated so much interest among later geometers that they have come to be known as the "classical problems": doubling the cube, *i.e.,* constructing a cube whose volume is twice that of a given cube; trisecting the angle; and squaring the circle. Even in the pre-Euclidean period the effort to construct a square equal in area to a given circle had begun. Some related results came from Hippocrates, others were reported from Antiphon and Bryson, and Euclid's theorem on the circle in *Elements,* Book XII, proposition 2, which states that circles are in the ratio of the squares of their diameters, was important for this search. But the first actual constructions (not, it must be noted, by means of the Euclidean tools, for this is impossible) came only in the 3rd century BC. The early history of angle trisection is obscure. Presumably it was attempted in the pre-Euclidean period, although solutions are known only from the 3rd century or later.

There are several successful efforts at doubling the cube, however, that date from the pre-Euclidean period. Hippocrates showed that the problem could be reduced to that of finding two mean proportionals: if for a given line a it is necessary to find x such that $x^3 = 2a^3$, lines x and y may be sought such that $a:x = x:y = y:2a;$ for then $a^3/x^3 = (a/x)^3 = (a/x)(x/y)(y/2a) = a/2a = 1/2$. (Note that the same argument holds for any multiplier, not just the number 2.) Thus, the cube can be doubled if it is possible to find the two mean proportionals x and y between the two given lines a and $2a$. Constructions of the problem of the two means were proposed by Archytas, Eudoxus, and Menaechmus in the 4th century BC. Menaechmus, for example, constructed three curves corresponding to these same proportions: $x^2 = ay$, $y^2 = 2ax$, and $xy = 2a^2$; the intersection of any two of them then produces the line x that solves the problem. Menaechmus' curves are conic sections: the first two are parabolas, the third a hyperbola. Thus, it is often claimed that Menaechmus originated the study of the conic sections. Indeed, Proclus and his older authority, Geminus (mid-1st century AD), appear to have held this view. The evidence does not indicate how Menaechmus actually conceived of the curves, however, so it is possible that the formal study of the conic sections as such did not begin until later, near the time of Euclid. Both Euclid and an older contemporary, Aristaeus, composed treatments (now lost) of the theory of conic sections.

In seeking the solutions of problems, geometers developed a special technique, which they called "analysis." They assumed the problem to have been solved and then, by investigating the properties of this solution, worked back to find an equivalent problem that could be solved on the basis of the givens. To obtain the formally correct solution of the original problem, then, geometers reversed the procedure: first the data were used to solve the equivalent problem derived in the analysis, and from the solution obtained the original problem was then solved. In contrast to analysis, this reversed procedure is called "synthesis."

Analysis and synthesis

Menaechmus' cube duplication is an example of analysis: he assumed the mean proportionals x and y and then discovered them to be equivalent to the result of intersecting the three curves whose construction he could take as known. (The synthesis consists of introducing the curves, finding their intersection, and showing that this solves the problem.) It is clear that geometers of the 4th century BC were well acquainted with this method, but Euclid provides only syntheses, never analyses, of the problems solved in the *Elements.* Certainly in the cases of the more complicated constructions, however, there can be little doubt that some form of analysis preceded the syntheses presented in the *Elements.*

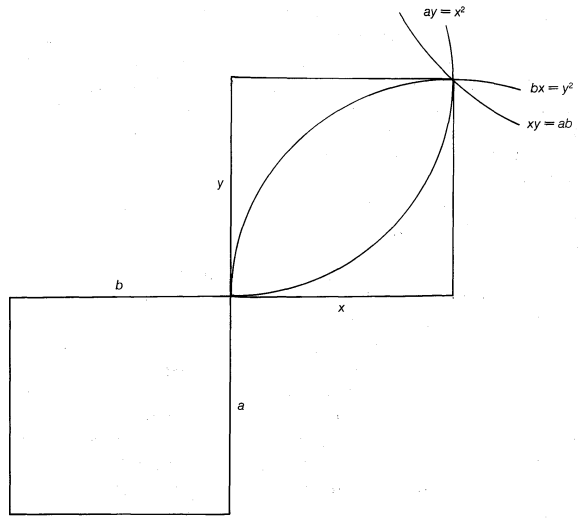


Figure 2: Cube duplication.

GEOMETRY IN THE 3RD CENTURY BC

The *Elements* was one of several major efforts by Euclid and others to consolidate the advances made over the 4th century BC. On the basis of these advances, Greek geometry entered its golden age in the 3rd century. This was a period rich with geometric discoveries, particularly in the solution of problems by analysis and other methods, and was dominated by the achievements of two figures: Archimedes of Syracuse (b. early 3rd century–d. 212 BC) and Apollonius of Perga (mid-3rd to early 2nd century BC).

Archimedes. Archimedes was most noted for his use of the Eudoxean method of exhaustion in the measurement of curved surfaces and volumes and for his applications of geometry to mechanics. To him is owed the first appearance and proof of the approximation $3\frac{1}{7}$ for the ratio of the circumference to the diameter of the circle (what is now designated π). Characteristically, Archimedes went beyond familiar notions, such as that of simple approximation, to more subtle insights, like the notion of bounds. For example, he showed that the perimeters of regular polygons circumscribed about the circle eventually become less than $3\frac{1}{7}$ the diameter as the number of their sides increases (Archimedes established the result for 96-sided polygons); similarly, the perimeters of the inscribed polygons eventually become greater than $3\frac{10}{71}$. Thus, these two values are upper and lower bounds, respectively, of π.

Archimedes' result bears on the problem of circle quadrature in the light of another theorem he proved: that the area of a circle equals the area of a triangle whose height equals the radius of the circle and whose base equals its circumference. He established analogous results for the sphere showing that the volume of a sphere is equal to that of a cone whose height equals the radius of the sphere and whose base equals its surface area; the surface area of the sphere he found to be four times the area of its greatest circle. Equivalently, the volume of a sphere is shown to be two-thirds that of the cylinder which just contains it (that is, having height and diameter equal to the diameter of the sphere), while its surface is also equal to two-thirds that of the same cylinder (that is, if the circles that enclose the cylinder at top and bottom are included). The Greek historian Plutarch (early 2nd century AD) relates that Archimedes requested the figure for this theorem to

Volume of the sphere

be engraved on his tombstone, which is confirmed by the Roman writer Cicero (1st century BC), who actually located the tomb when he was quaestor of Sicily in 75 BC.

Apollonius. The work of Apollonius of Perga extended the field of geometric constructions far beyond the range in the *Elements.* For example, Euclid in Book III shows how to draw a circle so as to pass through three given points or to be tangent to three given lines; Apollonius (in a work called *Tangencies,* which no longer survives) found the circle tangent to three given circles, or tangent to any combination of three points, lines, and circles. (The three-circle tangency construction, one of the most extensively studied geometric problems, has attracted more than 100 different solutions in the modern period.)

Apollonius is best known for his *Conics,* a treatise in eight books (Books I–IV survive in Greek, V–VII in a medieval Arabic translation; Book VIII is lost). The conic sections are the curves formed when a plane intersects the surface of a cone (or double cone); it is assumed that the surface of the cone is generated by the rotation of a line through a fixed point around the circumference of a circle which is in a plane not containing that point. (The fixed point is the vertex of the cone, and the rotated line its generator.) There are three basic types: if the cutting plane is parallel to one of the positions of the generator, it produces the "parabola"; if it meets the cone only on one side of the vertex, it produces an "ellipse" (of which the circle is a special case), but, if it meets both parts of the cone, a "hyperbola." Apollonius sets out in detail the properties of these curves. He shows, for example, that for given line segments a and b the parabola corresponds to the relation (in modern notation) $y^2 = ax$, the ellipse to $y^2 = ax - ax^2/b$, and the hyperbola to $y^2 = ax + ax^2/b$.

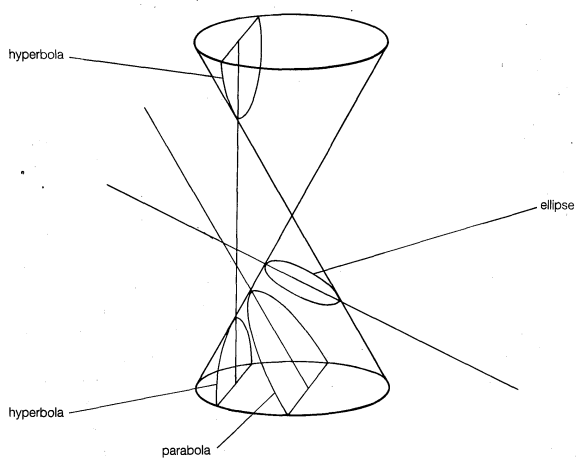


Figure 3: Conic sections.

Apollonius' treatise on conics in part consolidated more than a century of work before him and in part presented new findings of his own. As mentioned above, Euclid had already issued a textbook on the conics, while even earlier Menaechmus had played a role in their study. The names that Apollonius chose for the curves (the terms may be original with him) indicate yet an earlier connection. In the pre-Euclidean geometry *parabolē* referred to a specific operation, the "application" of a given area to a given line, in which the line x is sought such that $ax = b^2$ (where a and b are given lines); alternatively, x may be sought such that $x(a + x) = b^2$, or $x(a - x) = b^2$, and in these cases the application is said to be in "excess" (*hyperbolē*) or "defect" (*elleipsis*) by the amount of a square figure (namely, x^2). These constructions, which amount to a geometric solution of the general quadratic, appear in Books I, II, and VI of the *Elements* and can be associated in some form with the 5th-century Pythagoreans.

The *Conics*

Apollonius presented a comprehensive survey of the properties of these curves. A sample of the topics he covered includes the following: the relations satisfied by the diameters and tangents of conics (Book I); how hyperbolas are related to their "asymptotes," the lines they approach without ever meeting (Book II); how to draw tangents to given conics (Book II); relations of chords intersecting in conics (Book III); the determination of the number of ways in which conics may intersect (Book IV); how to draw "normal" lines to conics, that is, lines meeting them at right angles (Book V); and the congruence and similarity of conics (Book VI).

By Apollonius' explicit statement, his results are of principal use as methods for the solution of geometric problems via conics. While he actually solved only a limited set of problems, the solutions of many others can be inferred from his theorems. For instance, the theorems of Book III permit the determination of conics that pass through given points or are tangent to given lines. In another work (now lost) Apollonius solved the problem of cube duplication by conics (a solution related in some way to that given by Menaechmus); further, a solution of the problem of angle trisection given by Pappus may have come from Apollonius or have been influenced by his work.

With the advance of the field of geometric problems by Euclid, Apollonius, and their followers, it became appropriate to introduce a classifying scheme: those problems solvable by means of conics were called solid, while those solvable by means of circles and lines only (as assumed in Euclid's *Elements*) were called planar. Thus, one can double the square by planar means (as in *Elements,* Book II, proposition 14), but one cannot double the cube in such a way, although a solid construction is possible (as given above). Similarly, the bisection of any angle is a planar construction (as shown in *Elements,* Book I, proposition 9), but the general trisection of the angle is of the solid type. It is not known when the classification was first introduced or when the planar methods were assigned canonical status relative to the others, but it seems plausible to date this near Apollonius' time. Indeed, much of his work—books like the *Tangencies,* the *Vergings* (or *Inclinations*), and the *Plane Loci,* now lost but amply described by Pappus—turns on the project of setting out the domain of planar constructions in relation to solutions by other means. On the basis of the principles of Greek geometry it cannot be demonstrated, however, that it is impossible to effect by planar means certain solid constructions (like the cube duplication and angle trisection). These results were only established by algebraists in the 19th century (notably by the French mathematician Pierre Laurent Wantzel in 1837).

Linear problems

A third class of problems, called linear, embraced those solvable by means of curves other than the circle and the conics (in Greek, the word for "line," *grammē,* refers to all lines, whether curved or straight). For instance, one group of curves, the conchoids (from the Greek word for shell), are formed by marking off a certain length on a ruler and then pivoting it about a fixed point in such a way that one of the marked points stays on a given line; the other marked point traces out a conchoid. These curves can be used wherever a solution involves the positioning of a marked ruler relative to a given line (in Greek, such constructions are called *neuses,* or "vergings" of a line to a given point). For example, any acute angle (figured as

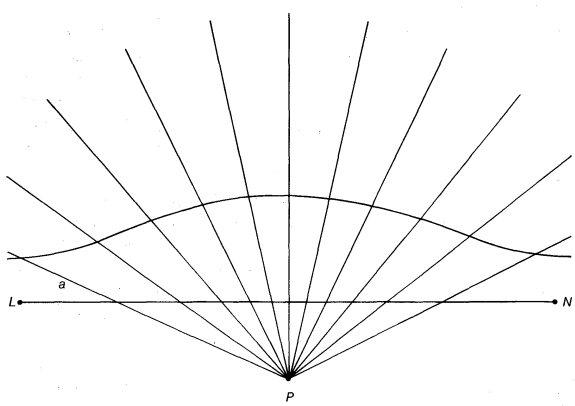


Figure 4: *Conchoid curve.*
From fixed point *P,* several lines are drawn. A standard distance (*a*) is marked along each line from line *LN,* and the connection of the points creates a conchoid curve.

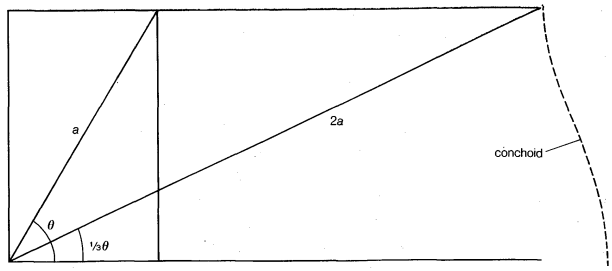


Figure 5: Angle trisection by means of a conchoid.

the angle between one side and the diagonal of a rectangle) can be trisected by taking a length equal to twice the diagonal and moving it about until it comes to be inserted between two other sides of the rectangle. If, instead, the appropriate conchoid relative to either of those sides is introduced, the required position of the line can be determined without the trial and error of a moving ruler. Because the same construction can be effected by means of a hyperbola, however, the problem is not linear but solid. Such uses of the conchoids were presented by Nicomedes (middle or late 3rd century BC), and their replacement by equivalent solid constructions appears to have come soon after, perhaps by Apollonius or his associates.

Some of the curves used for problem solving are not so reducible. For example, the Archimedean spiral couples uniform motion of a point on a half ray with uniform rotation of the ray around a fixed point at its end. Such curves have their principal interest as means for squaring the circle and trisecting the angle.

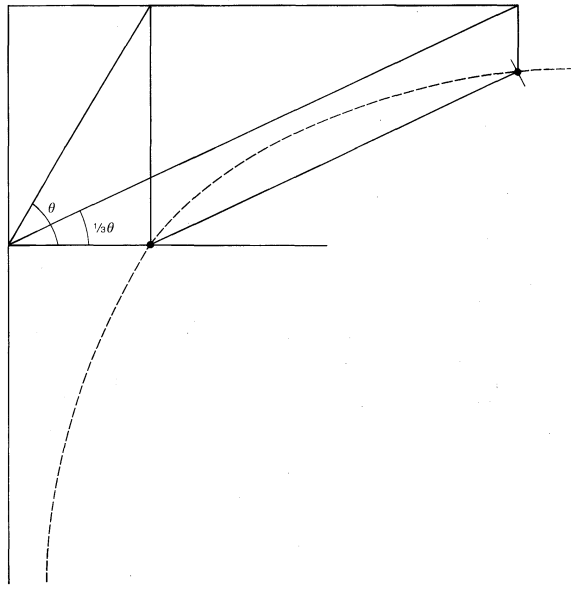


Figure 6: Angle trisection by means of a hyperbola.

Applied geometry. A major activity among geometers in the 3rd century BC was the development of geometric approaches in the study of the physical sciences, specifically, optics, mechanics, and astronomy. In each case the aim was to formulate the basic concepts and principles in terms of geometric and numerical quantities and then to derive the fundamental phenomena of the field by geometric constructions and proofs.

Optics

In optics, Euclid's textbook (called the *Optics*) set the precedent. Euclid postulated visual rays to be straight lines, and he defined the apparent size of an object in terms of the angle formed by the rays drawn from the top and the bottom of the object to the observer's eye. He then proved, for example, that nearer objects appear larger and appear to move faster and showed how to measure the height of distant objects from their shadows or reflected images, and so on. Other textbooks set out theorems on the phenomena of reflection and refraction (the field called catoptrics). The most extensive survey of optical phenomena is a treatise attributed to the astronomer Ptolemy (2nd century AD), which survives only in the form of an incomplete Latin translation (12th century) based on a lost Arabic translation. It covers the fields of geometric optics and catoptrics, as well as experimental areas, such as binocular vision, and more general philosophical principles (the nature of light, vision, and colour). Of a somewhat different sort are the studies of burning mirrors by Diocles (late 2nd century BC), who proved that the surface that reflects the rays from the Sun to a single point is a paraboloid of revolution. Constructions of such devices remained of interest as late as the 6th century AD, when Anthemius of Tralles, best known for his work as architect of the Hagia Sophia at Constantinople, compiled a survey of remarkable mirror configurations.

Mechanics was dominated by the work of Archimedes, who was the first to prove the principle of balance: that two weights are in equilibrium when they are inversely proportional to their distances from the fulcrum. From this principle he developed a theory of the centres of gravity of plane and solid figures. He was also the first to state and prove the principle of buoyancy—that floating bodies displace their equal in weight—and to use it for proving the conditions of stability of segments of spheres and paraboloids, solids formed by rotating a parabolic segment about its axis. Archimedes proved the conditions under which this solid will return to its initial position if tipped, in particular, for the positions now called "stable I" and "stable II," where the vertex faces up and down, respectively.

In his work *Method Concerning Mechanical Theorems,* Archimedes also set out a special "mechanical method" that he used for the discovery of results on volumes and centres of gravity. He employed the bold notion of constituting solids from the plane figures formed as their sections (*e.g.,* the circles that are the plane sections of spheres, cones, cylinders, and other solids of revolution), assigning to such figures a weight proportional to their area. For example, to measure the volume of a sphere, he imagined a balance beam, one of whose arms is a diameter of the sphere with the fulcrum at one endpoint of this diameter and the other arm an extension of the diameter to the other side of the fulcrum by a length equal to the diameter. Archimedes showed that the three circular cross sections made by a plane cutting the sphere and the associated cone and cylinder will be in balance (the circle in the cylinder with the circles in the sphere and cone) if the circle in the cylinder is kept in its original place, while the circles in the sphere and cone are placed with their centres of gravity at the opposite end of the balance. Doing this for all the sets of circles formed as cross sections of these solids by planes, he concluded that the solids themselves are in balance, the cylinder with the sphere and the cone together if the cylinder is left where it is, while the sphere and cone are placed with their centres of gravity at the opposite end of the balance. Since the centre of gravity of the cylinder is the midpoint of its axis, it follows that (*sphere* + *cone*):*cylinder* = 1:2 (by the inverse proportion of weights and distances). Since the volume of the cone is one-third that of the cylinder, however, the volume of the sphere is found to be one-sixth that of the cylinder. In similar manner, Archimedes worked out the volumes and centres of gravity of spherical segments and segments of the solids of revolution of conic sections (paraboloids, ellipsoids, and hyperboloids). The critical notions—constituting solids out of their plane sections and assigning weights to geometric figures—were not formally valid within the standard conceptions of Greek geometry,

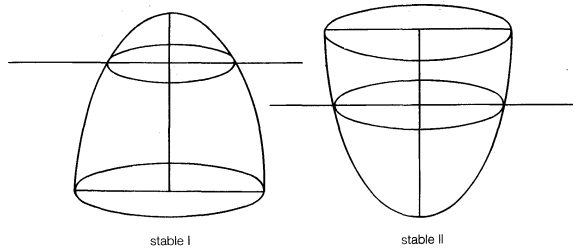


Figure 7: Paraboloids.

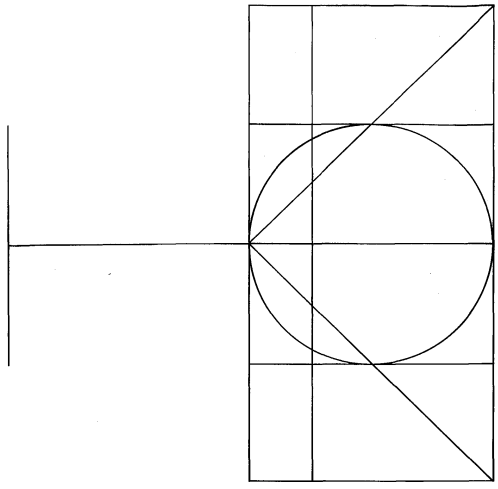


Figure 8: Measurement of the sphere.

and Archimedes admitted this. But he maintained that, although his arguments were not "demonstrations" (*i.e.*, proofs), they had value for the discovery of results about these figures.

Astronomy

The geometric study of astronomy has pre-Euclidean roots, with Eudoxus having developed a model for planetary motions around a stationary Earth. Accepting the principle—which, according to Eudemus, was first proposed by Plato—that only combinations of uniform circular motions are to be used, Eudoxus represented the path of a planet as the result of superimposing rotations of three or more concentric spheres whose axes are set at different angles. Although the fit with the phenomena was unsatisfactory, the curves thus generated (the *hippopede,* or "horse-fetter") continued to be of interest for their geometric properties, as is known through remarks by Proclus. Later geometers continued the search for geometric patterns satisfying the Platonic conditions. The simplest model, a scheme of circular orbits centred on the Sun, was introduced by Aristarchus of Samos (3rd century BC), but this was rejected by others, since a moving Earth was judged to be impossible on physical grounds. But Aristarchus' scheme could have suggested use of an "eccentric" model, in which the planets rotate about the Sun and the Sun in turn rotates about the Earth. Apollonius introduced an alternative "epicyclic" model, in which the planet turns about a point that itself orbits in a circle (the "deferent") centred at or near the Earth. As Apollonius knew, his epicyclic model is geometrically equivalent to an eccentric. These models were well adapted for explaining other phenomena of planetary motion. For instance, if the Earth is displaced from the centre of a circular orbit (as in the eccentric scheme), the orbiting body will appear to vary in speed (appearing faster when nearer the observer, slower when farther away), as is in fact observed for the Sun, Moon, and planets. By varying the relative sizes and rotation rates of the epicycle and deferent, in combination with the eccentric, a flexible device may be obtained for representing planetary motion.

LATER TRENDS IN GEOMETRY AND ARITHMETIC

Greek trigonometry and mensuration. After the 3rd century BC, mathematical research shifted increasingly away from the pure forms of constructive geometry toward areas related to the applied disciplines, in particular to astronomy. The necessary theorems on the geometry of the sphere (called spherics) were compiled into textbooks, such as the one by Theodosius (3rd or 2nd century BC), that consolidated the earlier work by Euclid and the work of Autolycus of Pitane (fl. *c.* 300 BC) on spherical astronomy. More significantly, in the 2nd century BC the Greeks first came into contact with the fully developed Mesopotamian astronomical systems and took from them many of their observations and parameters (for example, values for the average periods of astronomical phenomena). While retaining their own commitment to geometric models rather than adopting the arithmetic schemes of the Mesopotamians, the Greeks, nevertheless, followed the Mesopotamians' lead in seeking a predictive astronomy based on a combination of mathematical theory and observational parameters. They thus made it their goal not merely to describe but to calculate the angular positions of the planets on the basis of the numerical and geometric content of the theory. This major restructuring of Greek astronomy, in both its theoretical and practical respects, was due primarily to Hipparchus (2nd century BC), whose work was consolidated and further advanced by Ptolemy.

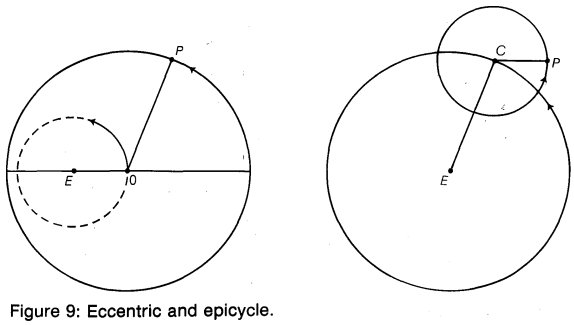


Figure 9: Eccentric and epicycle.

The measurement of angles

To facilitate their astronomical researches, the Greeks developed techniques for the numerical measurement of angles, a precursor of trigonometry, and produced tables suitable for practical computation. Early efforts to measure the numerical ratios in triangles were made by Archimedes and Aristarchus. Their results were soon extended, and comprehensive treatises on the measurement of chords (in effect, a construction of a table of values equivalent to the trigonometric sine) were produced by Hipparchus and by Menelaus of Alexandria (early 2nd century AD). These works are now lost, but the essential theorems and tables are preserved in Ptolemy's *Almagest* (Book I, chapter 10). For computing with angles, the Greeks adopted the Mesopotamian sexagesimal method in arithmetic, whence it survives in the standard units for angles and time employed to this day.

Number theory. Although Euclid handed down a precedent for number theory in Books VII–IX of the *Elements,* later writers made no further effort to extend the field of theoretical arithmetic in his demonstrative manner. Beginning with Nicomachus of Gerasa (fl. *c.* AD 100), several writers produced collections expounding a much simpler form of number theory. A favourite result is the representation of arithmetic progressions in the form of "polygonal numbers." For instance, if the numbers 1, 2, 3, 4, . . . are added successively, the "triangular" numbers 1, 3, 6, 10, . . . are obtained; similarly, the odd numbers 1, 3, 5, 7, . . . sum to the "square" numbers 1, 4, 9, 16, . . . , while the sequence 1, 4, 7, 10, . . . , with a constant difference of 3, sums to the "pentagonal" numbers 1, 5, 12, 22, ... In general, these results can be expressed in the form of geometric shapes formed by lining up dots in the appropriate two-dimensional configurations. In the ancient arithmetics, such results are invariably presented as particular cases, without any general notational method or general proof. The writers in this tradition are called neo-Pythagoreans, since they viewed themselves as continuing the Pythagorean school of the 5th century BC, and in the spirit of ancient Pythagoreanism they tied their numerical interests to a philosophical theory that was an amalgam of Platonic metaphysical and theological doctrines. With its exponent Iamblichus of Chalcis (4th century AD), neo-Pythagoreans became a prominent part of the revival of pagan religion in opposition to Christianity in late antiquity.

An interesting concept of this school of thought, which Iamblichus attributes to Pythagoras himself, is that of "amicable numbers"; two numbers are amicable if each is equal to the sum of the proper divisors of the other (for example, 220 and 284). Attributing virtues such as friendship and justice to numbers was characteristic of the Pythagoreans at all times.

Diophantus of Alexandria

Of much greater mathematical significance is the arithmetic work of Diophantus of Alexandria (active at an unknown time between the 2nd century BC and the 3rd

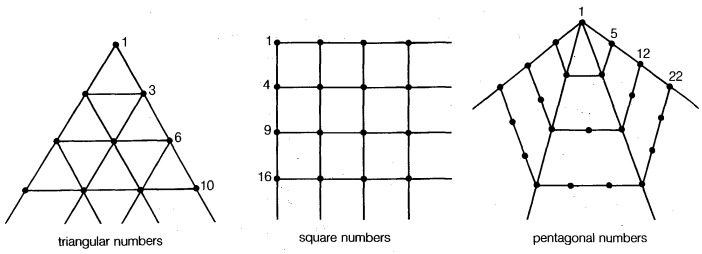


Figure 10: Polygonal arrays.

century AD). His writing, the *Arithmetica,* originally in 13 books (six survive in Greek, another four in the medieval Arabic translation), sets out hundreds of arithmetic problems with their solutions. For example, Book II, problem 8, seeks to express a given square number as the sum of two square numbers (here and throughout, the "numbers" are rational). Like those of the neo-Pythagoreans, his treatments are always of particular cases rather than general solutions; thus, in this problem the given number is taken to be 16 and the solutions worked out are $^{256}/_{25}$ and $^{144}/_{25}$. In this example, as is often the case, the solutions are not unique; indeed, in the very next problem Diophantus shows how a number given as the sum of two squares (*e.g.,* $13 = 4 + 9$) can be expressed differently as the sum of two other squares (for example, $13 = {}^{324}/_{25} + {}^{1}/_{25}$).

To find his solutions, Diophantus adopted an arithmetic form of the method of analysis. He first reformulated the problem in terms of one of the unknowns, and he then manipulated it as if it were known until an explicit value for the unknown emerged. He even adopted an abbreviated notational scheme to facilitate such operations, where, for example, the unknown is symbolized by a figure somewhat resembling the Roman letter *S.* (This is a standard abbreviation for the word number in ancient Greek manuscripts.) Thus, in the first problem discussed above, if S is one of the unknown solutions, then $16 - S^2$ is a square, supposing the other unknown to be $2S - 4$ (where the 2 is arbitrary but the 4 chosen because it is the square root of the given number 16), Diophantus found from summing the two unknowns ($[2S - 4]^2$ and S^2) that $4S^2 - 16S + 16 + S^2 = 16$, or $5S^2 = 16S$, that is, $S = {}^{16}/_{5}$. So one solution is $S^2 = {}^{256}/_{25}$, while the other solution is $16 - S^2$, or $^{144}/_{25}$.

Survival and influence of Greek mathematics. Notable in the closing phase of Greek mathematics were Pappus (early 4th century AD), and Theon (late 4th century) and his daughter Hypatia (d. 415). All were active in Alexandria as professors of mathematics and astronomy, and they produced extensive commentaries on the major authorities—Pappus and Theon on Ptolemy, Hypatia on Diophantus and Apollonius. Later, Eutocius of Ascalon (early 6th century) produced commentaries on Archimedes and Apollonius. While much of their output has since been lost, much survives. They proved themselves reasonably competent in technical matters but little inclined toward significant insights (their aim was usually to fill in minor steps assumed in the proofs, to append alternative proofs, and the like), and the level of originality was very low. But these scholars frequently preserved fragments of older works that are now lost, and their teaching and editorial efforts assured the survival of the works of Euclid, Archimedes, Apollonius, Diophantus, Ptolemy, and others that now do exist, either in Greek manuscripts or in medieval translations (Arabic, Hebrew, and Latin) derived from them.

The legacy of Greek mathematics, particularly in the fields of geometry and geometric science, was enormous. From an early period the Greeks formulated the objectives of mathematics not in terms of practical procedures but as a theoretical discipline committed to the development of general propositions and formal demonstrations. The range and diversity of their findings, especially those of the masters of the 3rd century BC, supplied geometers with subject matter for centuries thereafter, even though the tradition that was transmitted into the Middle Ages and Renaissance was incomplete and defective.

The rapid rise of mathematics in the 17th century was based in part on the conscious imitation of the ancient classics and on competition with them. In the geometric mechanics of Galileo and the infinitesimal researches of Kepler and Cavalieri, it is possible to perceive a direct inspiration from Archimedes. The study of the advanced geometry of Apollonius and Pappus stimulated new approaches in geometry—for example, the analytic methods of Descartes and the projective theory of Girard Desargues. Purists like Huygens and Newton insisted on the Greek geometric style as a model of rigour, just as others sought to escape its forbidding demands of completely worked-out proofs. The full impact of Diophantus' work is evident particularly with Pierre de Fermat in his researches in algebra and number theory. Although mathematics has today gone far beyond the ancient achievements, the leading figures of antiquity, like Archimedes, Apollonius, and Ptolemy, can still be rewarding reading for the ingenuity of their insights. (W.R.K.)

Mathematics in medieval Islām

Origins. In Hellenistic times and in late antiquity, scientific learning in the eastern part of the Roman world was spread over a variety of centres, and Justinian's closing of the pagan academies in Athens in 529 gave further impetus to this diffusion. An additional factor was the translation and study of Greek scientific and philosophical texts sponsored both by monastic centres of the various Christian churches in the Levant, Egypt, and Mesopotamia and by enlightened rulers of the Sāsanian dynasty in places like the medical school at Gondeshapur.

Indian astronomy and mathematics

Also important were developments in India in the first few centuries AD. Although the decimal system for whole numbers was apparently not known to the Indian astronomer Āryabhaṭa I (b. 476), it was used by his pupil Bhāskara I in 620, and by 670 the system had reached northern Mesopotamia, where the Nestorian bishop Severus Sebokht praised its Hindu inventors as discoverers of things more ingenious than those of the Greeks. Earlier, in the late 4th or early 5th century, the anonymous Hindu author of an astronomical handbook, the *Sūrya Siddhānta,* had tabulated the sine function (unknown in Greece) for every $3\frac{3}{4}°$ of arc from $3\frac{3}{4}°$ to 90°.

Within this intellectual context the rapid expansion of Islām took place between the time of Muḥammad's return to Mecca from his exile in Medina in 630 and the Muslim conquest of lands extending from Spain to the borders of China by 715. Not long afterward, Muslims began the acquisition of foreign learning, and, by the time of the caliph al-Manṣūr (d. 775), such Indian and Persian astronomical material as the *Brāhma-sphuṭa-siddhānta* and the *Shāh's Tables* had been translated into Arabic. The subsequent acquisition of Greek material was greatly advanced when the caliph al-Ma'mūn constructed a translation and research centre, the House of Wisdom, in Baghdad during his reign (813–833). Most of the translations were done from Greek and Syriac by Christian scholars, but the impetus and support for this activity came from Muslim patrons. These included not only the caliph but also wealthy individuals such as the three brothers known as the Banū Mūsā, whose treatises on geometry and mechanics formed an important part of the works studied in the Islāmic world.

Translation of Greek texts

Of Euclid's works, the *Elements,* the *Data,* the *Optics,* the *Phaenomena,* and *On Divisions* were translated. Of Archimedes' works only two—*Sphere and Cylinder* and *Measurement of the Circle*—are known to have been translated, but these were sufficient to stimulate independent researches from the 9th to the 15th century. On the other hand, virtually all of Apollonius' works were translated, and of Diophantus and Menelaus one book each, the *Arithmetica* and the *Sphaerica,* respectively, were translated into Arabic. Finally, the translation of Ptolemy's *Almagest* furnished important astronomical material.

Of the minor writings, Diocles' treatise on mirrors, Theodosius' *Spherics,* Pappus' work on mechanics, Ptolemy's *Planisphaerium,* and Hypsicles' treatises on regular polyhedra (the so-called Books XIV and XV of Euclid's *Elements*) were among those translated.

Mathematics in the 9th century. Thābit ibn Qurrah (836–901), a Sabian from Ḥarrān in northern Mesopotamia, was an important translator and reviser of these Greek works. In addition to translating works of the major Greek mathematicians (for the Banū Mūsā, among others), he was a court physician. He also translated Nicomachus of Gerasa's *Arithmetic* and discovered a beautiful rule for finding amicable numbers, a pair of numbers such that each number is the sum of the set of proper divisors of the other number. The investigation of such numbers formed a continuing tradition in Islām. Kamāl ad-Dīn al-Fārisī (d. *c.* 1320) gave the pair 17,926 and 18,416 as an example of Thābit's rule, and in the 17th century Muḥammad Bāqir Yazdī gave the pair 9,363,584 and 9,437,056.

One scientist typical of the 9th century was Muhammad ibn Mūsā al-Khwārizmī. Working in the House of Wisdom, he introduced Indian material in his astronomical works and also wrote an early book explaining Hindu arithmetic, the *Book of Addition and Subtraction According to the Hindu Calculation.* In another work, the *Book of Restoring and Balancing,* he provided a systematic introduction to algebra, including a theory of quadratic equations. Both works had important consequences for Islāmic mathematics. *Hindu Calculation* began a tradition of arithmetic books that, by the middle of the next century, led to the invention of decimal fractions (complete with a decimal point), and his *Restoring and Balancing* became the point of departure and model for later writers such as the Egyptian Abū Kāmil. Both books were translated into Latin, and *Restoring and Balancing* was the origin of the word algebra, from the Arabic word for "restoring" in its title (*al-jabr*). The *Hindu Calculation,* from a Latin form of the author's name, *algorismi,* yielded the word algorithm.

Al-Khwārizmī's algebra also served as a model for later writers in its application of arithmetic and algebra to the distribution of inheritances according to the complex requirements of Muslim religious law. This tradition of service to the Islāmic faith was an enduring feature of mathematical work in Islām and one which, in the eyes of many, justified the study of secular learning. In the same category are al-Khwārizmī's method of calculating the time of visibility of the new moon (which signals the beginning of the Muslim month) and the expositions by astronomers of methods for finding the direction to Mecca for the five daily prayers.

Mathematics in the 10th century. Islāmic scientists in the 10th century were involved in three major mathematical projects: the completion of arithmetic algorithms, the development of algebra, and the extension of geometry.

Arithmetic

The first of these projects led to the appearance of three complete numeration systems, one of which was the finger arithmetic used by the scribes and treasury officials. This ancient arithmetic system, which became known throughout the East and Europe, employed mental arithmetic and a system of storing intermediate results on the fingers as an aid to memory. (Its use of unit fractions recalls the Egyptian system.) During the 10th and 11th centuries capable mathematicians, such as Abū al-Wafā' (940–997/8), wrote on this system, but it was eventually replaced by the decimal system.

A second common system was the base-60 numeration inherited from the Babylonians via the Greeks and known as the arithmetic of the astronomers. Although astronomers used this system for their tables, they usually converted numbers to the decimal system for complicated calculations and then converted the answer back to sexagesimals.

The third system was Indian arithmetic, whose basic numeral forms, complete with the zero, eastern Islām took over from the Hindus. (Different forms of the numerals, whose origins are not entirely clear, were used in the western part of Islām.) The basic algorithms also came from India, but these were adapted by al-Uqlīdisī (*c.* 950) to pen and paper instead of the traditional dust board, a move that helped to popularize this system. Also, the arithmetic algorithms were completed in two ways: by the extension of root-extraction procedures, known to Hindus and Greeks only for square and cube roots, to roots of higher degree, and by the extension of the Hindu decimal system for whole numbers to include decimal fractions. These fractions appear simply as computational devices in the work of both al-Uqlīdisī and al-Baghdādī (*c.* 1000), but in subsequent centuries they received systematic treatment as a general method. As for extraction of roots, Abū al-Wafā' wrote a treatise (now lost) on the topic, and Omar Khayyam ('Umar al-Khayyāmī [1048–1131]) solved the general problem of extracting roots of any desired degree. Omar's treatise, too, is lost but the method is known from other writers, and it appears that a major step in its development was al-Karajī's 10th-century derivation by means of mathematical induction of the binomial theorem for whole-number exponents—*i.e.,* his discovery that

$$(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2}a^{n-1}b^2 + \frac{n(n-1)(n-2)}{2\cdot 3}a^{n-3}b^3 + \ldots + nab^{n-1} + b^n.$$

Algebra

During the 10th century Islāmic algebraists progressed from al-Khwārizmī's quadratic polynomials to the mastery of the algebra of expressions involving arbitrary positive or negative integral powers of the unknown. Several algebraists explicitly stressed the analogy between the rules for working with powers of the unknown in algebra and those for working with powers of 10 in arithmetic, and there was interaction between the development of arithmetic and algebra from the 10th to the 12th century. A 12th-century student of al-Karajī's works, as-Samaw'al, was able to approximate the quotient $(20x^2 + 30x)/(6x^2 + 12)$ as

$$3\tfrac{1}{3} + 5\left(\frac{1}{x}\right) - 6\tfrac{2}{3}\left(\frac{1}{x^2}\right) - 10\left(\frac{1}{x^3}\right) + 13\tfrac{1}{3} \ldots - 40\left(\frac{1}{x^7}\right)$$

and also gave a rule for finding the coefficients of the successive powers of $1/x$. Although none of this employed symbolic algebra, algebraic symbolism was in use by the 14th century in the western part of the Islāmic world. The context for this well-developed symbolism was, it seems, commentaries that were destined for teaching purposes, such as that of Ibn Qunfūdh (1330–1407) of Algeria on the algebra of Ibn al-Bannā' (1256–1321) of Morocco.

Other parts of algebra developed as well. Both Greeks and Hindus had studied indeterminate equations, and the translation of this material and the application of the newly developed algebra led to the investigation of Diophantine equations by writers like Abū Kāmil, al-Karajī, and Abū Ja'far al-Khāzin (first half of 10th century), as well as to attempts to prove a special case of what is now known as Fermat's last theorem, namely that there are no rational solutions to $x^3 + y^3 = z^3$. The great scientist Alhazen (Ibn al-Haytham [965–1041]) solved problems involving congruences by what is now called Wilson's theorem, which states that, if p is a prime, then p divides $(p-1)\times(p-2)\ldots\times 2\times 1 + 1$, and al-Baghdādī gave a variant of the idea of amicable numbers by defining two numbers to "balance" if the sums of their divisors are equal.

Geometry

However, not only arithmetic and algebra but geometry too underwent extensive development. Thābit ibn Qurrah, his grandson Ibrāhīm ibn Sinān (909–946), Abū Sahl al-Kūhī (d. *c.* 995), and Alhazen solved problems involving the pure geometry of conic sections, including the areas and volumes of plane and solid figures formed from them, and also investigated the optical properties of mirrors made from conic sections. Ibrāhīm ibn Sinān, Abu Sahl al-Kūhī, and Alhazen used the ancient technique of analysis to reduce the solution of problems to constructions involving conic sections. (Alhazen, for example, used this method to find the point on a convex spherical mirror at which a given object is seen by a given observer.) Thābit and Ibrāhīm showed how to design the curves needed for sundials. Abū al-Wafā', whose book on the arithmetic of the scribes is mentioned above, also wrote on geometric methods needed by artisans.

In addition, in the late 10th century Abū al-Wafā' and the prince Abū Naṣr Manṣur stated and proved theorems

of plane and spherical geometry that could be applied by astronomers and geographers, including the laws of sines and tangents. Abū Naṣr's pupil al-Bīrūnī (973–1050), who produced a vast amount of high-quality work, was one of the masters in applying these theorems to astronomy and to such problems in mathematical geography as the determination of latitudes and longitudes, the distances between cities, and the direction from one city to another.

Omar Khayyam. The mathematician and poet Omar Khayyam was born in Neyshābūr (in modern Iran) only a few years before al-Bīrūnī's death. He later lived in Samarkand and Eṣfahān, and his brilliant work there continued many of the main lines of development in 19th-century mathematics. Not only did he discover a general method of extracting roots of arbitrary high degree, but his *Algebra* contains the first complete treatment of the solution of cubic equations. Omar did this by means of conic sections, but he declared his hope that his successors would succeed where he had failed in finding an algebraic formula for the roots.

Euclid's parallel postulate

Omar was also a part of an Islāmic tradition, which included Thābit and Alhazen, of investigating Euclid's parallel postulate. To this tradition Omar contributed the idea of a quadrilateral with two congruent sides perpendicular to the base. The parallel postulate would be proved, Omar recognized, if he could show that the remaining two angles were right angles. In this he failed, but his question about the quadrilateral became the standard way of discussing the parallel postulate.

That postulate, however, was only one of the questions on the foundations of mathematics that interested Islāmic scientists. Another was the definition of ratios. Omar Khayyam, along with others before him, felt that the theory in Euclid's Book V was logically satisfactory but intuitively unappealing, so he proved that a definition known to Aristotle was equivalent to that given in Euclid. In fact, Omar argued that ratios should be regarded as "ideal numbers," and so he conceived of a much broader system of numbers than that used since Greek antiquity, that of the positive real numbers.

Islāmic mathematics to the 15th century. In the 12th century the physician as-Samaw'al continued and completed the work of al-Karajī in algebra and also provided a systematic treatment of decimal fractions as a means of approximating irrational quantities. In his method of finding roots of pure equations, $x^n = N$, he used what is now known as Horner's method to expand the binomial $(a + y)^n$. His contemporary, Sharaf ad-Dīn aṭ-Ṭūsī, late in the 12th century provided a method of approximating the positive roots of arbitrary equations, based on an approach virtually identical to that discovered by François Viète in 16th-century France. The important step here was less the general idea than the development of the numerical algorithms necessary to effect it.

Sharaf ad-Dīn was the discoverer of a device, called the linear astrolabe, that places him in another important Islāmic mathematical tradition, one that centred on the design of new forms of the ancient astronomical instrument known as the astrolabe. The astrolabe, whose mathematical theory is based on the stereographic projection of the sphere, was invented in late antiquity, but its extensive development in Islām made it the pocket watch of the medievals. In its original form, it required a different plate of horizon coordinates for each latitude, but in the 11th century the Spanish Muslim astronomer az-Zarqallu invented a single plate that worked for all latitudes. Slightly earlier, astronomers in the East had experimented with plane projections of the sphere, and al-Bīrūnī invented such a projection that could be used to produce a map of a hemisphere. The culminating masterpiece was the astrolabe of the Syrian Ibn ash-Shāṭir (1305–75), a mathematical tool that could be used to solve all the standard problems of spherical astronomy in five different ways.

Development of trigonometry

On the other hand, Muslim astronomers had developed other methods for solving these problems using the highly accurate trigonometric tables and the new trigonometric theorems they had developed. Out of these developments came the creation of trigonometry as a mathematical discipline, separate from its astronomical applications, by Naṣīr ad-Dīn aṭ-Ṭūsī at his observatory in Marāgheh in the 13th century. (It was there, too, that Naṣīr ad-Dīn's pupil, Quṭb ad-Dīn ash-Shīrāzī [1236–1311], and his pupil, Kamāl ad-Dīn Fārisī, using Alhazen's great work, the *Optics,* were able to give the first mathematically satisfactory explanation of the rainbow.)

Nasīr ad-Dīn's observatory was supported by a grandson of Genghis Khan, Hülegü, who sacked Baghdad in 1258. Ulugh Beg, the grandson of the Mongol conqueror Timur, founded an observatory at Samarkand in the early years of the 15th century. Ulugh Beg was himself a good astronomer, and his five (sexagesimal) place tables of sines and tangents for every minute of arc were one of the great achievements in numerical mathematics up to his time. He was also the patron of Jamshīd al-Kāshī (d. 1429), whose work *The Reckoners' Key* summarizes most of the arithmetic of his time and includes sections on algebra and practical geometry as well. Among al-Kāshī's works are a masterful computation of the value of 2π, which, when expressed in decimal fractions, is accurate to 16 places, as well as the application of a numerical method, now known as fixed-point iteration, for solving the cubic equation with sin 1° as a root. His work was indeed of a quality deserving Ulugh Beg's description as "known among the famous of the world."

Al-Kāshī lived almost five centuries after the first translations of Arabic material into Latin, and by his time the Islāmic mathematical tradition had given the West not only its first versions of many of the Greek classics but also a complete set of algorithms for Hindu-Arabic arithmetic, plane and spherical trigonometry, and the powerful tool of algebra. Although mathematical inquiry continued in Islām in the centuries after al-Kāshī's time, the mathematical centre of gravity was shifting to the West. That this was so is due, of course, in no small measure to what the Western mathematicians had learned from their Islāmic predecessors during the preceding centuries. (J.L.B.)

European mathematics during the Middle Ages and Renaissance

Until the 11th century only a small part of the Greek mathematical corpus was known in the West. Because almost no one could read Greek, what little was available came from the poor texts written in Latin in the Roman Empire, together with the very few Latin translations of Greek works. Of these the most important were the treatises by Boethius, who in about AD 500 made Latin redactions of a number of Greek scientific and logical writings. His *Arithmetic,* which was based on Nicomachus, was well known and was the means by which medieval scholars learned of Pythagorean number theory. Boethius and Cassiodorus provided the material for the part of the monastic education called the quadrivium: arithmetic, geometry, astronomy, and music theory. Together with the trivium (grammar, logic, rhetoric), these subjects formed the seven liberal arts, which were taught in the monasteries, cathedral schools, and, from the 12th century on, in the universities and which constituted the principal university instruction until modern times.

Monastic education

For monastic life it sufficed to know how to calculate with Roman numerals. The principal application of arithmetic was a method for determining the date of Easter, the computus, that was based on the lunar cycle of 19 solar years (*i.e.,* 235 lunar revolutions) and the 28-year solar cycle. Between the time of Bede (d. 735), when the system was fully developed, and about 1500, the computus was reduced to a series of verses that were learned by rote. Until the 12th century, geometry was largely concerned with approximate formulas for measuring areas and volumes in the tradition of the Roman surveyors. About AD 1000 the French scholar Gerbert of Aurillac, later Pope Sylvester II, introduced a type of abacus, in which numbers were represented by stones bearing Arabic numerals. Such novelties were known to very few.

The transmission of Greek and Arabic learning. In the 11th century a new phase of mathematics began with the translations from Arabic. Scholars throughout Europe went to Toledo, Córdoba, and elsewhere in Spain to translate

into Latin the accumulated learning of the Muslims. Along with philosophy, astronomy, astrology, and medicine, important mathematical achievements of the Greek, Indian, and Islāmic civilizations became available in the West. Particularly important were Euclid's *Elements,* the works of Archimedes, and al-Khwārizmī's treatises on arithmetic and algebra. Western texts called *algorismus* (a Latin form of the name al-Khwārizmī), introduced the Hindu-Arabic numerals and applied them in calculations. Thus modern numerals first came into use in universities and then became common among merchants and other laymen. It should be noted that, up to the 15th century, calculations were often performed with board and counters. Reckoning with Hindu-Arabic numerals was used by merchants at least from the time of Leonardo of Pisa (beginning of the 13th century) first in Italy, then in the trading cities of southern Germany and France, where *maestri d'abbaco* or *Rechenmeister* taught commercial arithmetic in the various vernaculars. Some schools were private, while others were run by the community.

The universities. Mathematics was studied from a theoretical standpoint in the universities. The universities of Paris and Oxford, which were founded relatively early (*c.* 1200), were centres for mathematics and philosophy. Of particular importance in these universities were the Arabic-based versions of Euclid, of which there were at least four by the 12th century. Of the numerous redactions and compendia which were made, that of Campanus (*c.* 1250; first printed in 1482) was easily the most popular, serving as a textbook for many generations. Such redactions of the *Elements* were made to help students not only to understand Euclid's textbook but also to handle other, particularly philosophical, questions suggested by passages in Aristotle. The ratio theory of the *Elements* provided a means of expressing the various relations of the quantities associated with moving bodies, relations that now would be expressed by formulas. Also in Euclid were to be found methods of analyzing infinity and continuity (paradoxically, because Euclid always avoided infinity).

Influence of Euclid's *Elements*

Studies of such questions led not only to new results but also to a new approach to what is now called physics. Thomas Bradwardine, who was active in Merton College, Oxford, in the first half of the 14th century, was one of the first medieval scholars to ask whether the continuum can be divided infinitely, or whether there are smallest parts (indivisibles). Among other topics, he compared different geometric shapes in terms of the multitude of points that were assumed to compose them, and from such an approach paradoxes were generated that were not to be solved for centuries. Another fertile question stemming from Euclid concerned the angle between a circle and a line tangent to it (called the horn angle): if this angle is not zero, a contradiction quickly ensues, but, if it is zero, then, by definition, there can be no angle. For the relation of force, resistance, and the speed of the body moved by this force, Bradwardine suggested an exponential law. Nicholas Oresme (d. 1382) extended Bradwardine's ideas to fractional exponents.

Another question having to do with the quantification of qualities, the so-called latitude of forms, began to be discussed at about this time in Paris and in Merton College, Oxford. Various Aristotelian qualities (*e.g.,* heat, density, and velocity) were assigned an intensity and extension, which were sometimes represented by the height and bases (respectively) of a geometric figure. The area of the figure was then considered to represent the quantity of the quality. In the important case in which the quality is the motion of a body, the intensity its speed, and the extension its time, the area of the figure was taken to represent the distance covered by the body. Uniformly accelerated motion starting at zero velocity gives rise to a triangular figure (see Figure 11). It was proved by the Merton school that the quantity of motion in such a case is equal to the quantity of a uniform motion at the speed achieved halfway through the accelerated motion; in modern formulation: $s = \frac{1}{2}at^2$ (Merton rule). Discussions like this certainly influenced Galileo indirectly and may have influenced the founding of coordinate geometry in the 17th century. Another important development in the scholastic "calculations" was the summation of infinite series.

The Merton rule

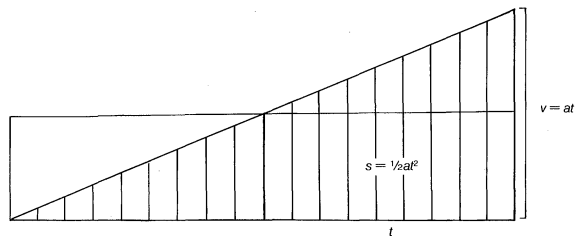


Figure 11: Uniformly accelerated motion; s = speed, a = acceleration, t = time, and v = velocity.

Basing his work on translated Greek sources, the German mathematician and astronomer Regiomontanus wrote the first book in the West on plane and spherical trigonometry independent of astronomy in about 1464 (printed in 1533). He also published tables of sines and tangents that were in constant use for more than two centuries.

The Renaissance. Italian artists and merchants influenced the mathematics of the late Middle Ages and the Renaissance in several ways. In the 15th century a group of Tuscan artists, including Filippo Brunelleschi, Leon Battista Alberti, and Leonardo da Vinci, incorporated linear perspective into their practice and teaching, about a century before the subject was formally treated by mathematicians. Italian *maestri d'abbaco* tried, albeit unsuccessfully, to solve nontrivial cubic equations. In fact, the first general solution was found by Scipione Del Ferro at the beginning of the 16th century and rediscovered by Niccolò Tartaglia several years later. The solution was published by Girolamo Cardano in his *Ars magna* in 1545, together with Lodovico Ferrari's solution of the quartic equation.

By 1380 an algebraic symbolism had been developed in Italy in which letters were used for the unknown, for its square, and for constants. The symbols used today for the unknown (for example, x), the square root sign, and the signs + and − came into general use in southern Germany beginning in about 1450. They were used by Regiomontanus and by Fridericus Gerhart and received an impetus in about 1486 at the University of Leipzig from Johann Widman. The idea of distinguishing between known and unknown quantities in algebra was first consistently applied by François Viète, with vowels for unknown and consonants for known quantities. Viète found some relations between the coefficients of an equation and its roots. This was suggestive of the idea, explicitly stated by Albert Girard in 1629 and proved by Gauss in 1799, that an equation of degree n has n roots. Complex numbers, which are implicit in such ideas, were gradually accepted about the time of Rafael Bombelli (d. 1572), who used them in connection with the cubic.

Apollonius' *Conics* and the investigations of areas (quadratures) and of volumes (cubatures) of Archimedes formed part of the humanistic learning of the 16th century. These studies strongly influenced the later developments of analytic geometry, the infinitesimal calculus, and the theory of functions, subjects that were developed in the 17th century. (Me.F.)

Mathematics in the 17th and 18th centuries

THE 17TH CENTURY

The 17th century, the period of the scientific revolution, witnessed the consolidation of Copernican heliocentric astronomy and the establishment of inertial physics in the work of Kepler, Galileo, Descartes, and Newton. This period was also one of intense activity and innovation in mathematics. Advances in numerical calculation, the development of symbolic algebra and analytic geometry, and the invention of the differential and integral calculus resulted in a major expansion of the subject areas of mathematics. By the end of the 17th century a program of research based in analysis had replaced classical Greek geometry at the centre of advanced mathematics. In the next century this program would continue to develop in close association with physics, more particularly mechanics and theoretical astronomy. The extensive use of

Development of analysis

analytic methods, the incorporation of applied subjects, and the adoption of a pragmatic attitude to questions of logical rigour distinguished the new mathematics from traditional geometry.

Institutional background. Until the middle of the 17th century, mathematicians worked alone or in small groups, publishing their work in books or communicating with other researchers by letter. At a time when people were often slow to publish, "invisible colleges," networks of scientists who corresponded privately, played an important role in coordinating and stimulating mathematical research. Marin Mersenne in Paris acted as a clearinghouse for new results, informing his many correspondents—including Fermat, Descartes, Blaise Pascal, Gilles Personne de Roberval, and Galileo—of challenge problems and novel solutions. Later in the century John Collins, librarian of London's Royal Society, performed a similar function among British mathematicians.

In 1660 the Royal Society of London was founded, to be followed in 1666 by the Academy of Sciences in France, in 1700 by the Berlin Academy, and in 1724 by the St. Petersburg Academy. The official publications sponsored by the academies, as well as independent journals such as the *Acta Eruditorum* (founded in 1682), made possible the open and prompt communication of research findings. Although universities in the 17th century provided some support for mathematics, they became increasingly ineffective as state-supported academies assumed direction of advanced research.

Numerical calculation. The development of new methods of numerical calculation was a response to the increased practical demands of numerical computation, particularly in trigonometry, navigation, and astronomy. New ideas spread quickly across Europe and resulted by 1630 in a major revolution in numerical practice.

Simon Stevin of Holland, in his short pamphlet *La Disme* (1585), introduced decimal fractions to Europe and showed how to extend the principles of Hindu-Arabic arithmetic to calculation with these numbers. Stevin emphasized the utility of decimal arithmetic "for all accounts that are encountered in the affairs of men," and he explained in an appendix how it could be applied to surveying, stereometry, astronomy, and mensuration. His idea was to extend the base-10 positional principle to numbers with fractional parts, with a corresponding extension of notation to cover these cases. In his system the number 237.578 was denoted

$$237\ ⓪\ 5\ ①\ 7\ ②\ 8\ ③,$$

in which the digits to the left of the zero are the integral part of the number. To the right of the zero are the digits of the fractional part, with each digit succeeded by a circled number that indicates the negative power to which 10 is raised. Stevin showed how the usual arithmetic of whole numbers could be extended to decimal fractions using rules that determined the positioning of the negative powers of 10.

In addition to its practical utility, *La Disme* was significant for the way it undermined the dominant style of classical Greek geometry in theoretical mathematics. Stevin's proposal required a rejection of the distinction in Euclidean geometry between magnitude, which is continuous, and number, which is a multitude of indivisible units. For Euclid, unity, or one, was a special sort of thing, not number but the origin, or principle, of number. The introduction of decimal fractions seemed to imply that the unit could be subdivided and that arbitrary continuous magnitude could be represented numerically; it implicitly supposed the concept of a general positive real number.

Logarithms

Tables of logarithms were first published in 1614 by the Scottish baron John Napier in his treatise *Mirifici Logarithmorum Canonis Descriptio* (*Description of the Marvelous Canon of Logarithms*). This work was followed (posthumously) five years later by another in which Napier set forth the principles used in the construction of his tables. The basic idea behind logarithms is that addition and subtraction are easier to perform than multiplication and division, which, as Napier observed, require a "tedious expenditure of time" and are subject to "slippery errors." By the law of exponents, $a^n a^m = a^{n+m}$, that is, in the multiplication of numbers the exponents are related additively. By correlating the geometric sequence of numbers a, a^2, a^3, . . . (a is called the base) and the arithmetic sequence 1, 2, 3, . . . and interpolating to fractional values, it is possible to reduce the problem of multiplication and division to one of addition and subtraction. To do this Napier chose a base that was very close to 1, differing from it by only $1/10^7$. The resulting geometric sequence therefore yielded a dense set of values, suitable for constructing a table.

In his work of 1619 Napier presented an interesting kinematic model to generate the geometric and arithmetic sequences used in the construction of his tables. Assume two particles move along separate lines from given initial points. The particles begin moving at the same instant with the same velocity. The first particle continues to move with a speed that is decreasing, proportional at each instant to the distance remaining between it and some given fixed point on the line. The second particle moves with a constant speed equal to its initial velocity. Given any increment of time, the distances traveled by the first particle in successive increments form a geometrically decreasing sequence. The corresponding distances traveled by the second particle form an arithmetically increasing sequence. Napier was able to use this model to derive theorems yielding precise limits to approximate values in the two sequences.

Napier's kinematic model indicated how skilled mathematicians had become by the early 17th century in analyzing nonuniform motion. Kinematic ideas, which appeared frequently in mathematics of the period, provided a clear and visualizable means for the generation of geometric magnitude. The conception of a curve traced by a particle moving through space later played a significant role in the development of the calculus.

Napier's ideas were taken up and revised by the English mathematician Henry Briggs, the first Savilian professor of geometry at Oxford. In 1624 Briggs published an extensive table of common logarithms, or logarithms to the base 10. Because the base was no longer close to 1, the table could not be obtained as simply as Napier's, and Briggs therefore devised techniques involving the calculus of finite differences to facilitate calculation of the entries. He also devised interpolation procedures of great computational efficiency to obtain intermediate values.

In Switzerland the instrument maker Joost Bürgi arrived at the idea for logarithms independently of Napier, although he did not publish his results until 1620. Four years later a table of logarithms prepared by Johannes Kepler appeared in Marburg. Both Bürgi and Kepler were astronomical observers, and Kepler included logarithmic tables in his famous Rudolphine Tables, astronomical tabulations of planetary motion derived using the assumption of elliptical orbits about the Sun.

Analytic geometry. The invention of analytic geometry was, next to the differential and integral calculus, the most important mathematical development of the 17th century. Originating in the work of the French mathematicians Viète, Fermat, and Descartes, it had by the middle of the century established itself as a major program of mathematical research.

Two tendencies in contemporary mathematics stimulated the rise of analytic geometry. The first was an increased interest in curves, resulting in part from the recovery and Latin translation of the classical treatises of Apollonius, Archimedes, and Pappus, and in part from the increasing importance of curves in such applied fields as astronomy, mechanics, optics, and stereometry. The second was the emergence a century earlier of an established algebraic practice in the work of the Italian and German algebraists and its subsequent shaping into a powerful mathematical tool by Viète at the end of the century.

Influence of Greek geometers

Viète was a prominent representative of the humanist movement in mathematics that set itself the project of restoring and furthering the achievements of the classical Greek geometers. In his *In artem analyticem isagoge* ("Introduction to the Analytic Arts"; 1591) Viète, as part

of his program of rediscovering the method of analysis used by the ancient Greek mathematicians, proposed new algebraic methods that employed variables, constants, and equations, but he saw this as an advancement over the ancient method, a view he arrived at by comparing the geometric analysis contained in Book VII of Pappus' *Collection* with the arithmetic analysis of Diophantus' *Arithmetica*. Pappus had employed an analytic method for the discovery of theorems and the construction of problems; in analysis, by contrast to synthesis, one proceeds from what is sought until one arrives at something known. In approaching an arithmetic problem by laying down an equation among known and unknown magnitudes and then solving for the unknown, one was, Viète reasoned, following an "analytic" procedure.

Viète's notation

Viète introduced the concept of algebraic variable, which he denoted using a capital vowel (*A, E, I, O, U*) as well as the concept of parameter (an unspecified constant quantity), denoted by a capital consonant (*B, C, D,* and so on). In his system the equation $5BA^2 - 2CA + A^3 = D$ would appear as *B*5 in *A* quad − *C* plano 2 in *A* + *A* cub aequatur *D* solido.

Viète retained the classical principle of homogeneity, according to which terms added together must all be of the same dimension. In the above equation, for example, each of the terms has the dimension of a solid or cube; thus the constant *C,* which denotes a plane, is combined with *A* to form a quantity having the dimension of a solid.

It should be noted that in Viète's scheme the symbol *A* is part of the expression for the object obtained by operating on the magnitude denoted by *A*. Thus operations on the quantities denoted by the variables are reflected in the algebraic notation itself. This innovation, considered by historians of mathematics to be a major conceptual advance in algebra, facilitated the study of the symbolic solution of algebraic equations and led to the creation of the first conscious theory of equations.

After Viète's death the analytic art was applied to the study of curves by his countrymen Fermat and Descartes. Both men were motivated by the same goal, to apply the new algebraic techniques to Apollonius' theory of loci as preserved in Pappus' *Collection.* The most celebrated of these problems consisted of finding the curve or locus traced by a point whose distances from several fixed lines satisfied a given relation.

Fermat adopted Viète's notation in his paper of 1636, "Ad Locos Planos et Solidos Isagoge" ("Introduction to Plane and Solid Loci"). The title of the paper refers to the ancient classification of curves as plane (straight lines, circles), solid (ellipses, parabolas, and hyperbolas), or linear (curves defined kinematically or by a locus condition). Fermat considered an equation among two variables. One of the variables represented a line measured horizontally from a given initial point, while the other represented a second line positioned at the end of the first line and inclined at a fixed angle to the horizontal. As the first variable varied in magnitude, the second took on a value determined by the equation, and the endpoint of the second line traced out a curve in space. By means of this construction Fermat was able to formulate the fundamental principle of analytic geometry:

> Whenever two unknown quantities are found in final equality, there results a locus fixed in place, and the endpoint of one of these unknown quantities describes a straight line or a curve.

The principle implied a correspondence between two different classes of mathematical objects, geometric curves and algebraic equations. In the paper of 1636 Fermat showed that, if the equation is a quadratic, then the curve is a conic section, that is, an ellipse, parabola, or hyperbola. He also showed that the determination of the curve given by an equation is simplified by a transformation involving a change of variables to an equation in standard form.

La Géométrie

Descartes's *La Géométrie* appeared in 1637 as an appendix to his famous *Discours de la méthode,* the treatise that presented the foundation of his philosophical system. Although supposedly an example from mathematics of his rational method, *La Géométrie* was a technical treatise understandable independently of philosophy. It was destined to become one of the most influential books in the history of mathematics.

In the opening sections of *La Géométrie* Descartes introduced two innovations. In place of Viète's notation he initiated the modern practice of denoting variables by letters at the end of the alphabet (*x, y, z*) and parameters by letters at the beginning of the alphabet (*a, b, c*) and of using exponential notation to indicate powers of *x* (x^2, x^3, . . .). More significantly conceptually, he set aside Viète's principle of homogeneity, showing by means of a simple construction how to represent multiplication and division of lines by lines; thus all magnitudes (lines, areas, and volumes) could be represented independently of their dimension in the same way.

Descartes's goal in *La Géométrie* was to achieve the construction of solutions to geometric problems by means of instruments that were acceptable generalizations of ruler and compass. Algebra was a tool to be used in this program:

> If, then, we wish to solve any problem, we first suppose the solution already effected, and give names to all the lines that seem necessary for its construction—to those that are unknown as well as to those that are known. Then, making no distinction in any way between known and unknown lines, we must unravel the difficulty in any way that shows most naturally the relations between these lines, until we find it possible to express a single quantity in two ways. This will constitute an equation, since the terms of one of these two expressions are together equal to the terms of the other.

In the problem of Apollonius, for example, one sought to find the locus of points whose distances from a collection of fixed lines satisfied a given relation. One used this relation to derive an equation and then obtained points on the curve given by the roots of the equation using a geometric procedure involving acceptable instruments of construction.

Descartes described instruments more general than the compass for drawing "geometric" curves. He stipulated that the parts of the instrument be linked together so that the ratio of the motions of the parts could be knowable. This restriction excluded "mechanical" curves generated by kinematic processes. The Archimedean spiral, for example, was generated by a point moving on a line as the line rotated uniformly about the origin. The ratio of the circumference to the diameter did not permit exact determination:

> the ratios between straight and curved lines are not known, and I even believe cannot be discovered by men, and therefore no conclusion based upon such ratios can be accepted as rigorous and exact.

Geometric curves

Descartes concluded that a geometric or nonmechanical curve was one whose equation $f(x, y) = 0$ was a polynomial of finite degree in two variables. He wished to restrict mathematics to the consideration of such curves.

Descartes's emphasis on construction reflected his classical orientation. His conservatism with respect to what curves were acceptable in mathematics further distinguished him as a traditional thinker. At the time of his death in 1650, he had been overtaken by events, as research moved away from questions of construction to problems of finding areas (then called problems of quadrature) and tangents. The geometric objects that were then of growing interest were precisely the mechanical curves that Descartes had wished to banish from mathematics.

Following the important results achieved in the 16th century by Cardano and the Italian algebraists, the theory of algebraic equations reached an impasse. The ideas needed to investigate equations of degree higher than four were slow to develop. The immediate historical influence of Viète, Fermat, and Descartes was to furnish algebraic methods for the investigation of curves. A vigorous school of research became established in Leiden around Frans van Schooten, a Dutch mathematician who edited and published in 1649 a Latin translation of *La Géométrie.* Van Schooten published a second two-volume translation of the same work in 1659–1661 that also contained mathematical appendices by three of his disciples, Johan de Witt, Johan Hudde, and Hendrick van Heuraet. The Leiden group of mathematicians, which also included

Christiaan Huygens, was in large part responsible for the rapid development of Cartesian geometry in the middle of the century.

The calculus. The historian Carl Boyer has called the calculus "the most effective instrument for scientific investigation that mathematics has ever produced." As the mathematics of variability and change, the calculus was the characteristic product of the scientific revolution. The subject was properly the invention of two mathematicians, the German Gottfried Wilhelm Leibniz and the Englishman Isaac Newton. Both men published their researches in the 1680s, Leibniz in 1684 in the recently founded journal *Acta Eruditorum* and Newton in 1687 in his great treatise *Principia Mathematica.* Although a bitter dispute over priority developed later among followers of the two men, it is now clear that they each arrived at the calculus independently.

The calculus developed from techniques to solve two types of problems, the determination of areas and volumes and the calculation of tangents to curves. In classical geometry Archimedes had advanced furthest in this part of mathematics, having used the method of exhaustion to establish rigorously various results on areas and volumes and having derived for some curves (*e.g.,* the spiral) significant results concerning tangents. In the early 17th century there was a sharp revival of interest in both classes of problems. The decades between 1610 and 1670, referred to in the history of mathematics as "the precalculus period," were at a time of remarkable activity in which researchers throughout Europe contributed novel solutions and competed with each other to arrive at important new methods.

The precalculus period. In his treatise *Geometria Indivisibilibus Continuorum* (1635) (*Geometry by Indivisibles of Continuum*) Cavalieri, a professor of mathematics at the University of Bologna, formulated a systematic method for the determination of areas and volumes. As had Archimedes, Cavalieri regarded a plane figure as being composed of a collection of indivisible lines, "all the lines" of the plane figure. The collection was generated by a fixed line moving through space parallel to itself. Cavalieri showed that these collections could be interpreted as magnitudes obeying the rules of Euclidean ratio theory. In proposition 4 of Book II, he derived the result that is written today as

Cavalieri

$$\int_0^1 x^2 dx = \frac{1}{3}:$$

> Let there be given a parallelogram in which a diagonal is drawn; then "all the squares" of the parallelogram will be triple "all the squares" of each of the triangles determined by the diagonal.

Cavalieri showed that this proposition could be interpreted in different ways, as asserting that the volume of a cone is one-third the volume of the circumscribed cylinder or that the area under a segment of a parabola is one-third the area of the associated rectangle. In a later treatise he generalized the result by proving

$$\int_0^1 x^n dx = \frac{1}{(n+1)}$$

for $n=3$ to $n=9$. To establish these results he introduced transformations among the variables of the problem, using a result equivalent to the binomial theorem for integral exponents. The ideas involved went beyond anything that had appeared in the classical Archimedean theory of content.

Although Cavalieri was successful in formulating a systematic method based on general concepts, his ideas were not easy to apply. The derivation of very simple results required intricate geometric considerations, and the turgid style of the *Geometria Indivisibilibus* was a barrier to its reception.

John Wallis presented a quite different approach to the theory of quadratures in his *Arithmetica Infinitorum* (1655) (*Arithmetic of Infinities).* Wallis, a successor to Henry Briggs as the Savilian professor of geometry at Oxford, was a champion of the new methods of arithmetic algebra that he had learned from his teacher William Oughtred. Wallis expressed the area under a curve as the sum of an infinite series and used clever and unrigorous inductions to determine its value. To calculate the area under the parabola,

$$\int_0^1 x^2 dx,$$

he considered the successive sums

$$\frac{0+1}{1+1} = \frac{1}{3} + \frac{1}{6}, \quad \frac{0+1+4}{4+4+4} = \frac{1}{3} + \frac{1}{12}, \quad \frac{0+1+4+9}{9+9+9+9} = \frac{1}{3} + \frac{1}{18}$$

and inferred by "induction" the general relation

$$\frac{0^2 + 1^2 + 2^2 \ldots + n^2}{n^2 + n^2 + n^2 \ldots + n^2} = \frac{1}{3} + \frac{1}{6n}.$$

By letting the number of terms be infinite, he obtained $1/3$ as the limiting value of the expression. With more complicated curves he achieved very impressive results, including the infinite expression now known as Wallis' product:

$$\frac{4}{\pi} = \frac{3}{2} \cdot \frac{3}{4} \cdot \frac{5}{4} \cdot \frac{5}{6} \cdot \frac{7}{6} \cdot \ldots .$$

Research on the determination of tangents, the other subject leading to the calculus, proceeded along different lines. In *La Géométrie* Descartes had presented a method that could in principle be applied to any algebraic or "geometric" curve—*i.e.,* any curve whose equation was a polynomial of finite degree in two variables. The method depended upon finding the normal, the line perpendicular to the tangent, using the algebraic condition that it be the unique radius to intersect the curve in only one point. Descartes's method was simplified by Hudde, a member of the Leiden group of mathematicians, and was published in 1659 in van Schooten's edition of *La Géométrie.*

A class of curves of growing interest in the 17th century were those generated kinematically by a point moving through space. The famous cycloidal curve, for example, was traced by a point on the perimeter of a wheel that rolled on a line without slipping or sliding. These curves were nonalgebraic and hence could not be treated by Descartes's method. Gilles de Roberval, professor at the Collège Royale in Paris, devised a method borrowed from dynamics to determine their tangents. In his analysis of projectile motion Galileo had shown that the instantaneous velocity of a particle is compounded of two separate motions: a constant horizontal motion and an increasing vertical motion due to gravity. If the motion of the generating point of a kinematic curve is likewise regarded as the sum of two velocities, then the tangent will lie in the direction of their sum. Roberval applied this idea to several different kinematic curves, obtaining results that were often ingenious and elegant.

Non-algebraic curves

In an essay of 1636 circulated among French mathematicians, Fermat presented a method of tangents adapted from a procedure he had devised to determine maxima and minima and used it to find tangents to several algebraic curves of the form $y = x^n$. His account was short and contained no explanation of the mathematical basis of the new method. It is possible to see in his procedure an argument involving infinitesimals, and Fermat has sometimes been proclaimed as the discoverer of the differential calculus. Modern historical study, however, suggests that he was working with concepts introduced by Viète and that his method was based on finite algebraic ideas.

Isaac Barrow, the Lucasian professor of mathematics at Cambridge, published in 1670 his *Lectiones Geometricae* (*Geometrical Lectures*), a treatise that more than any other anticipated the unifying ideas of the calculus. In it he adopted a purely geometric form of exposition to show how the determinations of areas and tangents are inverse problems. He began with a curve and considered the slope of its tangent corresponding to each value of the abscissa. He then defined an auxiliary curve by the condition that its ordinate be equal to this slope and showed that the area under the auxiliary curve corresponding to a given

abscissa is equal to the rectangle whose sides are unity and the ordinate of the original curve. When reformulated analytically, this result expresses the inverse character of differentiation and integration, the fundamental theorem of the calculus. Although Barrow's decision to proceed geometrically prevented him from taking the final step to a true calculus, his lectures influenced both Newton and Leibniz.

Newton and Leibniz. The essential insight of Newton and Leibniz was to use Cartesian algebra to synthesize the earlier results and to develop algorithms that could be applied uniformly to a wide class of problems. The formative period of Newton's researches was from 1665 to 1670, while Leibniz worked a few years later, in the 1670s. Their contributions differ in origin, development, and influence, and it is necessary to consider each man separately.

The son of an English farmer, Newton became in 1669 the Lucasian professor of mathematics at Cambridge University. Newton's earliest researches in mathematics grew in 1665 from his study of van Schooten's edition of *La Géométrie* and Wallis' *Arithmetica Infinitorum.* Using the Cartesian equation of the curve, he reformulated Wallis' results, introducing for this purpose infinite sums in the powers of an unknown x, now known as infinite series. Possibly under the influence of Barrow, he used infinitesimals to establish for various curves the inverse relationship of tangents and areas. The operations of differentiation and integration emerged in his work as analytic processes that could be applied generally to investigate curves.

Fluents and fluxions

Unusually sensitive to questions of rigour, Newton at a fairly early stage tried to establish his new method on a sound foundation using ideas from kinematics. A variable was regarded as a "fluent," a magnitude that flows with time; its derivative or rate of change with respect to time was called a "fluxion," denoted by the given variable with a dot above it. The basic problem of the calculus was to investigate relations among fluents and their fluxions. Newton finished a treatise on the method of fluxions as early as 1671, although it was not published until 1736. In the 18th century this method became the preferred approach to the calculus among British mathematicians, especially after the appearance in 1742 of Colin Maclaurin's influential *Treatise of Fluxions.*

Newton first published the calculus in Book I of his great *Philosophiae Naturalis Principia Mathematica* (1687) (*Mathematical Principles of Natural Philosophy*). Originating as a treatise on the dynamics of particles, the *Principia* presented an inertial physics that combined Galileo's mechanics and Kepler's planetary astronomy. It was written in the early 1680s at a time when Newton was reacting against Descartes's science and mathematics. Setting aside the analytic method of fluxions, Newton introduced in 11 introductory lemmas his calculus of first and last ratios, a geometric theory of limits that provided the mathematical basis of his dynamics.

Newton's use of the calculus in the *Principia* is illustrated by proposition 11 of Book I: If the orbit of a particle moving under a centripetal force is an ellipse with the centre of force at one focus, then the force is inversely proportional to the square of the distance from the centre. Because the planets were known by Kepler's laws to move in ellipses with the Sun at one focus, this result supported his inverse square law of gravitation. To establish the proposition, Newton derived an approximate measure for the force by using small lines defined in terms of the radius (the line from the force centre to the particle) and the tangent to the curve at a point. This result expressed geometrically the proportionality of force to vector acceleration. Using properties of the ellipse known from classical geometry, Newton calculated the limit of this measure and showed that it was equal to a constant times one over the square of the radius.

Newton avoided analytical processes in the *Principia* by expressing magnitudes and ratios directly in terms of geometric quantities, both finite and infinitesimal. His decision to eschew analysis constituted a striking rejection of the algebraic methods that had been important in his own early researches on the calculus. Although the *Principia* was of inestimable value for later mechanics, it would be reworked by researchers on the Continent and expressed in the mathematical idiom of the Leibnizian calculus.

Leibniz's interest in mathematics was aroused in 1672 during a visit to Paris, where the Dutch mathematician Christiaan Huygens introduced him to his work on the theory of curves. Under Huygens' tutelage Leibniz immersed himself for the next several years in the study of mathematics. He investigated relationships among the summing and differencing of finite and infinite sequences of numbers. Having read Barrow's geometric lectures, he devised a transformation rule to calculate quadratures, obtaining the famous infinite series for $\pi/4$:

$$\frac{\pi}{4} = \frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots .$$

Notation

Leibniz was interested in questions of logic and notation, of how to construct a *characteristica universalis* for rational investigation. After considerable experimentation he arrived by the late 1670s at an algorithm based on the symbols d and $\int$. He first published his research on differential calculus in 1684 in an article in the *Acta Eruditorum,* "Nova Methodus pro Maximis et Minimis, Itemque Tangentibus, qua nec Fractas nec Irrationales Quantitates Moratur, et Singulare pro illi Calculi Genus" ("A New Method for Maxima and Minima as Well as Tangents, Which Is Impeded Neither by Fractional nor by Irrational Quantities, and a Remarkable Type of Calculus for This"). In this article he introduced the differential dx satisfying the rules $d(x+y) = dx + dy$ and $d(xy) = xdy + ydx$ and illustrated his calculus with a few examples. Two years later he published a second article, "On a Deeply Hidden Geometry," in which he introduced and explained the symbol $\int$ for integration. He stressed the power of his calculus to investigate transcendental curves, the very class of "mechanical" objects Descartes had believed lay beyond the power of analysis, and derived a simple analytical formula for the cycloid.

Leibniz continued to publish results on the new calculus in the *Acta Eruditorum* and began to explore his ideas in extensive correspondence with other scholars. Within a few years he had attracted a group of researchers to promulgate his methods, including the brothers Johann and Jakob Bernoulli in Basel and the priest Pierre Varignon and Guillaume-François-Antoine de L'Hospital in Paris. In 1700 he convinced Frederick III of Prussia to establish the Brandenburg Society of Sciences (later renamed the Berlin Academy of Sciences), with himself appointed president for life.

Leibniz's vigorous espousal of the new calculus, the didactic spirit of his writings, and his ability to attract a community of researchers contributed to his enormous influence on subsequent mathematics. In contrast, Newton's slowness to publish and his personal reticence resulted in a reduced presence within European mathematics. Although the British school in the 18th century included capable researchers, Abraham De Moivre, James Stirling, Brook Taylor, and Maclaurin among them, they failed to establish a program of research comparable to that established by Leibniz's followers on the Continent. There is a certain tragedy in Newton's isolation and his reluctance to acknowledge the superiority of continental analysis. As the historian Michael Mahoney observed:

> Whatever the revolutionary influence of the *Principia,* mathematics would have looked much the same if Newton had never existed. In that endeavour he belonged to a community, and he was far from indispensable to it.

THE 18TH CENTURY

Institutional background. After 1700 a movement to found learned societies on the model of Paris and London spread throughout Europe and the American colonies. The academy was the predominant institution of science until it was displaced by the university in the 19th century. The leading mathematicians of the period, such as Leonhard Euler, Jean Le Rond d'Alembert, and Joseph-Louis Lagrange, pursued academic careers at St. Petersburg, Paris, and London.

The Paris Academy of Sciences

The Paris Academy of Sciences provides an informative study of the 18th-century learned society. The academy was divided into six sections, three for the mathematical and three for the physical sciences. The mathematical sections were for geometry, astronomy, and mechanics, the physical sections for chemistry, anatomy, and botany. Membership in the academy was divided by section, with each section contributing three pensionnaires, two associates, and two adjuncts. There was also a group of free associates, distinguished men of science from the provinces, and foreign associates, eminent international figures in the field. A larger group of 70 corresponding members had partial privileges, including the right to communicate reports to the academy. The administrative core consisted of a permanent secretary, treasurer, president, and vice president. In a given year the average total membership in the academy was 153.

Prominent characteristics of the academy included its small and elite membership, made up heavily of men from the middle class, and its emphasis on the mathematical sciences. In addition to holding regular meetings and publishing memoirs, the academy organized scientific expeditions and administered prize competitions on important mathematical and scientific questions.

The historian Roger Hahn has noted that the academy in the 18th century allowed "the coupling of relative doctrinal freedom on scientific questions with rigorous evaluations by peers," an important characteristic of modern professional science. Academic mathematics and science did, however, foster a stronger individualistic ethos than is usual today. A determined individual such as Euler or Lagrange could emphasize a given program of research through his own work, the publications of the academy, and the setting of the prize competitions. The academy as an institution may have been more conducive to the solitary patterns of research in a theoretical subject like mathematics than it was to the experimental sciences. The separation of research from teaching is perhaps the most striking characteristic that distinguished the academy from the model of university-based science which developed in the 19th century.

Analysis and mechanics. The scientific revolution had bequeathed to mathematics a major program of research in analysis and mechanics. The period from 1700 to 1800, "the century of analysis," witnessed the consolidation of the calculus and its extensive application to mechanics. With expansion came specialization, as different parts of the subject acquired their own identity: ordinary and partial differential equations, calculus of variations, infinite series, and differential geometry. The applications of analysis were also varied, including the theory of the vibrating string, particle dynamics, the theory of rigid bodies, the mechanics of flexible and elastic media, and the theory of compressible and incompressible fluids. Analysis and mechanics developed in close association, with problems in one giving rise to concepts and techniques in the other, and all the leading mathematicians of the period made important contributions to mechanics.

Applications of analysis

The close relationship between mathematics and mechanics in the 18th century had roots extending deep into Enlightenment thought. In the organizational chart of knowledge at the beginning of the preliminary discourse to the *Encyclopédie,* d'Alembert distinguished between "pure" mathematics (geometry, arithmetic, algebra, calculus) and "mixed" mathematics (mechanics, geometric astronomy, optics, art of conjecturing). Mathematics generally was classified as a "science of nature" and separated from logic, a "science of man." The modern disciplinary division between physics and mathematics and the association of the latter with logic had not yet developed.

Mathematical mechanics itself as it was practiced in the 18th century differed in important respects from later physics. The goal of modern physics is to explore the ultimate particulate structure of matter and to arrive at fundamental laws of nature to explain physical phenomena. The character of applied investigation in the 18th century was rather different. The material parts of a given system and their interrelationship were idealized for the purposes of analysis. A material object could be treated as a point-mass (a mathematical point at which it is assumed all the mass of the object is concentrated), as a rigid body, as a continuously deformable medium, and so on. The intent was to obtain a mathematical description of the macroscopic behaviour of the system rather than to ascertain the ultimate physical basis of the phenomena. In this respect the 18th-century viewpoint is closer to modern mathematical engineering than it is to physics.

Mathematical research in the 18th century was coordinated by the Paris, Berlin, and St. Petersburg academies, as well as by several smaller provincial scientific academies and societies. Although England and Scotland were important centres early in the century, with Maclaurin's death in 1746 the British flame was all but extinguished.

History of analysis. The history of analysis in the 18th century can be followed in the official memoirs of the academies and in independently published expository treatises. In the first decades of the century the calculus was cultivated in an atmosphere of intellectual excitement, as mathematicians applied the new methods to a range of problems in the geometry of curves. The brothers Johann and Jakob Bernoulli showed that the shape of a smooth wire along which a particle descends in the least time is the cycloid, a transcendental curve much studied in the previous century. Working in a spirit of keen rivalry, the two brothers arrived at ideas that would later develop into the calculus of variations. In his study of the rectification of the lemniscate, a ribbon-shaped curve discovered by Jakob Bernoulli in 1694, Giulio Carlo Fagnano (1682–1766) introduced ingenious analytic transformations that laid the foundation for the theory of elliptic integrals. Nikolaus I Bernoulli (1687–1759), the nephew of Johann and Jakob, proved the equality of mixed second-order partial derivatives and made important contributions to differential equations by the construction of orthogonal trajectories to families of curves. Pierre Varignon (1654–1722), Johann Bernoulli, and Jakob Hermann (1678–1733) continued to develop analytic dynamics as they adapted Leibniz's calculus to the inertial mechanics of Newton's *Principia.*

Geometric conceptions and problems predominated in the early calculus. This emphasis on the curve as the object of study provided coherence to what was otherwise a disparate collection of analytic techniques. With its continued development, the calculus gradually became removed from its origins in the geometry of curves, and a movement emerged to establish the subject on a purely analytic basis. In a series of textbooks published in the middle of the century, the Swiss mathematician Leonhard Euler systematically accomplished the separation of the calculus from geometry. In his *Introductio in Analysin Infinitorum* (1748; *Introduction to the Analysis of Infinities*) he made the notion of function the central organizing concept of analysis:

Foundations of the calculus

> A function of a variable quantity is an analytical expression composed in any way from the variable and from numbers or constant quantities.

Euler's analytic approach is illustrated by his introduction of the sine and cosine functions. Trigonometric tables had existed since antiquity, and the relations between sines and cosines were commonly used in mathematical astronomy. In the early calculus mathematicians had derived in their study of periodic mechanical phenomena the differential equation

$$\frac{dy}{dx} = \frac{1}{\sqrt{1-x^2}}$$

and they were able to interpret its solution geometrically in terms of lines and angles in the circle. Euler was the first to introduce the sine and cosine functions as quantities whose relation to other quantities could be studied independently of any geometric diagram.

Euler's analytic approach to the calculus received support from his younger contemporary Joseph-Louis Lagrange, who, following Euler's death in 1783, replaced him as the leader of European mathematics. In 1755 the 19-year-old Lagrange wrote to Euler to announce the discovery of a new algorithm in the calculus of variations, a subject to

which Euler had devoted an important treatise 11 years earlier. Euler had used geometric ideas extensively and had emphasized the need for analytic methods. Lagrange's idea was to introduce the new symbol δ into the calculus and to experiment formally until he had devised an algorithm to obtain the variational equations. Mathematically quite distinct from Euler's procedure, his method required no reference to the geometric configuration. Euler immediately adopted Lagrange's idea, and in the next several years the two men systematically revised the subject using the new techniques.

In 1766 Lagrange was invited by the Prussian king, Frederick II the Great, to become mathematics director of the Berlin Academy. During the next two decades he wrote important memoirs on nearly all of the major areas of mathematics. In 1788 he published his famous *Mécanique analytique,* a treatise that used variational ideas to present mechanics from a unified analytic viewpoint. In the preface Lagrange wrote:

> One will find no Figures in this Work. The methods that I present require neither constructions nor geometrical or mechanical reasonings, but only algebraic operations, subject to a regular and uniform course. Those who admire Analysis, will with pleasure see Mechanics become a new branch of it, and will be grateful to me for having extended its domain.

Following the death of Frederick the Great, Lagrange traveled to Paris to become a pensionnaire of the Academy of Sciences. With the establishment of the École Polytechnique (French: "Polytechnic School") in 1794 he was asked to deliver the lectures on mathematics. There was a concern in European mathematics at the time to place the calculus on a sound basis, and Lagrange used the occasion to develop his ideas for an algebraic foundation of the subject. The lectures were published in 1797 under the title *Théorie des fonctions analytiques* ("Theory of Analytical Functions"), a treatise whose contents were summarized in its longer title "Containing the Principles of the Differential Calculus Disengaged from All Consideration of Infinitesimals, Vanishing Limits or Fluxions and Reduced to the Algebraic Analysis of Finite Quantities." Lagrange published a second treatise on the subject in 1801, a work that appeared in a revised and expanded form in 1806.

The theory of functions of a real variable

The range of subjects presented and the consistency of style distinguished Lagrange's didactic writings from other contemporary expositions of the calculus. He began with Euler's notion of a function as an analytic expression composed of variables and constants. He defined the "derived function" or derivative $f'(x)$ of $f(x)$ to be the coefficient of i in the Taylor expansion of $f(x + i)$. Assuming the general possibility of such expansions, he attempted a rather complete theory of the differential and integral calculus, including extensive applications to geometry and mechanics. Lagrange's lectures represented the most advanced development of the 18th-century analytic conception of the calculus.

Beginning with Augustin-Louis Cauchy in the 1820s, later mathematicians used the concept of limit to establish the calculus on an arithmetic basis. The algebraic viewpoint of Euler and Lagrange was rejected. To arrive at a proper historical appreciation of their work it is necessary to reflect on the meaning of analysis in the 18th century. Since Viète, analysis had referred generally to mathematical methods that employed equations, variables, and constants. With the extensive development of the calculus by Leibniz and his school, analysis became identified with all calculus-related subjects. In addition to this historical association, there was a deeper sense in which analytic methods were fundamental to the new mathematics. An analytic equation implied the existence of a relation that remained valid as the variables changed continuously in magnitude. Analytic algorithms and transformations presupposed a correspondence between local and global change, the basic concern of the calculus. It is this aspect of analysis that fascinated Euler and Lagrange and caused them to see in it the "true metaphysics" of the calculus.

OTHER DEVELOPMENTS

During the period 1600–1800 significant advances occurred in the theory of equations, foundations of Euclidean geometry, number theory, projective geometry, and probability theory. These subjects, which became mature branches of mathematics only in the 19th century, never rivaled analysis and mechanics as programs of research.

Theory of equations. After the dramatic successes of Niccolò Fontana Tartaglia and Lodovico Ferrari in the 16th century the theory of equations developed slowly, as problems resisted solution by known techniques. In the later 18th century the subject experienced an infusion of new ideas. Interest was concentrated on two problems. The first was to establish the existence of a root of the general polynomial equation of degree n. The second was to express the roots as algebraic functions of the coefficients, or to show why it was not in general possible to do so.

The proposition that the general polynomial with real coefficients has a root of the form $a + b\sqrt{-1}$ became known later as the fundamental theorem of algebra. By 1742 Euler recognized that roots appear in conjugate pairs; if $a + b\sqrt{-1}$ is a root, then so is $a - b\sqrt{-1}$. Thus, if $a + b\sqrt{-1}$ is a root of $f(x) = 0$, then $f(x) = (x^2 - 2ax - a^2 - b^2)g(x)$. The fundamental theorem was therefore equivalent to asserting that a polynomial may be decomposed into linear and quadratic factors. This result was of considerable importance for the theory of integration, since by the method of partial fractions it ensured that a rational function, the quotient of two polynomials, could always be integrated in terms of algebraic and elementary transcendental functions.

The fundamental theorem of algebra

Although d'Alembert, Euler, and Lagrange worked on the fundamental theorem, the first successful proof was developed by Carl Friedrich Gauss in his doctoral dissertation of 1799. Earlier researchers had investigated special cases or had concentrated on showing that all possible roots were of the form $a \pm b\sqrt{-1}$. Gauss tackled the problem of existence directly. Expressing the unknown in terms of the polar variables r and θ, he showed that a solution of the polynomial would lie at the intersection of two curves of the form $T(r, \theta) = 0$ and $U(r, \theta) = 0$. By a careful and rigorous investigation he proved that the two curves intersect.

Gauss's demonstration of the fundamental theorem initiated a new approach to the question of mathematical existence. In the 18th century mathematicians were interested in the nature of particular analytic processes or the form that given solutions should take. Mathematical entities were regarded as things that were given, not as things whose existence needed to be established. Because analysis was applied in geometry and mechanics, the formalism seemed to possess a clear interpretation that obviated any need to consider questions of existence. Gauss's demonstration was the beginning of a change of attitude in mathematics, of a shift to the rigorous, internal development of the subject.

The problem of expressing the roots of a polynomial as functions of the coefficients was addressed by several mathematicians independently around 1770. The Cambridge mathematician Edward Waring published treatises in 1762 and 1770 on the theory of equations. In 1770 Lagrange presented a long expository memoir on the subject to the Berlin Academy, and in 1771 Alexandre Vandermonde submitted a paper to the French Academy of Sciences. Although the ideas of the three men were related, Lagrange's memoir was the most extensive and most influential historically.

Lagrange presented a detailed analysis of the solution by radicals of second-, third-, and fourth-degree equations and investigated why these solutions failed when the degree was greater than or equal to five. He introduced the novel idea of considering functions of the roots and examining the values they assumed as the roots were permuted. He was able to show that the solution of an equation depends on the construction of a second resolvent equation, but he was unable to provide a general procedure for solving the resolvent when the degree of the original equation was greater than four. Although his theory left the subject in an unfinished condition, it provided a solid basis for future work. The search for a general solution to the polynomial equation would provide the greatest single impetus for the transformation of algebra in the 19th century.

The efforts of Lagrange, Vandermonde, and Waring illustrate how difficult it was to develop new concepts in algebra. The history of the theory of equations belies the view that mathematics is subject to an almost automatic technical development. Much of the later algebraic work would be devoted to devising terminology, concepts, and methods necessary to advance the subject.

Foundations of geometry. Although the emphasis of mathematics after 1650 was increasingly on analysis, foundational questions in classical geometry continued to arouse interest. Attention centred on the fifth postulate of Book I of the *Elements,* which Euclid had used to prove the existence of a unique parallel through a point to a given line. Since antiquity, Greek, Islāmic, and European geometers had attempted unsuccessfully to show that the parallel postulate need not be a postulate but could instead be deduced from the other postulates of Euclidean geometry. During the period 1600–1800 mathematicians continued these efforts by trying to show that the postulate was equivalent to some result that was considered self-evident. Although the decisive breakthrough to non-Euclidean geometry would not occur until the 19th century, researchers did achieve a deeper and more systematic understanding of the classical properties of space.

The parallel postulate

Interest in the parallel postulate developed in the 16th century after the recovery and Latin translation of Proclus' commentary on Euclid's *Elements.* The Italian researchers Christopher Clavius in 1574 and Giordano Vitale in 1680 showed that the postulate is equivalent to asserting that the line equidistant from a straight line is a straight line. In 1693 John Wallis, Savilian professor of geometry at Oxford, attempted a different demonstration, proving that the axiom follows from the assumption that to every figure there exists a similar figure of arbitrary magnitude.

In 1733 the Italian Girolamo Saccheri published his *Euclides ab Omni Naevo Vindicatus* (*Euclid Cleared of Every Flaw*). This was an important work of synthesis in which he provided a complete analysis of the problem of parallels in terms of Omar Khayyam's quadrilateral. Using the Euclidean assumption that straight lines do not enclose an area, he was able to exclude geometries that contain no parallels. It remained to prove the existence of a unique parallel through a point to a given line. To do this Saccheri adopted the procedure of reductio ad absurdum; he assumed the existence of more than one parallel and attempted to derive a contradiction. After a long and detailed investigation, he was able to convince himself (mistakenly) that he had found the desired contradiction.

In 1766 Johann Heinrich Lambert of the Berlin Academy composed *Die Theorie der Parallellinien* ("The Theory of Parallel Lines"; published 1786), a penetrating study of the fifth postulate in Euclidean geometry. Among other theorems Lambert proved is that the parallel axiom is equivalent to the assertion that the sum of the angles of a triangle is equal to two right angles. He combined this fact with Wallis' result to arrive at an unexpected characterization of classical space. According to Lambert, if the parallel postulate is rejected, it follows that for every angle θ less than $2R/3$ (R is a right angle) an equilateral triangle can be constructed with corner angle θ. By Wallis' result any triangle similar to this triangle must be congruent to it. It is therefore possible to associate with every angle a definite length, the side of the corresponding equilateral triangle. Since the measurement of angles is absolute, independent of any convention concerning the selection of units, it follows that an absolute unit of length exists. Hence, to accept the parallel postulate is to deny the possibility of an absolute concept of length.

The final 18th-century contribution to the theory of parallels was Adrien-Marie Legendre's textbook *Éléments de géométrie,* the first edition of which appeared in 1794. Legendre presented an elegant demonstration that purported to show that the sum of the angles of a triangle is equal to two right angles. He believed that he had conclusively established the validity of the parallel postulate. His work attracted a large audience and was influential in informing readers of the new ideas in geometry.

The 18th-century failure to develop a non-Euclidean geometry was rooted in deeply held philosophical beliefs. In his *Critique of Pure Reason* (1781), Immanuel Kant had emphasized the synthetic a priori character of mathematical judgments. From this standpoint, statements of geometry and arithmetic were necessarily true propositions with definite empirical content. The existence of similar figures of different size, or the conventional character of units of length, appeared self-evident to mathematicians of the period. As late as 1824 Simon Laplace wrote:

The influence of Kant

> Thus the notion of space includes a special property, self-evident, without which the properties of parallels cannot be rigorously established. The idea of a bounded region, *e.g.,* the circle, contains nothing which depends on its absolute magnitude. But if we imagine its radius to diminish, we are brought without fail to the diminution in the same ratio of its circumference and the sides of all the inscribed figures. This proportionality appears to me a more natural postulate than that of Euclid, and it is worthy of note that it is discovered afresh in the results of the theory of universal gravitation.
>
> (C.G.F.)

Mathematics in the 19th and 20th centuries

Most of the powerful abstract mathematical theories in use today originated in the 19th century, so that any historical account of the period should be supplemented by reference to detailed treatments of these topics. Moreover, mathematics grew so much during this period that any account must necessarily be selective. Nonetheless, some broad features stand out. The growth of mathematics as a profession was accompanied by a sharpening division between mathematics and the physical sciences, and contact between the two subjects takes place today across a clear professional boundary. One result of this separation has been that mathematics, no longer able to rely on its scientific import for its validity, developed markedly higher standards of rigour. It was also freed to develop in directions that had little to do with applicability. Some of these pure creations have turned out to be surprisingly applicable, while the attention to rigour has led to a wholly novel conception of the nature of mathematics and logic. Moreover, many outstanding questions in mathematics yielded to the more conceptual approaches that came into vogue.

Projective geometry. The French Revolution provoked a radical rethinking of education in France, and mathematics was given a prominent role. The École Polytechnique was established in 1794 with the ambitious task of preparing all candidates for the specialist civil and military engineering schools of the republic. Mathematicians of the highest calibre were involved; the result was a rapid and sustained development of the subject. The inspiration for the École was that of Gaspard Monge, who believed strongly that mathematics should serve the scientific and technical needs of the state. To that end he devised a syllabus that promoted his own descriptive geometry, which was useful in the design of forts, gun emplacements, and machines and which was employed to great effect in the Napoleonic survey of Egyptian historical sites.

In Monge's descriptive geometry, three-dimensional objects are described by their orthogonal projections onto a horizontal and a vertical plane, the plan and elevation of the object. A pupil of Monge, Jean-Victor Poncelet, was taken prisoner in Napoleon's retreat from Moscow and sought to keep up his spirits while in jail in Saratov by thinking over the geometry he had learned. He dispensed with the restriction to orthogonal projections and decided to investigate what properties figures have in common with their shadows.

Poncelet

There are several of these properties: a straight line casts a straight shadow, and a tangent to a curve casts a shadow that is tangent to the shadow of the curve. But some properties are lost: the lengths and angles of a figure bear no relation to the lengths and angles of its shadow. Poncelet felt that the properties that survive are worthy of study, and, by considering only those properties that a figure shares with all its shadows, Poncelet hoped to put truly geometric reasoning on a par with algebraic geometry.

In 1822 Poncelet published the *Traité des propriétés projectives des figures* (*Treatise on the Projective Properties of Figures*). From his standpoint, every conic section is

equivalent to a circle, so his treatise contained a unified treatment of the theory of conic sections. It also established several new results. Geometers who took up his work divided into two groups: those who accepted his terms and those who, finding them obscure, reformulated his ideas in the spirit of algebraic geometry. On the algebraic side it was taken up in Germany by August Ferdinand Möbius, who seems to have come to his ideas independently of Poncelet, and then by Julius Plücker. They showed how rich was the projective geometry of curves defined by algebraic equations and thereby gave an enormous boost to the algebraic study of curves, comparable to the original impetus provided by Descartes. Germany also produced synthetic projective geometers, notably Jakob Steiner (born in Switzerland but educated in Germany) and Karl George Christian von Staudt, who emphasized what can be understood about a figure from a careful consideration of all its transformations.

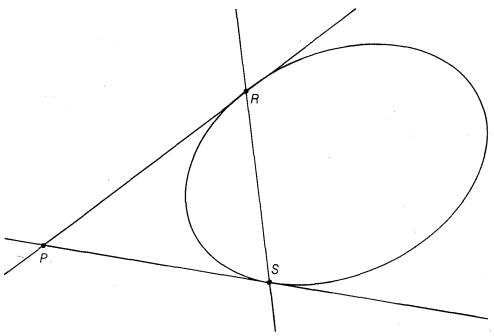


Figure 12: Duality associates with the point *P* the line *RS*, and vice versa.

Within the debates about projective geometry emerged one of the few synthetic ideas to be discovered since the days of Euclid, that of duality. This associates with each point a line and with each line a point, in such a way that (1) three points lying in a line give rise to three lines meeting in a point, and, conversely, three lines meeting in a point give rise to three points lying on a line, and (2) if one starts with a point (or a line) and passes to the associated line (point) and then repeats the process, one returns to the original point (line). One way of doing this (presented by Poncelet) is to pick an arbitrary conic and then to associate with a point P lying outside the conic the line that joins the points R and S at which the tangents through P to the conic touch the conic (see Figure 12). A second method is needed for points on or inside the conic. The feature of duality that makes it so exciting is that one can apply it mechanically to every proof in geometry, interchanging "point" and line" and "collinear" and "concurrent" throughout, and so obtain a new result. Sometimes a result turns out to be equivalent to the original, sometimes to its converse, but at a single stroke the number of theorems was more or less doubled.

Making the calculus rigorous. Monge's educational ideas were opposed by Lagrange, who favoured a more traditional and theoretical diet of advanced calculus and rational mechanics (the application of the calculus to the study of the motion of solids and liquids). Eventually Lagrange won, and the vision of mathematics that was presented to the world was that of an autonomous subject that was also applicable to a broad range of phenomena by virtue of its great generality, a view that has persisted to the present day.

During the 1820s Augustin-Louis Cauchy lectured at the École Polytechnique on the foundations of the calculus. Since its invention it had been generally agreed that the calculus gave correct answers, but no one had been able to give a satisfactory explanation of why this was so. Cauchy rejected Lagrange's algebraic approach and proved that Lagrange's basic assumption that every function had a power series expansion was in fact false. Newton had suggested a geometric or dynamic basis for calculus, but this ran the risk of introducing a vicious circle when the calculus was applied to mechanical or geometric problems. Cauchy proposed basing the calculus on a sophisticated and difficult interpretation of the idea of two points or numbers being arbitrarily close together. Although his students disliked the new approach, and Cauchy was ordered to teach material that the students could actually understand and use, his methods gradually became established and refined to form the core of the modern rigorous calculus, a subject now called mathematical analysis.

Mathematical analysis

Traditionally, the calculus had been concerned with the two processes of differentiation and integration and the reciprocal relation that exists between them. Cauchy provided a novel underpinning by stressing the importance of the concept of continuity, which is more basic than either. He showed that, once the concepts of a continuous function and limit are defined, the concepts of a differentiable function and an integrable function can be defined in terms of them. Unfortunately, neither of these concepts is easy to grasp, and the much-needed degree of precision they bring to mathematics has proved difficult to appreciate. Roughly speaking, a function is continuous at a point in its domain if small changes in the input around the specified value only produce small changes in the output.

Thus, in Figure 13 the familiar graph of a parabola $y = x^2$ is continuous around the point $x = 0$; as x varies by small amounts, so necessarily does y. On the other hand, the graph of the function that takes the value 0 when x is negative or zero, and the value 1 when x is positive, plainly has a discontinuous graph at the point $x = 0$, and it is indeed discontinuous there according to the definition. If x varies from 0 by any small positive amount, the value of the function jumps by the fixed amount 1, which is not an arbitrarily small amount.

Cauchy said that a function $f(x)$ tends to a limiting value 1 as x tends to the value a whenever the value of the difference $f(x) - f(a)$ becomes arbitrarily small as the difference $x - a$ itself becomes arbitrarily small. He then showed that, if $f(x)$ is continuous at a, the limiting value of the function as x tended to a was indeed $f(a)$. The crucial feature of this definition is that it defines what it means for a variable quantity to tend to something entirely without reference to ideas of motion.

Cauchy then said a function $f(x)$ is differentiable at the point a if, as x tends to a (which it is never allowed to reach), the value of the quotient $\{f(x) - f(a)\}/(x - a)$ tends to a limiting value, called the derivative of the function $f(x)$ at a. To define the integral of a function $f(x)$ between the values a and b, Cauchy went back to the primitive idea of integral as the measure of the area under the graph of the function. He approximated this area by rectangles and said that, if the sum of the areas of the rectangles tends to a limit as their number increases indefinitely and

Definition of the integral

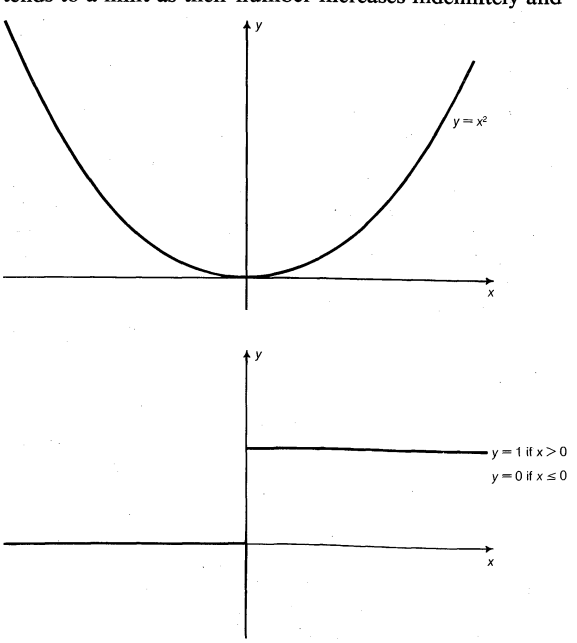


Figure 13: Continuous and discontinuous functions.

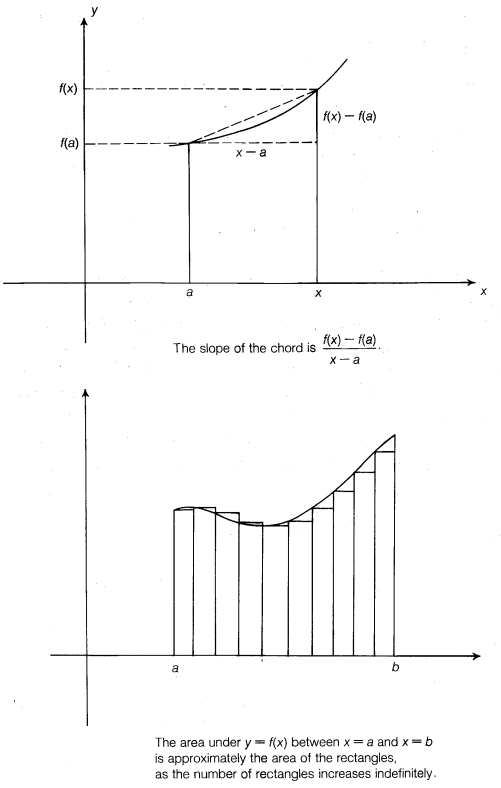


Figure 14: Differentiation and integration.

if this limiting value is the same however the rectangles are obtained, then the function is integrable. Its integral is the common limiting value. After he had defined the integral independently of the differential calculus, Cauchy had to prove that the processes of integrating and differentiating are mutually inverse. This he did, giving for the first time a rigorous foundation to all the elementary calculus of his day.

Fourier series. The other crucial figure of the time in France was Joseph Fourier. His major contribution, presented in *La Théorie analytique de la chaleur* (1822; *The Analytical Theory of Heat*), was to the theory of heat diffusion in solid bodies. He proposed that any function could be written as an infinite sum of the trigonometric functions cosine and sine, for example,

$$f(x) = a_0 + a_1 \sin x + a_2 \sin 2x + \ldots .$$

While expressions of this kind had been written down earlier, what was new in Fourier's treatment was the degree of attention given to their convergence. He investigated the question, "Given the function $f(x)$, for what range of values of x does the expression above sum to a finite number?" It turned out that the answer depends on the coefficients a_n, and Fourier gave rules for obtaining them of the form:

$$a_n = \int_{-\pi}^{\pi} f(x) \sin (nx) dx.$$

Had Fourier's work been entirely correct, it would have brought all functions into the calculus, making possible the solution of many kinds of differential equations and greatly extending the theory of mathematical physics. But his arguments were unduly naive: after Cauchy, it was not clear that the function $f(x) \sin (nx)$ was necessarily integrable. When Fourier's ideas were finally published, they were eagerly taken up, but the more cautious mathematicians, notably the influential German mathematician Peter Gustav Lejeune Dirichlet, wanted to rederive Fourier's conclusions in a more rigorous way. Fourier's methodology was widely accepted, but questions about its validity in detail were to occupy mathematicians for the rest of the century.

Elliptic functions. The theory of functions of a complex variable was also being decisively reformulated. At the start of the 19th century, complex numbers were discussed from a quasi-philosophical standpoint by several French writers, notably Jean Robert Argand. A consensus emerged that complex numbers should be thought of as pairs of real numbers, with suitable rules for their addition and multiplication so that the pair (0, 1) was a square root of −1. The underlying meaning of such a number pair was given by its geometric interpretation either as a point in a plane or as a directed segment joining the coordinate origin to the point in question. (This representation is sometimes called the Argand diagram.) In 1827, while revising an earlier manuscript for publication, Cauchy showed how the problem of integrating functions of two variables can be illuminated by a theory of functions of a single complex variable, which he was then developing. But the decisive influence on the growth of the subject came from the theory of elliptic functions.

The Argand diagram

The study of elliptic functions originated in the 18th century, when many authors studied integrals of the form

$$\int_0^x \frac{p(t)dt}{\sqrt{q(t)}},$$

where $p(t)$ and $q(t)$ are polynomials in t and $q(t)$ is of degree 3 or 4 in t. Such integrals arise naturally, for example, as an expression for the length of an arc of an ellipse (whence the name). These integrals cannot be evaluated explicitly; they do not define a function that can be obtained from the rational and trigonometric functions, a difficulty that added to their interest. Elliptic integrals were intensively studied for many years by the French mathematician Legendre, who was able to calculate tables of values for such expressions as functions of their upper endpoint, x. But the topic was completely transformed in the late 1820s by the independent but closely overlapping discoveries of two young mathematicians, the Norwegian Niels Henrik Abel and the German Carl Gustav Jacob Jacobi. These men showed that, if one allowed the variable x to be complex and the problem were inverted, so that the object of study became

$$u = \int_0^x \frac{p(t)dt}{\sqrt{q(t)}},$$

considered as defining a function x of a variable u, then a remarkable new theory could be discovered. The new function, for example, possessed a property that generalized the basic property of periodicity of the trigonometric functions sine and cosine: $\sin(x) = \sin(x + 2\pi)$. Any function of the kind just described has two distinct periods, ω_1 and ω_2:

$$x(u) = x(u + \omega_1) = x(u + \omega_2).$$

These new functions, the elliptic functions, aroused a considerable degree of interest from the first because many interesting and novel things could be said about them. The analogy with trigonometric functions ran very deep (indeed the trigonometric functions turned out to be special cases of elliptic functions), but their greatest influence was on the burgeoning general study of functions of a complex variable. The theory of elliptic functions became the paradigm of what could be discovered by allowing variables to be complex instead of real. But their natural generalization to functions defined by more complicated

Elliptic functions

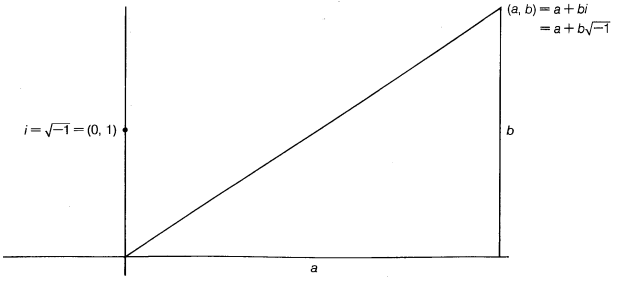


Figure 15: The plane of complex numbers (Argand diagram).

integrands, although it yielded partial results, resisted analysis until the second half of the 19th century.

The theory of numbers. While the theory of elliptic functions typifies the 19th century's enthusiasm for pure mathematics, some contemporary mathematicians said that the simultaneous developments in number theory carried that enthusiasm to excess. Nonetheless, during the 19th century the algebraic theory of numbers grew from being a minority interest to its present central importance in pure mathematics. The earlier investigations of Fermat had eventually drawn the attention of Euler and Lagrange. Euler proved some of Fermat's unproved claims and discovered many new and surprising facts; Lagrange not only supplied proofs of many remarks that Euler had merely conjectured but also worked them into something like a coherent theory. For example, it was known to Fermat that the numbers which can be written as the sum of two squares are either the number 2, squares themselves, primes of the form $4n + 1$, or else products of these numbers. Thus 29, which is $4 \times 7 + 1$, is $5^2 + 2^2$, but 35, which is not of this form, cannot be written as the sum of two squares. Euler had proved this result and had gone on to consider similar cases, such as primes of the form $x^2 + 2y^2$, or $x^2 + 3y^2$. But it was left to Lagrange to provide a general theory covering all expressions of the form $ax^2 + bxy + cy^2$, quadratic forms, as they are called.

Lagrange's theory of quadratic forms had made considerable use of the idea that a given quadratic form could often be simplified to another with the same properties but with smaller coefficients. To do this in practice, it was often necessary to consider whether a given integer left a remainder that was a square when it was divided by another given integer. (For example, 48 leaves a remainder of 4 upon division by 11, and 4 is a square.) Legendre discovered a remarkable connection between the question, "Does the integer p leave a square remainder on division by q?" and the seemingly unrelated question, "Does the integer q leave a square remainder upon division by p?" He saw, in fact, that, when p and q are primes, both questions have the same answer unless both primes are of the form $4n - 1$. Because this observation connects two questions in which the integers p and q play mutually opposite roles, it became known as the law of quadratic reciprocity. Legendre also gave an effective way of extending his law to cases when p and q are not prime.

All this work set the scene for the emergence of Gauss, whose *Disquisitiones Arithmeticae* not only consummated what had gone before but also directed number theorists in new and deeper directions. He rightly showed that Legendre's proof of the law of quadratic reciprocity was fundamentally flawed and gave the first rigorous proof. His work suggested that there were profound connections between the original question and other branches of number theory, a fact that he perceived to be of signal importance for the subject. He extended Lagrange's theory of quadratic forms by showing how two quadratic forms can be "multiplied" to obtain a third. Later mathematicians were to rework this into an important example of the theory of finite commutative groups. And in the long final section of his book Gauss gave the theory that lay behind his first discovery as a mathematician: that a regular 17-sided figure can be constructed by circle and straightedge alone.

The discovery that the regular "17-gon" is so constructible was the first such discovery since the Greeks, who had known only of the equilateral triangle, the square, the regular pentagon, the regular 15-sided figure, and the figures that can be obtained from these by successively bisecting all the sides. But what was of much greater significance than the discovery was the theory that underpinned it, the theory of what are now called algebraic numbers. It may be thought of as an analysis of how complicated a number may be while yet being amenable to an exact treatment.

The theory of algebraic numbers

The simplest numbers to understand and use are the integers and the rational numbers; the irrational numbers seem to pose problems. Famous among these is $\sqrt{2}$. It cannot be written as a finite or repeating decimal (because it is not rational), but it can be manipulated algebraically very easily. It is only necessary to replace every occurrence of $(\sqrt{2})^2$ by 2. In this way expressions of the form $m + n\sqrt{2}$, where m and n are integers, can be handled arithmetically. These expressions have many properties akin to those of whole numbers, and one can even define prime numbers of this form; therefore they are called algebraic integers. In this case they are obtained by grafting onto the rational numbers a solution of the polynomial equation $x^2 - 2 = 0$. In general an algebraic integer is any solution, real or complex, of a polynomial equation with integer coefficients in which the coefficient of the highest power of the unknown is 1.

Gauss's theory of algebraic integers led to the question of determining when a polynomial of degree n with integer coefficients can be solved given the solvability of polynomial equations of lower degree but with coefficients that are algebraic integers. For example, he regarded the coordinates of the 17 vertices of a regular 17-sided figure as complex numbers satisfying the equation $x^{17} - 1 = 0$, and thus as algebraic integers. One such integer is $z = 1$. He showed that the rest are obtained by solving a succession of four quadratic equations. Because solving a quadratic equation is equivalent to performing a construction with a ruler and compass, as Descartes had shown long before, Gauss had shown how to construct the regular 17-gon.

Inspired by Gauss's works on the theory of numbers, a growing school of mathematicians was drawn to the subject. Like Gauss, the German mathematician Ernst Eduard Kummer sought to generalize the law of quadratic reciprocity to deal with questions about third, fourth, and higher powers of numbers. He found that his work led him in an unexpected direction, toward a partial resolution of Fermat's last theorem. This result asserts that there are no solutions in positive integers to the equation $x^n + y^n = z^n$ if $n > 2$. It was claimed by Fermat when he was young, but he gave no proof, and only partial successes have been obtained since then.

Kummer's approach was to develop the theory of algebraic integers. If it could be shown that the equation had no solution in algebraic integers, then a fortiori there could be no solution in ordinary integers. He was eventually able to establish the truth of Fermat's last theorem for a large class of prime exponents n (those satisfying some technical conditions needed to make the proof work). This was the first significant breakthrough in the study of the theorem, and it remains unmatched to the present day. Together with the earlier work of the French mathematician Sophie Germain, it has enabled mathematicians to establish Fermat's last theorem for every value of n from 3 to 1,000,000. However, Kummer's way around the difficulties he encountered further propelled the theory of algebraic integers into the realm of abstraction. It amounted to the suggestion that there should be yet other types of integers, but many found these ideas obscure.

In Germany Richard Dedekind patiently created a new approach, in which each new number (called ideal) was defined by means of a suitable set of algebraic integers in such a way that it was the common divisor of the set of algebraic integers used to define it. Dedekind's work was slow to gain approval, yet it illustrates several of the most profound features of modern mathematics. It was clear to Dedekind that the ideal algebraic integers were the work of the human mind. Their existence can neither be based on nor deduced from the existence of physical objects, analogies with natural processes, or some process of abstraction from more familiar things. A second feature of Dedekind's work was its reliance on the idea of sets of objects, such as sets of numbers, even sets of sets. Dedekind's work showed how basic the naive conception of a set could be. The third crucial feature of his work was its emphasis on the structural aspects of algebra. The presentation of number theory as a theory about objects that can be manipulated (in this case, added and multiplied) according to certain rules akin to those governing ordinary numbers was to be a paradigm of the more formal theories of the 20th century.

The ideal algebraic integers

The theory of equations. Another subject that was transformed in the 19th century was the theory of equations. Ever since Tartaglia and Ferrari in the 16th century had found rules giving the solutions of cubic and quartic

equations in terms of the coefficients of the equations, formulas had unsuccessfully been sought for equations of the fifth and higher degrees. At stake was the existence of a formula that expresses the roots of a quintic equation in terms of the coefficients. This formula moreover, must involve only the operations of addition, subtraction, multiplication, and division, together with the extraction of roots, since that was all that had been required for the solution of quadratic, cubic, and quartic equations. If such a formula were to exist, the quintic would accordingly be said to be solvable by radicals.

In 1770 Lagrange had analyzed all the successful methods he knew for equations of degrees 2, 3, and 4, in an attempt to see why they worked and how they could be generalized. His analysis of the problem in terms of permutations of the roots was promising, but he became more and more doubtful as the years went by that his complicated line of attack could be carried through. The first valid proof that the general quintic is not solvable by radicals was offered only after his death, in a startlingly short paper by Abel, written in 1824.

Abel also showed by example that some quintic equations were solvable by radicals and that some equations could be solved unexpectedly easily. For example, the equation $x^5 - 1 = 0$ has one root $x = 1$, but the remaining four roots can be found just by extracting square roots, not fourth roots as might be expected. He therefore raised the question, "What equations of degree higher than 4 are solvable by radicals?"

Abel died in 1829 at the age of 26 and did not resolve the problem he had posed. Almost at once, however, the astonishing prodigy Évariste Galois burst upon the Parisian mathematical scene. He submitted an account of his novel theory of equations to the Academy of Sciences in 1829, but the manuscript was lost. A second version was also lost and was not found among Fourier's papers when Fourier, the secretary of the academy, died in 1830. Galois was killed in a duel in 1832, at the age of 20, and it was not until his papers were published in Joseph Liouville's *Journal de mathématiques* in 1846 that his work began to receive the attention it deserved. His theory eventually made the theory of equations into a mere part of the theory of groups. Galois emphasized the group (as he called it) of permutations of the roots of an equation. This move took him away from the equations themselves, instead turning toward the markedly more tractable study of permutations. To any given equation there corresponds a definite group, with a definite collection of subgroups. To explain which equations were solvable by radicals and which were not, Galois analyzed the ways in which these subgroups were related to one another: solvable equations gave rise to what are now called a chain of normal subgroups with cyclic quotients. This technical condition makes it clear how far mathematicians had gone from the familiar questions of 18th-century mathematics, and it marks a transition characteristic of modern mathematics: the replacement of formal calculation by conceptual analysis. This is a luxury available to the pure mathematician that the applied mathematician faced with a concrete problem cannot always afford.

The theory of groups

According to this theory, a group is a set of objects that one can combine in pairs in such a way that the resulting object is also in the set. Moreover, this way of combination has to obey the following rules (here objects in the group are denoted a, b, etc., and the combination of a and b is written $a * b$):

1. There is an element, e, such that $a * e = a = e * a$ for every element a in the group. This element is called the identity element of the group.
2. For every element a there is an element, written a^{-1}, with the property that $a * a^{-1} = e = a^{-1} * a$. The element a^{-1} is called the inverse of a.
3. For every a, b, and c in the group the associative law holds: $(a * b) * c = a * (b * c)$.

Examples of groups include the integers with $*$ interpreted as addition and the positive rational numbers with $*$ interpreted as multiplication. An important property shared by some groups but not all is commutativity: for every element a and b, $a * b = b * a$. The rotations of an object in the plane around a fixed point form a commutative group, but the rotations of a three-dimensional object around a fixed point form a noncommutative group.

Gauss. A convenient way to assess the situation in mathematics in the mid-19th century is to look at the career of its greatest exponent, the last man to be called the "Prince of Mathematics," Carl Friedrich Gauss. In 1801, the same year in which he published his *Disquisitiones Arithmeticae,* he rediscovered the asteroid Ceres (which had disappeared behind the Sun not long after it had first been discovered and before its orbit was precisely known). He was the first to give a sound analysis of the method of least squares in the analysis of statistical data. Gauss did important work in potential theory and, with the German physicist Wilhelm Weber, built the first electric telegraph. He helped conduct the first survey of the Earth's magnetic field and did both theoretical and field work in cartography and surveying. He was a polymath who almost singlehandedly embraced what elsewhere was being put asunder: the world of science and the world of mathematics. It is his purely mathematical work, however, that in its day and ever since has been regarded as the best evidence of his genius.

Gauss's writings transformed the theory of numbers. His theory of algebraic integers lay close to the theory of equations as Galois was to redefine it. More remarkable are the extensive writings on the theory of elliptic functions dating from 1797 to the 1820s but unpublished at his death. In 1827 he published his crucial discovery that the curvature of a surface can be defined intrinsically—that is, solely in terms of properties defined within the surface and without reference to the surrounding Euclidean space. This result was to be decisive in the acceptance of non-Euclidean geometry. All his work displays a sharp concern for rigour and a refusal to rely on intuition or physical analogy, which was to serve as an inspiration to his successors. His emphasis on achieving full conceptual understanding, which may have led to his dislike of publication, was by no means the least influential of his achievements.

Gauss's work on the theory of elliptic functions

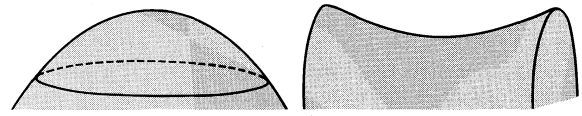


Figure 16: Intrinsic curvature of a surface.

Non-Euclidean geometry. Perhaps it was this desire for conceptual understanding that made Gauss reluctant to publish the fact that he was led more and more "to doubt the truth of geometry," as he put it. For if there was a logically consistent geometry differing from Euclid's only because it made a different assumption about the behaviour of parallel lines, it too could apply to physical space, and so the truth of (Euclidean) geometry could no longer be assured a priori, as Kant had thought.

Gauss's investigations into the new geometry went further than any one else's before him, but he did not publish them. The honour of being the first to proclaim the existence of a new geometry belongs to two others, who did so in the late 1820s: Nicolay Ivanovich Lobachevsky in Russia and János Bolyai in Hungary. Because the similarities in the work of these two men far exceed the differences, it is convenient to describe their work together.

Both men made an assumption about parallel lines that differed from Euclid's and proceeded to draw out its consequences. This way of working cannot guarantee the consistency of one's findings, so strictly speaking they could not prove the existence of a new geometry in this way. Both men described a three-dimensional space different from Euclidean space, couching their findings in the language of trigonometry. The formulas they obtained were exact analogues of the formulas that describe triangles drawn on the surface of a sphere, with the usual trigonometric functions replaced by those of hyperbolic trigonometry. (The functions hyperbolic cosine, written cosh, and hyperbolic sine, written sinh [see Figure 17], are defined as follows: $\cosh x = \frac{1}{2}\{e^x + e^{-x}\}$, and $\sinh x = \frac{1}{2}\{e^x - e^{-x}\}$. They are called hyperbolic because of their use in describ-

The hyperbolic functions

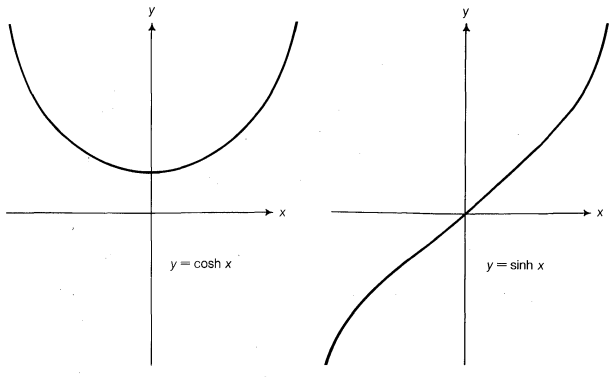


Figure 17: The hyperbolic functions cosh x and sinh x.

ing the hyperbola. Their names derive from the evident analogy with the trigonometric functions, which Euler showed satisfy these equations: $\cos x = \frac{1}{2}\{e^{ix} + e^{-ix}\}$, and $\sin x = (\frac{1}{2}i)\{e^{ix} - e^{-ix}\}$.) The formulas were what gave their work the precision needed to give conviction in the absence of a sound logical structure. Both men observed that it had become an empirical matter to determine the nature of space, Lobachevsky even going so far as to conduct astronomical observations, although these proved inconclusive.

The work of Bolyai and Lobachevsky was poorly received. Gauss endorsed what they had done, but so discreetly that most mathematicians did not find out his true opinion on the subject until he was dead. The main obstacle each man faced was surely the shocking nature of their discovery. It was easier, and in keeping with 2,000 years of tradition, to continue to believe that Euclidean geometry was correct and that Bolyai and Lobachevsky had somewhere gone astray, like many an investigator before them.

The turn toward acceptance came in the 1860s after Bolyai and Lobachevsky had died. The Italian mathematician Eugenio Beltrami decided to investigate Lobachevsky's work and to place it, if possible, within the context of differential geometry as redefined by Gauss. He therefore moved independently in the direction already taken by Bernhard Riemann. He investigated the surface of constant negative curvature and found that on such a surface triangles obeyed the formulas of hyperbolic trigonometry that Lobachevsky had discovered were appropriate to his form of non-Euclidean geometry. Thus he gave the first rigorous description of a geometry other than Euclid's. Beltrami's account of the surface of constant negative curvature was ingenious. He said that it was an abstract surface which he could describe by drawing maps of it, much as one might describe a sphere by means of the pages of a geographic atlas. He did not claim to have constructed the surface embedded in Euclidean three-dimensional space; later Hilbert showed that it cannot be done.

Riemann. When Gauss died in 1855, his post at Göttingen was taken by Peter Gustav Lejeune Dirichlet. One mathematician who found the presence of Dirichlet a stimulus to research was Riemann, and his few short contributions to mathematics were among the most influential of the century. Riemann's first paper, his doctoral thesis (1851) on the theory of complex functions, provided the foundations for a geometric treatment of functions of a complex variable. His main result guaranteed the existence of a wide class of complex functions satisfying only modest general requirements and so made it clear that complex functions could be expected to occur widely in mathematics. More importantly, it was achieved by yoking together the complex theory with the theory of harmonic functions and potential theory. The theories of complex and harmonic functions were henceforth inseparable.

Riemann then wrote on the theory of Fourier series and their integrability. His paper was directly in the tradition that ran from Cauchy and Fourier to Dirichlet, and it marked a considerable step forward in the precision with which the concept of integral can be defined. In 1854 he took up a subject that much interested Gauss, the hypotheses lying at the basis of geometry.

The study of geometry has always been one of the central concerns of mathematicians. It was the language, and the principal subject matter, of Greek mathematics, the mainstay of elementary education in the subject, and it has an obvious visual appeal. It seems easy to apply, for one can proceed from a base of naively intelligible concepts. In keeping with the general trends of the century, however, it was just the naive concepts that Riemann chose to refine. What he proposed as the basis of geometry was far more radical and fundamental than anything that had gone before.

Riemann's work on the foundations of geometry

Riemann took his inspiration from Gauss's discovery that the curvature of a surface is intrinsic, and he argued that one should therefore ignore Euclidean space and treat each surface by itself. A geometric property, he argued, was one that was intrinsic to the surface. To do geometry, it was enough to be given a set of points and a way of measuring lengths along curves in the surface. For this, traditional ways of applying the calculus to the study of curves could be made to suffice. But Riemann did not stop with surfaces. He proposed that geometers study spaces of any dimension in this spirit, even, he said, spaces of infinite dimension.

Several profound consequences followed from this view. It dethroned Euclidean geometry, which now became just one of many geometries. It allowed the geometry of Bolyai and Lobachevsky to be recognized as the geometry of a surface of constant negative curvature, thus resolving doubts about the logical consistency of their work. It highlighted the importance of intrinsic concepts in geometry. It helped open the way to the study of spaces of many dimensions. Last, but not least, Riemann's work ensured that any investigation of the geometric nature of physical space would thereafter have to be partly empirical. One could no longer say that physical space is Euclidean because there is no geometry but Euclid's. This finally destroyed any hope that questions about the world could be answered by a priori reasoning.

In 1857 Riemann published several papers applying his very general methods for the study of complex functions to various parts of mathematics. One of these papers solved the outstanding problem of extending the theory of elliptic functions to the integration of any algebraic function. It opened up the theory of complex functions of several variables and showed how Riemann's novel topological ideas were essential in the study of complex functions. (In subsequent lectures Riemann showed how the special case of the theory of elliptic functions could be regarded as the study of complex functions on a torus.)

Another paper dealt with the question of how many prime numbers there are that are less than any given number x. The answer is a function of x, and Gauss had conjectured on the basis of extensive numerical evidence that this function was approximately $x/\ln(x)$. This turned out to be true, but it was not proved until 1896, when both the Belgian mathematician Charles Jean de la Vallée-Poussin and the French mathematician Jacques-Salomon Hadamard independently proved it. It is remarkable that a question about integers led to a discussion of functions of a complex variable, but similar connections had previously been made by Dirichlet. Riemann took the expression $\Pi(1 - p^{-s})^{-1} = \Sigma n^{-s}$, introduced by Euler the century before, where the infinite product is taken over all prime numbers p and the sum over all whole numbers n and treated it as a function of s. The infinite sum makes sense whenever s is real and greater than 1. Riemann proceeded to study this function when s is complex (now called the Riemann zeta function), and he thereby not only helped clarify the question of the distribution of primes but also was led to several other remarks that later mathematicians were to find of exceptional interest. One remark has continued to elude proof and remains one of the greatest conjectures in mathematics: the claim that the nonreal zeros of the zeta function are complex numbers whose real part is always equal to $\frac{1}{2}$.

The Riemann zeta function

Riemann's influence. In 1859 Dirichlet died and Riemann became a full professor, but he was already ill with tuberculosis, and in 1862 his health broke. He died in 1866. His work, however, exercised a growing influence

on his successors. His work on trigonometric series, for example, led to a deepening investigation of the question of when a function is integrable. Attention was concentrated on the nature of the sets of points at which functions and their integrals (when these existed) had unexpected properties. The conclusions that emerged were at first obscure, but it became clear that some properties of point sets were important in the theory of integration, while others were not. (These other properties proved to be a vital part of the emerging subject of topology.) The properties of point sets that matter in integration have to do with the size of the set. If one can change the values of a function on a set of points without changing its integral, it is said that the set is of negligible size. The naive idea is that integrating is a generalization of counting: negligible sets do not need to be counted. Around the turn of the century the French mathematician Henri-Léon Lebesgue managed to systematize this naive idea into a new theory about the size of sets, which included integration as a special case. In this theory, called measure theory, there are sets that can be measured, and they either have positive measure or are negligible (they have zero measure), and there are sets that cannot be measured at all.

The first success for Lebesgue's theory was that, unlike the Cauchy–Riemann integral, it obeyed the rule that, if a sequence of functions $f_n(x)$ tended suitably to a function $f(x)$, then the sequence of integrals $\int f_n(x)dx$ tended to the integral $\int f(x)dx$. This has made it the natural theory of the integral when dealing with questions about trigonometric series. Another advantage is that it is very general. For example, in probability theory it is desirable to estimate the likelihood of certain outcomes of an experiment. By imposing a measure on the space of all possible outcomes, the Russian mathematician Andrey Kolmogorov was the first to put probability theory on a rigorous mathematical footing.

Another example is provided by a remarkable result discovered by the 20th-century American mathematician Norbert Wiener: within the set of all continuous functions on an interval, the set of differentiable functions has measure zero. In probabilistic terms, therefore, the chance that a function taken at random is differentiable has probability zero. In physical terms, this means that, for example, a particle moving under Brownian motion almost certainly is moving on a nondifferentiable path. This discovery clarified Einstein's fundamental ideas about Brownian motion (displayed by the continual motion of specks of dust in a fluid under the constant bombardment of surrounding molecules). The hope of physicists is that Richard Feynman's theory of quantum electrodynamics will yield to a similar measure-theoretic treatment, for it has the disturbing aspect of a theory that has not been made rigorous mathematically but that accords excellently with observation.

Applications of measure theory

Yet another setting for Lebesgue's ideas was to be the theory of Lie groups. The Hungarian mathematician Alfréd Haar showed how to define the concept of measure so that functions defined on Lie groups could be integrated. This became a crucial part of Hermann Weyl's way of representing a Lie group as acting linearly on the space of all (suitable) functions on the group (for technical reasons, "suitable" means that the square of the function is integrable with respect to a Haar measure on the group).

Differential equations. Another field that developed considerably in the 19th century was the theory of differential equations. The pioneer in this direction once again was Cauchy. Above all, he insisted that one should prove that solutions do indeed exist; it is not a priori obvious that every ordinary differential equation has solutions. The methods that Cauchy proposed for these problems fitted naturally into his program of providing rigorous foundations for all the calculus. The solution method he preferred, although the less general of his two approaches, worked equally well in the real and complex cases. It established the existence of a solution equal to the one obtainable by traditional power series methods using newly developed techniques in his theory of functions of a complex variable.

The harder part of the theory of differential equations concerns partial differential equations, those for which the unknown function is a function of several variables. In the early 19th century there was no known method of proving that a given second- or higher-order partial differential equation had a solution, and there was not even a method of writing down a plausible candidate. In this case progress was to be much less marked. Cauchy found new and more rigorous methods for first-order partial differential equations, but the general case eluded treatment.

An important special case was successfully prosecuted, that of dynamics. Dynamics is the study of the motion of a physical system under the action of forces. Working independently of each other, William Rowan Hamilton in Ireland and Jacobi in Germany showed how problems in dynamics could be reduced to systems of first-order partial differential equations. From this base grew an extensive study of certain partial differential operators. These are straightforward generalizations of a single partial differentiation $(\partial/\partial x)$ to a sum of the form

$$a_1\frac{\partial}{\partial x_1}+\ldots+a_n\frac{\partial}{\partial x_n},$$

where the a's are functions of the x's. The effect of performing several of these in succession can be complicated, but Jacobi and the other pioneers in this field found that there are formal rules which such operators tend to satisfy. This enabled them to shift attention to these formal rules, and gradually an algebraic analysis of this branch of mathematics began to emerge.

The most influential worker in this direction was the Norwegian, Sophus Lie. Lie, and independently Wilhelm Killing in Germany, came to suspect that the systems of partial differential operators they were studying came in a limited variety of types. Once the number of independent variables was specified (which fixed the dimension of the system), a large class of examples, including many of considerable geometric significance, seemed to fall into a small number of patterns. This suggested that the systems could be classified, and such a prospect naturally excited mathematicians. After much work by Lie and Killing and later by the French mathematician Élie-Joseph Cartan, they were classified. Initially, this discovery aroused interest because it produced order where previously the complexity had threatened chaos and because it could be made to make sense geometrically. The realization that there were to be major implications of this work for the study of physics lay well in the future.

The classification of systems of partial differential operators

Linear algebra. Differential equations, whether ordinary or partial, may profitably be classified as linear or nonlinear; linear differential equations are those for which the sum of two solutions is again a solution. The equation giving the shape of a vibrating string is linear, which provides the mathematical reason why a string may simultaneously emit more than one frequency. The linearity of an equation makes it easy to find all its solutions, so in general linear problems have been tackled successfully, while nonlinear equations continue to be difficult. Indeed, in many linear problems there can be found a finite family of solutions with the property that any solution is a sum of them (suitably multiplied by arbitrary constants). Obtaining such a family, called a basis, and putting them into their simplest and most useful form, was an important source of many techniques in the field of linear algebra.

Consider, for example, the system of linear differential equations

$$\frac{dy_1}{dx}=ay_1+by_2,\quad \frac{dy_2}{dx}=cy_1+dy_2.$$

It is evidently much more difficult to study than the system $dy_1/dx=\alpha y_1$, $dy_2/dx=\beta y_2$, whose solutions are (constant multiples of) $y_1=\exp(\alpha x)$ and $y_2=\exp(\beta x)$. But if a suitable linear combination of y_1 and y_2 can be found so that the first system reduces to the second, then it is enough to solve the second system. The existence of such a reduction is determined by the array (called a matrix) of the four numbers $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$. In 1858 the English mathematician Arthur

Cayley began the study of matrices in their own right when he noticed that they satisfy polynomial equations. The matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, for example, satisfies the equation $A^2 - (a+d)A + (ad - bc) = 0$. Moreover, if this equation has two distinct roots, say, α and β, then the sought-for reduction will exist and the coefficients of the simpler system will indeed be those roots α and β. If the equation has a repeated root, then the reduction usually cannot be carried out. In either case, the difficult part of solving the original differential equation has been reduced to elementary algebra.

Vector spaces

The study of linear algebra begun by Cayley and continued by Leopold Kronecker includes a powerful theory of vector spaces. These are sets whose elements can be added together and multiplied by arbitrary numbers, such as the family of solutions of a linear differential equation. A more familiar example is that of three-dimensional space. If one picks an origin, then every point in space can be labeled by the line segment (called a "vector") joining it to the origin. Matrices appear as ways of representing linear transformations of a vector space—*i.e.*, transformations that preserve sums and multiplication by numbers: the transformation T is linear if, for any vectors $\boldsymbol{u}$, $\boldsymbol{v}$, $T(\boldsymbol{u} + \boldsymbol{v}) = T(\boldsymbol{u}) + T(\boldsymbol{v})$ and for any scalar λ, $T(\lambda \boldsymbol{v}) = \lambda T(\boldsymbol{v})$. When the vector space is finite-dimensional, linear algebra and geometry form a potent combination. Vector spaces of infinite dimensions are also studied.

The theory of vector spaces is useful in other ways. Vectors in three-dimensional space represent such physically important concepts as velocities and forces. Such an assignment of vector to point is called a vector field; examples include electric and magnetic fields. Scientists such as Maxwell and J. Willard Gibbs took up vector analysis and were able to extend vector methods to the calculus. They introduced in this way measures of how a vector field varies infinitesimally, which, under the names div, grad, and curl, have become the standard tools in the study of electromagnetism and potential theory. To the modern mathematician, div, grad, and curl form part of a theory to which Stokes's theorem (a special case of which is Green's theorem) is central. This result, named after two leading English applied mathematicians of the 19th century, generalizes the fundamental theorem of the calculus to functions of several variables. The fundamental theorem asserts that

$$\int_b^a f'(x) = f(a) - f(b),$$

which can be read as saying that the integral of the derivative of some function in an interval is equal to the difference in the values of the function at the endpoints of the interval. Generalized to a part of a surface or space, this asserts that the integral of the derivative of some function over a region is equal to the value of the function over the boundary of the region. In symbols, this says that $\int d\omega = \int \omega$, where the first integral is taken over the region in question and the second integral over its boundary, while $d\omega$ is the derivative of ω.

The foundations of geometry. By the late 19th century the hegemony of Euclidean geometry had been challenged by non-Euclidean geometry and projective geometry. The first notable attempt to reorganize the study of geometry was made by the German mathematician Felix Klein and published at Erlangen in 1872. In his *Erlanger Programm* Klein proposed that Euclidean and non-Euclidean geometry be regarded as special cases of projective geometry. In each case the common features that, in Klein's opinion, made them geometries were that there was a set of points, called a "space," and a group of transformations by means of which figures could be moved around in the space without altering their essential properties. For example, in Euclidean plane geometry the space is the familiar plane, and the transformations are rotations, reflections, translations, and their composites, none of which change either length or angle, the basic properties of figures in Euclidean geometry. Different geometries would have different spaces and different groups, and the figures would have different basic properties.

The *Erlanger Programm*

Klein produced an account that unified a large class of geometries, roughly speaking all those which were homogeneous in the sense that every piece of the space looked like every other piece of the space. This excluded, for example, geometries on surfaces of variable curvature, but it produced an attractive package for the rest and gratified the intuition of those who felt that somehow projective geometry was basic. It continued to look like the right approach when Lie's ideas appeared and there seemed to be a good connection between Lie's classification and the types of geometry organized by Klein.

Mathematicians could now ask why they had believed Euclidean geometry to be the only one when, in fact, many different geometries existed. The first to take up this question successfully was the German mathematician Moritz Pasch, who argued in 1882 that the mistake had been to rely too heavily on physical intuition. In his view an argument in mathematics should depend for its validity not on the physical interpretation of the terms involved but upon purely formal criteria. Indeed, the principle of duality did violence to the sense of geometry as a formalization of what one believed about (physical) points and lines; one did not believe that these terms were interchangeable.

The ideas of Pasch caught the attention of the German mathematician David Hilbert, who, with the French mathematician Henri Poincaré, came to dominate mathematics at the turn of the century. In wondering why it was that mathematics, and in particular geometry, produced correct results, he came to feel more and more that it was not because of the lucidity of its definitions. Rather, mathematics worked because its (elementary) terms were meaningless. What kept it heading in the right direction was its rules of inference. Proofs were valid because they were constructed through the application of the rules of inference, according to which new assertions could be declared to be true simply because they could be derived, by means of these rules, from the axioms or previously proved theorems. The theorems and axioms were viewed as formal statements that expressed the relationships between these terms.

The rules governing the use of mathematical terms were arbitrary, Hilbert argued, and each mathematician could choose them at will, provided only that the choices made were self-consistent. A mathematician produced abstract systems unconstrained by the needs of science, and, if scientists found an abstract system that fit one of their concerns, they could apply the system secure in the knowledge that it was logically consistent.

The axiomatization of geometry

Hilbert first became excited about this point of view (presented in his *Grundlagen der Geometrie* ["Foundations of Geometry"], 1899) when he saw that it led not merely to a clear way of sorting out the geometries in Klein's hierarchy according to the different axiom systems they obeyed but to new geometries as well. For the first time there was a way of discussing geometry that lay beyond even the very general terms proposed by Riemann. Not all of these geometries have continued to be of interest, but the general moral that Hilbert first drew for geometry he was shortly to draw for the whole of mathematics.

The foundations of mathematics. By the late 19th century the debates about the foundations of geometry had become the focus for a running debate about the nature of the branches of mathematics. Cauchy's work on the foundations of the calculus, completed by the German mathematician Karl Weierstrass in the late 1870s, left an edifice that rested on concepts such as that of the natural number (the integers 1, 2, 3, and so on) and on certain constructions involving them. The algebraic theory of numbers and the transformed theory of equations had focused attention on abstract structures in mathematics. Questions that had been raised about numbers since Babylonian times turned out to be best cast theoretically in terms of entirely modern creations whose independence from the physical world was beyond dispute. Finally, geometry, far from being a kind of abstract physics, was now seen as dealing with meaningless terms obeying arbitrary systems of rules. Although there had been no conscious

plan leading in that direction, the stage was set for a consideration of questions about the fundamental nature of mathematics.

Similar currents were at work in the study of logic, which had also enjoyed a revival during the 19th century. The work of the English mathematician George Boole and the American Charles Sanders Peirce had contributed to the development of a symbolism adequate to explore all elementary logical deductions. Significantly, Boole's book on the subject was called *An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities* (1854). In Germany, the logician Gottlob Frege had directed keen attention to such fundamental questions as what it means to define something and what sorts of purported definitions actually do define.

Cantor. All of these debates came together through the pioneering work of the German mathematician Georg Cantor on the concept of a set. Cantor had begun work in this area because of his interest in Riemann's theory of trigonometric series, but the problem of what characterized the set of all real numbers came to occupy him more and more. He began to discover unexpected properties of sets. For example, he could show that the set of all algebraic numbers, and a fortiori the set of all rational numbers, is countable in the sense that there is a one-to-one correspondence between the integers and the members of each of these sets by means of which for any member of the set of algebraic numbers (or rationals), no matter how large, there is always a unique integer it may be placed in correspondence with. But, more surprisingly, he could also show that the set of all real numbers is not countable. So, although the set of all integers and the set of all real numbers are both infinite, the set of all real numbers is a strictly larger infinity. This was in complete contrast to the prevailing orthodoxy, which proclaimed that infinite could only mean "larger than any finite amount."

Set theory

Here the concept of number was being extended and undermined at the same time. The concept was extended because it was now possible to count and order sets that the set of integers was too small to measure; and it was undermined because even the integers ceased to be basic undefined objects. Cantor himself had given a way of defining real numbers as certain infinite sets of rational numbers. Rational numbers were easy to define in terms of the integers, but now integers could be defined by means of sets. One way was given by Frege in *Die Grundlagen der Arithmetik* (1884). He regarded two sets as the same if they contained the same elements. So in his opinion there was only one empty set (today symbolized by Ø), the set with no members. A second set could be defined as having only one element by letting that element be the empty set itself (symbolized by {Ø}), a set with two elements by letting them be the two sets just defined (*i.e.*, {Ø, {Ø}}), and so on. Having thus defined the integers in terms of the primitive concepts "set" and "element of," Frege agreed with Cantor that there was no logical reason to stop, and he went on to define infinite sets in the same way Cantor had. Indeed, Frege was clearer than Cantor about what sets and their elements actually were.

Frege's proposals went in the direction of a reduction of all mathematics to logic. He hoped that every mathematical term could be defined precisely and manipulated according to agreed, logical rules of inference. This, the "logicist" program, was dealt an unexpected blow by the English mathematician and philosopher Bertrand Russell in 1902, who pointed out unexpected complications with the naive concept of a set. Nothing seemed to preclude the possibility that some sets were elements of themselves while others were not, but asked Russell, "what then of the set of all sets that were not elements of themselves?" If it is an element of itself, then it is not (an element of itself), but, if it is not, then it is—a paradox. Either the idea of a set as an arbitrary collection of already defined objects was flawed, or else the idea that one could legitimately form the set of all sets of a given kind was incorrect. Frege's program never recovered from this blow, and the theories of Russell, which he developed together with Alfred North Whitehead in their *Principia Mathematica* (1910–13), that went in the same direction never found lasting appeal with mathematicians.

Greater interest attached to the ideas that Hilbert and his school began to advance. It seemed to them that what had worked once for geometry could work again for all of mathematics. Rather than attempt to define things so that problems could not arise, they suggested that it was possible to dispense with definitions and cast all of mathematics in an axiomatic structure using the ideas of set theory. Indeed, the hope was that the study of logic could be embraced in this spirit, thus making logic a branch of mathematics, the opposite of Frege's intention. There was considerable progress in this direction, and there emerged both a powerful school of mathematical logicians (notably in Poland) and an axiomatic theory of sets that avoided Russell's paradoxes and the others which had sprung up.

The axiomatic school

In the 1920s Hilbert put forward his most detailed proposal for establishing the validity of mathematics. According to his theory of proofs, everything was to be put into an axiomatic form, allowing the rules of inference to be only those of elementary logic, and only those conclusions that could be reached from this finite set of axioms and rules of inference were to be admitted. He proposed that a satisfactory system would be one which was consistent, complete, and decidable. By consistent Hilbert meant that it should be impossible to derive both a statement and its negation; by complete, that every properly written statement should be such that either it or its negation was derivable from the axioms; by decidable, that one should have an algorithm which determines of any given statement whether it or its negation is provable. Such systems did exist, for example, the first-order predicate calculus, but none had been found capable of allowing mathematicians to do interesting mathematics.

Hilbert's program, however, did not last long. In 1931 the Austrian-born American mathematician and logician Kurt Gödel showed that there was no system of Hilbert's type within which the integers could be defined and which was both consistent and complete. Later Gödel and, independently, the English mathematician Alan Turing showed that decidability was also unattainable. Perhaps paradoxically, the effect of this dramatic discovery was to alienate mathematicians from the whole debate. Instead, mathematicians, who may not have been too unhappy with the idea that there is no way of deciding the truth of a proposition automatically, learned to live with the idea that not even mathematics rests on rigorous foundations. Progress since has been in other directions. An alternative axiom system for set theory was later put forward by the Hungarian-born American mathematician John von Neumann, which he hoped would help resolve contemporary problems in quantum mechanics. There was also a renewal of interest in statements that are both interesting mathematically and independent of the axiom system in use. The first of these was the American mathematician Paul Cohen's surprising resolution in 1963 of the continuum hypothesis, which was Cantor's conjecture that the set of all subsets of the rational numbers was of the same size as the set of all real numbers. This turns out to be independent of the usual axioms for set theory, so there are set theories (and therefore types of mathematics) in which it is true and others in which it is false.

The continuum hypothesis

Mathematical physics. At the same time that mathematicians were attempting to put their own house in order, they were also looking with renewed interest at contemporary work in physics. The man who did the most to rekindle their interest was Poincaré. Poincaré showed that dynamic systems described by quite simple differential equations, such as the solar system, can nonetheless yield the most random-looking, chaotic behaviour. He went on to explore ways in which mathematicians can nonetheless say things about this chaotic behaviour and so pioneered the way in which probabilistic statements about dynamic systems can be found to describe what otherwise defies our intelligence.

Poincaré later turned to problems of electrodynamics. After many years' work, the Dutch physicist Hendrik Antoon Lorentz had been led to an apparent dependence of length and time on motion, and Poincaré was pleased to notice

that the transformations that Lorentz proposed as a way of converting one observer's data into another's formed a group. This appealed to Poincaré and strengthened his belief that there was no sense in a concept of absolute motion; all motion was relative. Poincaré thereupon gave an elegant mathematical formulation of Lorentz's ideas, which fitted them into a theory in which the motion of the electron is governed by Maxwell's equations. Poincaré, however, stopped short of denying the reality of the ether or of proclaiming that the velocity of light is the same for all observers, so credit for the first truly relativistic theory of the motion of the electron rests with Einstein and his special theory of relativity (1905).

Einstein's special theory is so called because it treats only the special case of uniform relative motion. The much more important case of accelerated motion and motion in a gravitational field was to take a further decade and to require a far more substantial dose of mathematics. Einstein only changed his estimate of the value of pure mathematics, which he had hitherto disdained, when he discovered that many of the questions he was led to had already been formulated mathematically and had been solved. He was most struck by theories derived from the study of geometry in the sense in which Riemann had formulated it.

By 1915 a number of mathematicians were interested in reapplying their discoveries to physics. The leading institution in this respect was the University of Göttingen, where Hilbert had unsuccessfully attempted to produce a general theory of relativity before Einstein, and it was there that many of the leaders of the coming revolution in quantum mechanics were to study. There, too, went many of the leading mathematicians of their generation, notably John von Neumann and Hermann Weyl, to study with Hilbert. In 1904 Hilbert had turned to the study of integral equations. These arise in many problems where the unknown is itself a function of some variable, and especially in those parts of physics that are expressed in terms of extremal principles (such as the principle of least action). The extremal principle usually yields information about an integral involving the sought-for function, hence the name "integral equation." Hilbert's contribution was to bring together many different strands of contemporary work and to show how they could be elucidated if cast in the form of arguments about objects in certain infinite-dimensional vector spaces.

Integral equations

The extension to infinite dimensions was not a trivial task, but it brought with it the opportunity to use geometric intuition and geometric concepts to analyze problems about integral equations. Hilbert left it to his students to provide the best abstract setting for his work, and thus was born the concept of a Hilbert space. Roughly speaking this is an infinite-dimensional vector space in which it makes sense to speak of the lengths of vectors and the angles between them; useful examples include certain spaces of sequences and certain spaces of functions. Operators defined on these spaces are also of great interest; their study forms part of the field of functional analysis.

When in the 1920s mathematicians and physicists were seeking ways to formulate the new quantum mechanics, von Neumann proposed that the subject be written in the language of functional analysis. The quantum mechanical world of states and observables, with its mysterious wave packets that were sometimes like particles and sometimes like waves depending on how they were observed, went very neatly into the theory of Hilbert spaces. Functional analysis has ever since grown with the fortunes of particle physics.

Algebraic topology. The early 20th century saw the emergence of a number of theories whose power and utility resides in large part in their generality. Typically, they are marked by an attention to the set or space of all examples of a particular kind. (Functional analysis is such an endeavour.) One of the most energetic of these general theories was that of algebraic topology. In this subject, a variety of ways are developed for replacing a space by a group and a map between spaces by a map between groups. It is like using X rays: information is lost, but the shadowy image of the original space may turn out to contain, in an accessible form, enough information to solve the question at hand.

Interest in this kind of research came from various directions. Galois's theory of equations was an example of what could be achieved by transforming a problem in one branch of mathematics into a problem in another, more abstract branch. Another impetus came from Riemann's theory of complex functions. He had studied algebraic functions, that is, loci defined by equations of the form $f(x, y) = 0$, where f is a polynomial in x whose coefficients are polynomials in y. When x and y are complex variables, the locus can be thought of as a real surface spread out over the x plane of complex numbers (today called a Riemann surface). To each value of x there correspond a finite number of values of y. Such surfaces are not easy to comprehend, and Riemann had proposed to draw curves along them in such a way that, if the surface was cut open along them, it could be opened out into a polygonal disk (see Figure 18). He was able to establish a profound connection between the minimum number of curves needed to do this for a given surface and the number of functions (becoming infinite at specified points) that the surface could then support.

Riemann surfaces

The natural problem was to see how far Riemann's ideas could be applied to the study of spaces of higher dimension. Here two lines of inquiry developed. One emphasized what could be obtained from looking at the projective geometry involved. This point of view was fruitfully applied by the Italian school of algebraic geometers. It ran into problems, which it was not wholly able to solve, having to do with the singularities a surface can possess. Whereas a locus given by $f(x, y) = 0$ may intersect itself only at isolated points, a locus given by an equation of the form $f(x, y, z) = 0$ may intersect itself along curves, a problem that caused considerable difficulties. The second approach emphasized what can be learned from the study of integrals along paths on the surface. This approach, pursued by Charles-Émile Picard and by Poincaré, provided a rich generalization of Riemann's original ideas.

On this base, conjectures were made and a general theory produced, first by Poincaré and then by the American engineer-turned-mathematician Solomon Lefschetz (1884–1972), concerning the nature of manifolds of arbitrary dimension. Roughly speaking a manifold is the n-dimensional generalization of the idea of a surface; it is a space any small piece of which looks like a piece of n-dimensional space. Such an object is often given by a single algebraic equation in $n + 1$ variables. At first their work was concerned with how these manifolds may be decomposed into pieces, counting the number of pieces and decomposing them in their turn. The result was a list of numbers, called Betti numbers in honour of the Italian mathematician Enrico Betti, who had taken the first steps of this kind of extend Riemann's work. It was only in the late 1920s that the German mathematician Emmy Noether suggested how the Betti numbers might be thought of as measuring the size of certain groups. At her

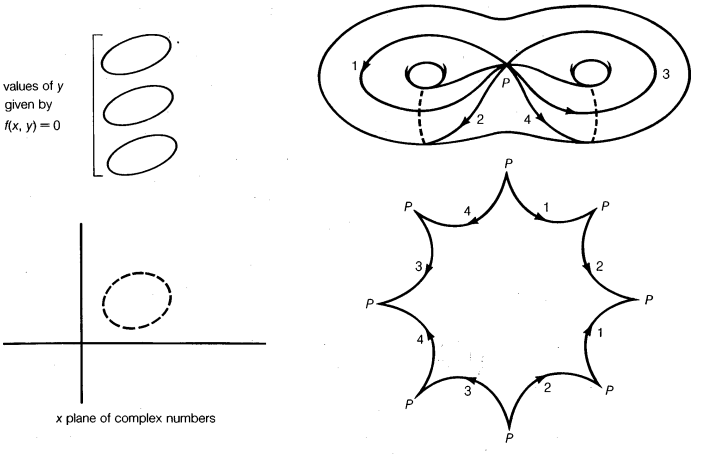


Figure 18: (Left) Pieces of a surface given by $f(x, y) = 0$; (right) if the surface is cut along the curves, an octagon is obtained.

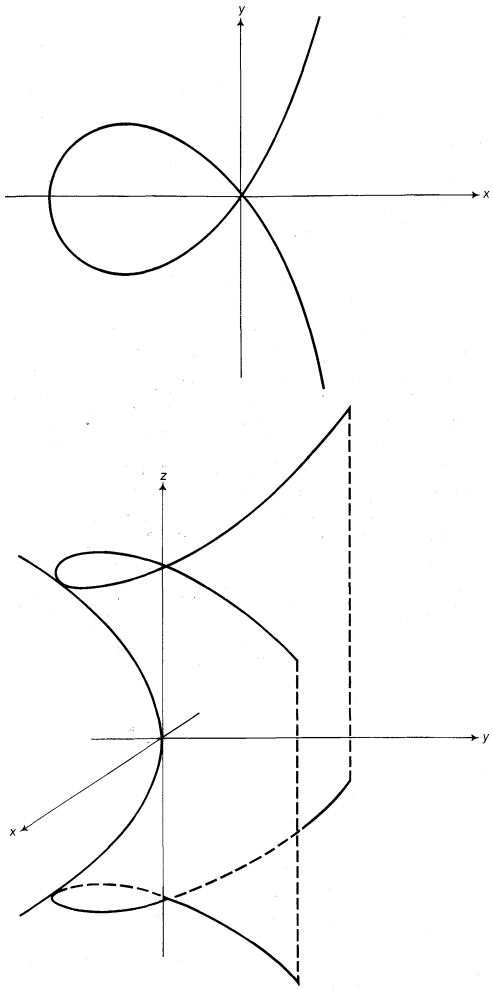


Figure 19: (Top) $f(x, y) = y^2(y + 1) - x^2 = 0$ intersects itself at $(x, y) = (0, 0)$; (bottom) $E(x, y, z) = 0 = y^2(y + z^2) - x^2$ intersects itself along the z-axis, but the origin is a triple self-intersection.

instigation, a number of people then produced a theory of these groups, the so-called homology and cohomology groups of a space.

Homology and cohomology

Two objects that can be deformed into one another will have the same homology and cohomology groups. To assess how much information is lost when a space is replaced by its algebraic topological picture, Poincaré asked the crucial converse question: according to what algebraic conditions is it possible to say that a space is topologically equivalent to a sphere? He showed by an ingenious example that having the same homology is not enough and proposed a more delicate index, which has since grown into the branch of topology called homotopy theory. Being more delicate, it is both more basic and more difficult. There are usually standard methods for computing homology and cohomology groups, and they are completely known for many spaces. In contrast, there is scarcely an interesting class of spaces for which all the homotopy groups are known. And Poincaré's original question continues to resist answer, but only in the dimension in which he raised it: there is still no algebraic criterion for recognizing the 3-sphere.

Developments in pure mathematics. The interest in axiomatic systems at the turn of the century led to axiom systems for the known algebraic structures, that for the theory of fields, for example, being developed by the German mathematician Ernst Steinitz in 1910. The theory of rings (structures in which it is possible to add, subtract, and multiply but not necessarily divide) was much harder to formalize. It is important for two reasons: the theory of algebraic integers forms part of it, because algebraic integers naturally form into rings; and (as Kronecker and Hilbert had argued) algebraic geometry forms another part. The rings that arise there are rings of functions definable on the curve, surface, or manifold, or are definable on specific pieces of it.

Problems in number theory and algebraic geometry are often very difficult, and it was the hope of mathematicians such as Noether, who laboured to produce a formal, axiomatic theory of rings, that by working at a more rarified level the essence of the concrete problems would remain while the distracting special features of any given case would fall away. This would make the formal theory both more general and easier. To a surprising extent these mathematicians were successful—so much so that it has become usual today to teach the general theory without ever getting around to the rich special cases it contains.

A further twist to the development came with the work of the American mathematician Oscar Zariski (1899–1986), who had studied with the Italian school of algebraic geometers but came to feel that their method of working was imprecise. He worked out a detailed program whereby every kind of geometric configuration could be redescribed in algebraic terms. His work succeeded in producing a rigorous theory, although some, notably Lefschetz, felt that the geometry had been lost sight of in the process.

The study of algebraic geometry was amenable to the topological methods of Poincaré and Lefschetz so long as the manifolds were defined by equations whose coefficients were complex numbers, and so it was quite well developed there. But with the creation of an abstract theory of fields it was natural to want a theory of varieties defined by equations with coefficients in an arbitrary field. This was provided for the first time by the French mathematician André Weil (b. 1906), in his *Foundations of Algebraic Geometry* (1946), in a way that drew on Zariski's work without suppressing the intuitive appeal of geometric concepts. Weil's theory of polynomial equations in any number of indeterminates and with coefficients in an arbitrary field is the proper setting for any investigation that seeks to determine what properties of a geometric object can be derived solely by algebraic means. But it falls tantalizingly short of one topic of importance: the solution of polynomial equations in integers. This was the topic that Weil took up next, and he was led to conjectures that brought together the new techniques of algebraic topology and algebraic geometry.

Weil's theory of polynomial equations

The central difficulty is that in a field it is possible to divide, but in a ring it is not. The integers form a ring but not a field (dividing 1 by 2 does not yield an integer). But Weil showed that simplified versions (posed over a field) of any question about integer solutions to polynomials could be profitably asked. This transferred the questions to the domain of algebraic geometry. To count the number of solutions, Weil proposed that, since the questions were now geometric, they should be amenable to the techniques of algebraic topology. This was an audacious move, since there was no suitable theory of algebraic topology available, but Weil could see what such a theory should do and therefore conjectured what results it should yield. The difficulty of Weil's conjectures may be judged by the fact that the last of them was a generalization to this setting of the famous Riemann hypothesis about the zeta function, and they rapidly became the focus of international attention.

Weil, along with Claude Chevalley, Henri Cartan, Jean Dieudonné, and others, created a group of young French mathematicians who began to publish a virtual encyclopaedia of mathematics under the name Nicolas Bourbaki, taken by Weil from an obscure general of the Franco-German War. Bourbaki became a self-selecting group of young mathematicians who were strong on algebra, and the individual Bourbaki members were interested in the Weil conjectures. In the end, they succeeded completely. A new kind of algebraic topology was developed, and the Weil conjectures proved. The generalized Riemann hypothesis was the last to surrender, being established by the Belgian mathematician Pierre Deligne in the early 1970s. Strangely, its resolution still leaves the original Riemann hypothesis unsolved.

Bourbaki was a key figure in the rethinking of structural mathematics. People who worked in algebraic topology increasingly found that it was dependent on a battery of

cumbersome techniques, and the key ideas risked being swamped by the artifices needed to make it work. For this reason Samuel Eilenberg, a Polish-born American mathematician who belonged for a time to Bourbaki, and the American mathematician Norman Steenrod put forward an axiomatization of algebraic topology. Together with another American mathematician, Saunders MacLane, Eilenberg then situated this axiomatization in a broader context. The end result was that most types of mathematical structures could be seen to arise naturally in families,

Categories

called categories. There is the category of all groups, for example, and the category of all topological spaces. In a given category it is usually possible to pass from one object to another by suitable kinds of map. For example, it may turn out that one group can be mapped to another by a map that preserves multiplication, or perhaps one topological space can be mapped to another by a continuous map (one that leaves nearby points near one another). To do algebraic topology was to transfer a problem posed in one category (that of topological spaces) to another (usually that of commutative groups or rings). When he created the right algebraic topology for the Weil conjectures, the German-born French mathematician Alexandre Grothendieck, a Bourbaki of enormous energy, produced a new description of algebraic geometry. In his hands it became infused with the language of category theory. The route to algebraic geometry became steeper than ever, but the views from the summit have a naturalness and a profundity that have brought many experts to prefer it to the earlier formulations, including Weil's.

Grothendieck's formulation makes algebraic geometry the study of equations defined over rings rather than fields. Accordingly, it raises the possibility that questions about the integers can be answered directly. Strong schools of arithmetic algebraic geometers, as they are called, now exist in the United States, France, Germany, and the Soviet Union. In 1983 they achieved their greatest success to date when the West German mathematician Gerd Faltings, using all the resources of Grothendieck's legacy, resolved an old conjecture of the English number theorist Louis Mordell. That conjecture stated that almost all polynomial equations that define curves have at most finitely many rational solutions; the cases excluded from the conjecture are the simple ones that are much better understood. Faltings' work has important implications for work on Fermat's last theorem. If it can be made to yield computable estimates for the number of points in any given case, it is thought likely that Fermat's theorem could be proved.

Mathematical physics and the theory of groups. In the 1910s the ideas of Lie and Killing were taken up by the French mathematician Elie-Joseph Cartan, who simplified their theory and rederived the classification of what came to be called the classical complex Lie algebras. The simple Lie algebras, out of which all the others in the classification are made, were all representable as algebras of matrices, and in a sense Lie algebra is the abstract setting for matrix algebra. Connected to each Lie algebra there were a small number of Lie groups, and there was a canonical simplest one to choose in each case. The groups had an even simpler geometric interpretation than the corresponding algebras, for they turned out to describe motions that leave certain properties of figures unaltered. For example, in Euclidean three-dimensional space, rotations leave unaltered the distances between points; the set of all rotations about a fixed point turns out to form a Lie group, and it is one of the Lie groups in the classification. The theory of Lie algebras and Lie groups shows that there are only a few sensible ways to measure properties of figures in a linear space and that these methods yield groups of motions leaving the figures, which are (more or less) groups of matrices, unaltered. The result is a powerful theory that could be expected to apply to a wide range of problems in geometry and physics.

The leader in the endeavours to make Cartan's theory, which was confined to Lie algebras, yield results for a corresponding class of Lie groups was the German-American

Hermann Weyl

mathematician Hermann Weyl. He produced a rich and satisfying theory for the pure mathematician and wrote extensively on differential geometry and group theory and its applications to physics. He attempted to produce a theory that would unify gravitation and electromagnetism. His theory met with criticism from Einstein and was generally regarded as unsuccessful; it is only recently that similar unified field theories have met with any acceptance. Nonetheless, Weyl's approach indicates how the theory of Lie groups can enter into physics in a substantial way.

In any physical theory the endeavour is to make sense of observations. Different observers make different observations. If they differ in choice and direction of their coordinate axes, they give different coordinates to the same points, and so on. Yet the observers agree on certain consequences of their observations: in Newtonian physics and Euclidean geometry they agree on the distance between points. Special relativity explains how observers in a state of uniform relative motion differ about lengths and times, but agree on a quantity called the interval. In each case they are able to do so because the relevant theory presents them with a group of transformations that converts one observer's measurements into another's and leaves the appropriate basic quantities invariant. What Weyl proposed was a group that would permit observers in nonuniform relative motion, and whose measurements of the same moving electron would differ, to convert their measurements and thus permit the (general) relativistic study of moving electric charges.

In the 1950s the American physicists Chen Ning Yang and Robert L. Mills gave a successful treatment of the so-called strong interaction in particle physics from the Lie group point of view. Twenty years later mathematicians took up their work, and a dramatic resurgence of interest in Weyl's theory began. These new developments, which had the incidental effect of enabling mathematicians to escape the problems in Weyl's original approach, were the outcome of lines of research that had originally been conducted with little regard for physical questions. Not for the first time, mathematics was to prove surprisingly or, as the Hungarian-born American physicist Eugene Wigner said, "unreasonably effective," in science.

Cartan had investigated how much may be accomplished in differential geometry using the idea of moving frames of reference. This work, which was partly inspired by Einstein's theory of general relativity, was also a development of the ideas of Riemannian geometry that had originally so excited Einstein. In the modern theory, one imagines a space (usually a manifold) made up of overlapping coordinatized pieces. On each piece, one supposes some functions to be defined, which might in applications be the values of certain physical quantities. Rules are given for interpreting these quantities where the pieces overlap. The data are thought of as a bundle of information provided at each point. For each function defined on each patch, it is supposed that at each point a vector space is available as mathematical storage space for all its possible values. Because a vector space is attached at each point, the theory is called the theory of vector bundles. Other kinds of space may be attached, thus entering the more general theory of fibre bundles. The subtle and vital point is that it is possible to create quite different bundles which nonetheless look similar in small patches. An example of this is illustrated in Figure 20. The cylinder and the Möbius band look alike in small pieces but are topologically distinct, since it is possible to give a standard sense of direction to all the lines in the cylinder but not to those in the Möbius band. Both spaces can be thought of as one-dimensional vector bundles over the circle, but they are very different. The cylinder is regarded as a "trivial" bundle, the Möbius band as a twisted one.

The theory of vector bundles

In the 1940s and '50s a vigorous branch of algebraic topology established the main features of the theory of bundles. Then, in the 1960s, work chiefly by Grothendieck and the English mathematician Michael Atiyah showed how the study of vector bundles on spaces could be regarded as the study of cohomology theory (called K theory). More significantly still, in the 1960s Atiyah, the American Isadore Singer, and others found ways of connecting this work to the study of a wide variety of questions involving partial differentiation, culminating in the celebrated

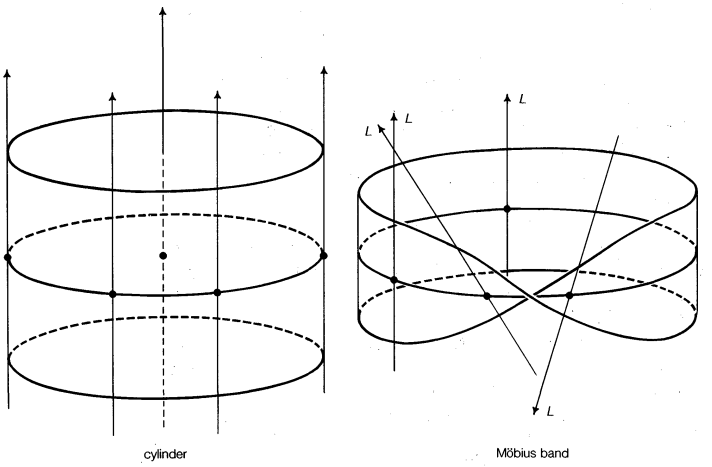


Figure 20: *Vector bundles.*
As the circle is followed clockwise around the Möbius band, the line *L* twists through the half a turn, so the lines cannot be consistently made to point in the same direction.

Atiyah–Singer theorem for elliptic operators. (Elliptic is a technical term for the type of operator studied in potential theory.) There are remarkable implications for the study of pure geometry, and much attention has been directed to the problem of how the theory of bundles embraces the theory of Yang and Mills, which it does precisely because there are nontrivial bundles, and to the question of how it can be made to pay off in large areas of theoretical physics. These include the theories of superspace and supergravity and the string theory of fundamental particles, which involves the theory of Riemann surfaces in novel and unexpected ways. (J.J.G.)

Mathematics in China and Japan

When speaking of mathematics in East Asia, it is necessary to take into account its development in China, Korea, and Japan as a whole. Mathematicians from these countries can be considered as part of the same working community. Moreover, these scholars usually wrote with Chinese characters and thus could read one another's texts.

The common basis for the development of mathematics in East Asia was books that were written in China from the 1st through the 13th century, to which most of the subsequent works refer. For this period, the two main sources, the references found in the texts and those given in the bibliographies compiled for the dynastic annals, indicate that there are many lacunae in the books that have survived. The oldest works extant survived because they became official books.

In the 17th century few ancient Chinese mathematical works were known, and those that were known were not fully understood. Thereafter, as Chinese mathematicians became aware of European achievements, they began to look for such works throughout the country. Editions of the texts they found began to appear at the end of the 18th century and have become the main sources for the history of Chinese mathematics. Discoveries of new sources are now rare, contrary to the situation for Arabic mathematics, for example. Nevertheless, in the 20th century a mathematical book was discovered in a grave dating from the 2nd or 3rd century, which pushed back by more than three centuries what was previously known about the subject.

CHINESE MATHEMATICS TO THE 13TH CENTURY

The *Nine Chapters*

Outline of the history. The *Nine Chapters on the Mathematical Procedures* (or the *Nine Chapters*) was probably compiled in the 1st century AD. This book gathered and organized many mathematical achievements from preceding periods. It played an important part in the development of mathematics in China, for all Chinese mathematicians refer to it and most of the subjects they have worked on stem from it. Its format, which was adopted by most subsequent authors, consists of problems for which a numerical answer and a procedure for solution are given. These problems are arranged in classes that come under the heading of a general method. Many scholars wrote commentaries on the *Nine Chapters,* adding explanations and proofs, rewriting procedures, and suggesting new formulas. The most important of the commentaries to survive, attributed to Liu Hui (3rd century), contains the richest set of proofs within this tradition.

Some of the books written subsequently are known because, gathered together with the *Nine Chapters* and the astronomical treatise *Mathematical Classic of the Gnomon of Chou* and commented on in the 7th century by a group under the leadership of Li Ch'un-feng, they became the *Ten Classics of Mathematics,* the manual for officials trained in the then newly established office of mathematics. Although some people were thus officially trained as mathematicians, no major breakthrough seems to have been achieved until the 11th century.

At that time (1084) the *Ten Classics of Mathematics* was edited and printed, and this seems to have been related to renewed activity in mathematics during the 11th and 12th centuries, known today only through later quotations. This period probably paved the way for the major achievements of Chinese tradition, as they are known today only through the few books that have come down from the second half of the 13th century. China was then divided into two countries, North and South, and achievements by mathematicians of both sides are known: in the South, those of Ch'in Chiu-shao and Yang Hui (who were minor officials), and in the North, those of Li Yeh (a recluse scholar) and Chu Shih-chieh (a wandering teacher). Their contributions seem to have been arrived at independently but they attest to a common basis.

From the period after this, no valuable works survive, and there is no evidence that any important mathematical works were written. Still, some major works of the 13th century are recorded in the encyclopaedia compiled under the Yung-lo emperor (1402–24), but commentaries on these books written by the end of the 15th century show that by this time they were no longer understood.

It was only in the 16th century that the abacus appears to have come into widespread use, and most of the books of the period discuss it. But it is not possible to date the moment when it actually appeared in China.

The following discussion of the evolution of mathematical subjects within the Chinese tradition emphasizes several common characteristics of most of the achievements: a specific use of algorithms and the importance given to position, to configurations of numbers, and to parallelisms between procedures.

The Nine Chapters. This book presupposes mathematical knowledge about how to represent numbers and how to perform the four arithmetic operations. The numbers are written in Chinese characters, but, for most of the procedures described, the computations are to be performed on a surface, perhaps on the ground (see below). Most probably, as can be inferred from later accounts, the numbers were represented with counting rods, according to a decimal place-value system. Such a representation can be moved and modified within a computation. Indian and Arabic mathematicians computed (among other ways) on erasable surfaces, and it appears that arithmetic operations were performed in China in ways comparable to these, although no written computations were recorded until much later. As will be seen, setting up the computations in this way greatly influenced later mathematical developments.

Computational methods

The *Nine Chapters* contains a number of mathematical achievements, already in a mature form, that are presented by most of the subsequent books without substantial changes.

Arithmetic of fractions. Division is a central operation in the *Nine Chapters.* The numbers used throughout are of the type that results from division: "integer plus fraction." Fractions are defined to be a natural part of the result of a division, the remainder of the division being taken as the numerator and the divisor as the denominator. Thus, dividing 17 by 5, one obtains 3 as the quotient and a remainder of 2. This gives rise to the fraction $^2/_5$. The fractions are thus always less than one, and their arithmetic is

described through the use of division. To get the sum of a set of fractions, one is instructed to "multiply the numerators by the denominators that do not correspond to them, add to get the dividend. Multiply the denominators all together to get the divisor. Perform the division. If there is a remainder, name it with the divisor." The sum of a set of fractions is itself thus the result of a division, of the form integer plus fraction. All the operations are described in a similar way. The book also contains explanations of such algorithms as the "rule of three," "sharing in unequal parts," "false double position," and so on.

Formulas for the computations of areas and volumes. The *Nine Chapters* gives formulas for elementary plane and solid figures as well as for the area of the circle, segment of a circle, and sphere. All these formulas are expressed by lists of operations to perform on the data to get the result, an algorithmic expression that is common in ancient traditions (see below on the remarkable elaboration of algorithms found in the book). For example, to compute the area of the circle, the following algorithm is given: "multiply the diameter by itself, triple this, divide by four." This algorithm amounts to taking 3 for π. Commentators added improved values for π along with some derivations. Liu Hui gave two other values for π, one slightly low ($^{157}/_{50}$), and one high ($^{3,927}/_{1,250}$). The *Nine Chapters* also provides the correct formula for the area of the circle: "multiplying half the diameter and half the circumference, one gets the area," which Liu Hui proved.

Solution of systems of simultaneous linear equations. The chapter devoted to this topic provides a general procedure. For instance, one of its problems, on the yields from three grades of grain, leads to a system of three linear equations in three unknowns, and the procedure for the solution arranges the data on the computing surface in a table, as in Figure 21a. The coefficients of the first equation are arranged in the right column. The coefficients of the second and third equations are arranged in the middle and left columns, respectively, and so on. Consequently, the numbers of the first row all correspond to the first unknown. This is an instance of a positional notation, in which the position of a number in a numerical configuration has a mathematical meaning. The main tool for the solution is the use of column reduction to obtain an equivalent configuration (see Figure 21). Next, the unknown of the third line is found by division, and hence the second and the first unknowns as well.

Positional notation

The description of the algorithm uses the structure that the configuration gives to the data on the counting surface in an essential way. Because this procedure implies a column-to-column subtraction, it gives rise to negative numbers. Detailed methods of computing with positive and negative coefficients are used to solve problems involving two to seven unknowns. This seems to be the first occurrence of negative numbers in the history of mathematics. The *Nine Chapters* also contains some knowledge that was to be reworked later.

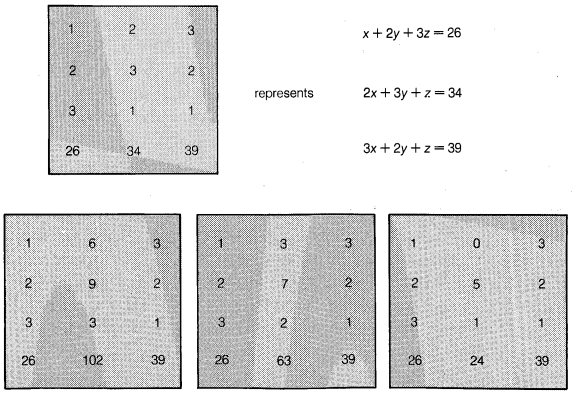


Figure 21: *The first example of a system of linear equations in the Nine Chapters.*
(Top) The coefficients of each equation are arranged in columns in the box; the coefficients of the middle column are multiplied by 3 (the first position in the right column), and the right column is then subtracted as many times as necessary to produce 0 in the first position of the middle column. (Bottom) The operation is repeated to obtain a "triangular matrix" from which the unknowns are then obtained one after the other.

Square root and cube root extraction. Algorithms in the *Nine Chapters* for finding integer parts of roots on the counting surface are based on the same idea as the arithmetic ones used today. They are described using the technical terms of division, in such a way that both roots appear to be derived from this operation. These algorithms are set up on the surface in the same way as is a division ("quotient" on the top; under it, the "dividend"; one row below, the "divisor"; at the bottom, auxiliary computations). Moreover, the two algorithms are written out, sentence by sentence, parallel to each other, so that their similarities and differences become clear.

In the writing of algorithms, there are three basic operations: iteration, conditional statements, and assignation of variables. Once the integer parts are found, the algorithms supply a fractional approximation of the root. (To find the square root of A, for example, if the integer part of the root found by the algorithm is a, the result is given as $[a + (A - a^2)/2a]$.)

In commenting on these algorithms, Liu Hui gives new approximations and suggests that one should continue computing the digits in the same way, setting 10 as denominator for the first digit, 100 as denominator for the second digit, and so on; he thus gives the root in terms of decimal fractions.

In the *Nine Chapters* a problem is solved with a quadratic equation, which appears to be conceived of as an arithmetic operation depending on square root extraction. The procedure for solving the equation is part of the procedure for root extraction. However, the equation is thought to have only one root. The theory of equations developed in China within that framework, until its apogee in the 13th century. (The solution by radicals that Babylonian mathematicians had already explored has not been found in the Chinese texts that survive).

Problems involving right-angled triangles. The so-called Pythagorean theorem is given in the *Nine Chapters.* Formulas are provided to solve problems on right-angled triangles such as the following: "given the base, and the sum of the height and of the hypotenuse, find the height and the hypotenuse." The diameter of the inscribed circle and the side of the inscribed square are given as well.

The Commentary of Liu Hui. Liu Hui's commentary on the *Nine Chapters* is the most important text dating from before the 13th century that contains proofs. He gives proofs for the formulas of volumes that are presented in the *Nine Chapters,* and he adds new formulas for the same volumes. He also organizes these formulas, given one after the other without comment in the *Nine Chapters,* into a system in which proofs for one formula use only formulas that have been established independently. Starting from the formula of the parallelepiped, Liu Hui proves the formulas of "standard blocks," from which he then deduces other formulas. He proceeds with a small set of proof techniques, including dissection, decomposition and recomposition into known pieces, and Cavalieri's principle (which states that, if two solids of the same height are such that their corresponding sections at any level have the same areas, then they have the same volume), and so on.

Proof techniques

Liu Hui compares the algorithms given in the algebraic sections of the *Nine Chapters* with one another and demonstrates how the same formal operations, which he calls the "key-steps" of computations, are brought into play in different algorithms. Thus, in the comparison between the addition of fractions and the solution of systems of simultaneous linear equations mentioned above, Liu Hui shows that sets of numbers are involved (numerator and denominator for a fraction, the coefficients of an equation for systems of equations) which share the property that all the numbers of a set can be multiplied by the same number without altering the mathematical meaning of that set. Both algorithms proceed by multiplying the sets of numbers that enter into a problem, each by an appropriate factor, in such a way that some corresponding numbers of the sets are made equal and the other numbers multiplied to keep intact the meaning of the whole

sets. (In the case of fractions, the denominators are made equal and the numerators changed appropriately; for systems of equations, the procedure is the same as if two numbers in the same row but in different columns were made equal by an appropriate multiplication, so that one of them can be eliminated through a column-to-column subtraction; the whole columns are then multiplied by the same number so that the equations remain true). Liu Hui proceeds from these analogies to state new algorithms for the same problems.

The Ten Classics of Mathematics. The *Ten Classics of Mathematics* contains explications of the mathematical knowledge presupposed by the *Nine Chapters* (the numeration system, arithmetic operations, etc.). Most of the subjects presented rely on the algorithms presented in the *Nine Chapters:* the rule of three, sharing in unequal parts, false double position, systems of simultaneous linear equations, and so on. These algorithms are not stated in as full generality in the *Ten Classics* as they are in the *Nine Chapters,* but they are given in the solutions of particular problems. Nevertheless, it is possible to see the ongoing evolution of some of these subjects, for example, the root extraction, the solution of equations, and the summation of series. The mathematical classics by Sun Tzu and Chang Ch'iu-chien (probably written before the 5th century) employ new ways of describing the algorithms of square root and cube root extraction. The underlying procedures are the same and they are still described in parallel ways, but this parallelism shows more clearly than before the mathematical object under which both of them are actually subsumed and which is thus responsible for their similarity: namely, equation. What has changed in the description is that, just as division involves a single divisor, square root extraction is shown to have two divisors, and cube root extraction three divisors. (These divisors actually are coefficients of the equations underlying the root extractions.) The divisors are shown to play similar roles in the algorithms. Moreover, in setting up the algorithms, the divisors are arranged one above the other, yielding a positional notation for the underlying equations: the row in which a number occurs is associated with the power of the unknown whose coefficient it is.

The *Ten Classics* contains discussions of topics that are not mentioned in the *Nine Chapters* but that were to be the subject of some of the highest mathematical achievements of the Sung and Yüan dynasties (960–1368). For example, the *Mathematical Classics by Sun Tzu* presents this congruence problem: Suppose one has an unknown amount of objects. If one counts them by 3s, there remain 2 of them. If one counts them by 5s, there remain 3 of them. If one counts them by 7s, there remain 2 of them. How many objects are there? The procedure used to solve the problem is difficult to understand, however, because it is described in a very condensed manner.

Congruence problems

Some major developments from the 11th century to the 13th century. *Theory of root extraction and of equations.* In the 11th century Chia Hsien is said to have given an algorithm to find a fourth root involving a method similar to the so-called Ruffini–Horner method. In this algorithm the successive numbers that are put in each of the rows (actually the coefficients) are obtained through computations that involve only the numbers located in the rows below. Again the algorithm makes use of the configuration given to this set of numbers in an essential way. In addition, the procedures used to compute the numbers of any row are basically the same. The representation of its computation still involves a positional notation for the underlying equations. This new algorithm can be applied for square and cube root extraction as well; it is only necessary to determine the number of rows (*i.e.,* coefficients) that each of these computations involves. It was again through a reformulation of the previous root extraction procedures that a new improvement was obtained: these procedures were shown to be special cases of the same general algorithm.

The 12th-century scholar Liu I explored a method of finding roots of quadratic equations that have positive or negative coefficients; such equations arise in geometric problems. These coefficients, whatever their sign, are entered in the table of the root extraction, and the square root algorithm is adapted to each situation.

Later, in Ch'in Chiu-shao's *Mathematics in Nine Chapters* (1247), a similar algorithm is used to find "the" root of any equation. By that time, general equations were used and were represented by a positional notation (see the example in the next section), which seems to indicate that it was the slow evolution of the algorithms of root extraction and their comparison that produced the concept of equation. Similar methods (with a slightly different notation) were well known to Li Yeh, and in his *Sea-Mirror of Circle Measurements,* written only one year after Ch'in completed his book, he takes the search for the root of equations for granted. Li Yeh lived in North China, while Ch'in Chiu-shao lived in the South, and he is thought to have worked without knowing Ch'in's achievement. It is thus highly probable that these methods were well known before the middle of the 13th century.

From Chia Hsien on, another method was known for finding the root of an equation, using the coefficients of what is now called Pascal's triangle (see Figure 22) and the same positional representation.

By permission of the Syndics of Cambridge University Library

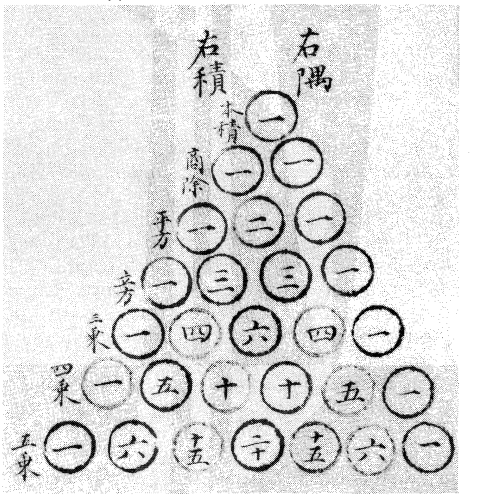


Figure 22: A Chinese representation of Pascal's triangle.

The method of the "celestial unknown." Li Yeh's book also contains a method, unknown to Ch'in Chiu-shao, which seems to have flourished in North China for some decades. This method explains how to use polynomial arithmetic to find equations to solve a problem. Li Yeh's book is the oldest surviving work that explains this method, but it was probably not the first to deal with it. In this book polynomials are also arranged according to a positional notation. Thus $x^2 - 3x + 5 + 7/x^2$ is represented as

$$\begin{matrix} 7 \\ 0 \\ 5 \\ -3 \\ 1 \end{matrix}.$$

A character is added next to the 5 to indicate that it is a constant term. The location of the coefficient indicates the power of the indeterminate with which it is associated. This indeterminate is called the celestial unknown.

It is known that some mathematicians used this representation for polynomials in two or three unknowns. In his book *The Jade Mirror of the Four Unknowns,* Chu Shih-chieh makes use of four unknowns. Starting from the centre, in the two horizontal and the two vertical directions, he puts in increasing order of their powers what comes from each of the four unknowns. In such problems, where there is more than one unknown, he has to use a method of elimination of a common unknown to two equations.

Polynomials in several unknowns

Indeterminate analysis. Ch'in Chiu-shao's book also contains algorithms for the general congruence problem, some examples of which are given in Sun Tzu's treatise, where its solution was too obscure to be understood. This

problem amounts to determining a number, the remainders of which are known when it is divided by given numbers (called moduli). There is no extant work between Sun's treatise and Ch'in's book of 1247 that reveals how this algorithm was elaborated. Such problems seem to have been worked out because of calendrical computation. Ch'in Chiu-shao introduces his discussion of these problems by saying that his goal is to clarify several procedures used by astronomers who were applying them without understanding them. His solution is known today as the Chinese remainder theorem. He deals with the case when moduli are relatively prime, and then reduces the case when they are not to it. The first case is easily solved when x can be found that satisfies the congruence $xa \equiv 1 \pmod{b}$, a and b being two given relatively prime numbers (suppose $a < b$). Ch'in gives an algorithm for this, using a sequence of quotients in searching for the greatest common divisor of a and b, which is also the sequence of convergents for the continued fraction for b/a. Having them, he is then able to compute x.

In addition, the summation of series was developed to a greater extent during this period.

The decline of the Sung-Yüan mathematics. Little is known about what happened after Chu Shih-chieh. In the 16th century a mathematician commenting on Li Yeh's *Sea-Mirror of Circle Measurements* no longer understood the method of the celestial unknown. By the 17th century it seems to have been completely forgotten. Rods were then no longer used as a counting tool, so perhaps Chinese algebraic positional notations, deprived of the instrument on which they were based, could not be understood. On the other hand, there was a rapid diffusion of the abacus, for which many books were written, including the *Systematic Treatise on Arithmetic* by Ch'eng Ta-wei (1592). This book explains in detail arithmetic on the abacus. Editions of this work remained popular until the 20th century.

When the Jesuits arrived in China at the end of the 16th century, they found people interested in science (so that they were accepted in China because of their scientific knowledge) but unaware of the Chinese past in mathematics. An era of translations of Western works started then (the six first books of Euclid's *Elements* were translated by a Jesuit and a Chinese in 1607), followed by a period when Chinese mathematicians attempted to find ancient books, to understand them, and to make a synthesis of the Chinese and Western traditions. In the 18th century, with the help of Western algebra, Mei Chüeh-ch'eng could understand again the ancient texts dealing with the method of the celestial unknown.

JAPAN IN THE 17TH CENTURY

The introduction of Chinese books. Very little is known of the history of Japanese mathematics before the 17th century. Through contacts with Chinese civilization, first mediated through Korea and then directly, there was, beginning in the seventh century, a flow of Chinese science to Japan. For example, the *Ten Classics of Mathematics* was introduced, along with the counting rods. Yet no Japanese book before the end of the 16th century is known to deal with mathematics. At that time another phase of importation began: the abacus and the *Systematic Treatise on Arithmetic* became known in Japan, where, however, they did not supplant the use of rods. Moreover, many books were taken from Korea to Japan, and perhaps in that way two Chinese books, the *Mathematical Treatise by Yang Hui* and the *Introduction to the Study of Mathematics* (written by Chu Shih-chieh in 1299), arrived in Japan. In those books, Japanese scholars could find the solution for systems of simultaneous linear equations, the search for the root of an equation according to the methods used in China in the 13th century, and applications of the method of the celestial unknown, although these were not easily understandable. Books about calendrical computations, containing mathematical knowledge, were also imported. Chinese mathematics greatly influenced the development of Japanese mathematics (for example, its algebraic orientation) and defined the context in which Japanese tradition opened to European mathematics.

Influence of Chinese texts

Later, at the beginning of the Tokugawa period (1603–1867), contacts with foreigners were limited to the trade with Chinese and Dutch boats through the harbour of Nagasaki. Some Chinese books, which may have contained Western knowledge, as well as Dutch books entered Japan secretly, but it is difficult to state how much, or what kind of, mathematical knowledge went through that channel.

The elaboration of Chinese methods. Although not the first mathematical book written in Japan, *Jingoki,* published in 1627 by Mitsuyoshi Yoshida, seems to be the first book that played an important part in the Japanese tradition. Inspired by the *Systematic Treatise on Arithmetic,* it described in Japanese the use of the soroban, an improvement of the Chinese abacus, and introduced some Chinese knowledge. Its many editions contributed to popularizing mathematics; most of the works on mathematics in Japan were written in Chinese and could not be widely read. In its enlarged edition of 1641, *Jingoki* introduced the method of performing computations with counting rods, which were no longer used in China. Moreover, Yoshida added "difficult problems" (*idai*), inspired by its Chinese source and left without solutions, which he recommended be posed to mathematicians. This initiated a tradition of challenges, reminiscent of those that took place in Europe during the Renaissance, that stimulated strongly the development of mathematics. In this context, in the 1650s, mathematicians, relying on counting-rod computations and looking for new methods of solution, began to understand by themselves the methods of Chinese algebra, as they were hinted at in the *Introduction to the Study of Mathematics,* which was reprinted in Japan in 1658.

These methods became a systematic tool for the solutions of any problem—what settled the algebraic framework for a mathematical development where equations and problems are essentially linked—and many books were published that introduced mathematicians to the method of finding the roots of equations or to the use of polynomial algebra to set down equations in order to solve problems. One of them, the *Kokon sanpoki* (1671), by Kazuyuki Sawaguchi, pointed out that "erroneous" problems could have more than one solution (equations could have more than one root) and left unanswered difficult problems involving simultaneous equations of the nth degree. The equations for their solutions were published in 1674 by Seki Kōwa (Takakazu), who was later referred to as the founder of the Japanese tradition of mathematics, called Wasan, that was based on Chinese mathematics. Seki founded a school of mathematics that became the most important one in Japan. As in the other schools, disciples had to keep the school methods secret, and only the best among them knew most of these methods. Only slowly did they publish their secrets, which hindered the free circulation of ideas (at this time, mathematics was widely practiced in Japan as a leisure activity) and which makes any attribution very difficult.

Wasan

Explanations about how to use Seki's equations to derive Sawaguchi's problems were published in 1685 by one of his disciples, Takebe Katahiro. Seki had designed for this purpose a "literal" written algebra using characters, thus liberating mathematicians from counting rods; he kept for equations the positional notation with respect to one unknown, the coefficients being expressed in terms of other unknowns. In establishing equations among several unknowns for the solution of a problem, he had to introduce procedures equivalent to computations of determinants in order to eliminate unknowns between simultaneous equations. Further research elaborated these procedures.

Seki devised a classification of problems that amounted to a classification of equations, which took into consideration negative roots and multiple roots, noticed by Sawaguchi; for this purpose he adapted the Chinese algorithms from the 13th century. Seki and his disciples thus improved upon Chinese methods in many ways, opening new directions for the development of mathematics in Japan (for example, in their work on infinite series, the subject of research by contemporary European scientists as well).

(K.C.C.)

BIBLIOGRAPHY

General sources: Two standard texts are CARL B. BOYER, *A History of Mathematics,* 2nd ed. edited by UTA C. MERZBACH

(1989); and, on a more elementary level, HOWARD EVES, *An Introduction to the History of Mathematics,* 5th ed. (1983). Discussions of the mathematics of various periods may be found in O. NEUGEBAUER, *The Exact Sciences in Antiquity,* 2nd ed. (1957, reissued 1969); MORRIS KLINE, *Mathematical Thought from Ancient to Modern Times* (1972); and BARTEL L. VAN DER WAERDEN, *Science Awakening,* 4th ed. (1975; originally published in Dutch, 1950). See also KENNETH O. MAY, *Bibliography and Research Manual of the History of Mathematics* (1973); and JOSEPH W. DAUBEN, *The History of Mathematics from Antiquity to the Present: A Selective Bibliography* (1985). A good source for biographies of mathematicians is CHARLES COULSTON GILLISPIE (ed.), *Dictionary of Scientific Biography,* 16 vols. (1970–80, reissued 16 vol. in 8, 1981). Those wanting to study the writings of the mathematicians themselves will find the following source books useful: HENRIETTA O. MIDONICK (ed.), *The Treasury of Mathematics: A Collection of Source Material in Mathematics* (1965); JOHN FAUVEL and JEREMY GRAY (eds.), *The History of Mathematics: A Reader* (1987); D.J. STRUIK (ed.), *A Source Book in Mathematics, 1200–1800* (1969); and DAVID EUGENE SMITH, *A Source Book in Mathematics* (1929; reissued in 2 vol., 1959). A study of the development of numeric notation can be found in GEORGES IFRAH, *From One to Zero* (1985; originally published in French, 1981).

Mathematics in ancient Mesopotamia: Editions of mathematical tablets include O. NEUGEBAUER (ed. and trans.), *Mathematische Keilschrift-Texte,* 3 vol. (1935–37, reprinted 3 vol. in 2, 1973); and F. THUREAU-DANGIN (ed. and trans.), *Textes mathématiques babyloniens* (1938). O. NEUGEBAUER and A. SACHS, *Mathematical Cuneiform Texts* (1945), is the principal English edition of mathematical tablets. A brief look at Babylonian mathematics is contained in the first chapter of ASGER AABOE, *Episodes from the Early History of Mathematics* (1964), pp. 5–31.

Mathematics in ancient Egypt: Editions of the basic texts are T. ERIC PEET (ed. and trans.), *The Rhind Mathematical Papyrus* (1923, reprinted 1970); A.B. CHACE *et al.* (eds. and trans.), *The Rhind Mathematical Papyrus,* 2 vol. (1927–29); and W.W. STRUVE (V.V. STRUVE) (ed.), *Mathematischer papyrus des staatlichen Museums der schönen Künste in Moskau* (1930). A brief but useful summary appears in G.J. TOOMER, "Mathematics and Astronomy," ch. 2 in J.R. HARRIS (ed.), *The Legacy of Egypt,* 2nd ed. (1971), pp. 27–54. For an extended account of Egyptian mathematics, see RICHARD J. GILLINGS, *Mathematics in the Time of the Pharaohs* (1972, reprinted 1982).

Greek mathematics: Critical editions of Greek mathematical texts include *The Thirteen Books of Euclid's Elements,* trans. by THOMAS L. HEATH, 2nd ed. rev., 3 vol. (1926, reprinted 1956); *The Works of Archimedes,* trans. by THOMAS L. HEATH (1897, reprinted 1953); E.J. DIJKSTERHUIS, *Archimedes,* trans. from Dutch (1956, reprinted 1987); THOMAS L. HEATH, *Apollonius of Perga: Treatise on Conic Sections* (1896, reissued 1961), and *Diophantus of Alexandria: A Study in the History of Greek Algebra,* 2nd ed. (1910, reprinted 1964); ROSHDI RASHED (trans.), *Les Arithmétiques* (1984–), of which vol. 3 and 4 contain Books IV–VII of Diophantus; and JACQUES SESIANO, *Books IV to VII of Diophantus' "Arithmetica" in the Arabic Translation Attributed to Qusṭā ibn Lūq̄* (1982). General surveys are THOMAS L. HEATH, *A History of Greek Mathematics,* 2 vol. (1921, reprinted 1981); JACOB KLEIN, *Greek Mathematical Thought and the Origin of Algebra* (1968; originally published in German, 1934); and WILBUR RICHARD KNORR, *The Ancient Tradition of Geometric Problems* (1986). Special topics are examined in O.A.W. DILKE, *Mathematics and Measurement* (1987); ÁRPÁD SZABÓ, *The Beginnings of Greek Mathematics* (1978; originally published in German, 1969); and WILBUR RICHARD KNORR, *The Evolution of the Euclidean Elements: A Study of the Theory of Incommensurable Magnitudes and Its Significance for Early Greek Geometry* (1975).

Mathematics in medieval Islām: Sources for Arabic mathematics include J.P. HOGENDIJK (ed. and trans.), *Ibn Al-Haytham's Completion of the Conics,* trans. from Arabic (1985); MARTIN LEVEY and MARVIN PETRUCK (eds. and trans.), *Principles of Hindu Reckoning,* trans. from Arabic (1965), the only extant text of Kūshyār ibn Labbān's work; MARTIN LEVEY (ed. and trans.), *The Algebra of Abū Kāmil,* trans. from Arabic and Hebrew (1966), with a 13th-century Hebrew commentary by Mordecai Finzi; DAOUD S. KASIR (ed. and trans.), *The Algebra of Omar Khayyam,* trans. from Arabic (1931, reprinted 1972); FREDERIC ROSEN (ed. and trans.), *The Algebra of Mohammed ben Musa,* trans. from Arabic (1831, reprinted 1986); and A.S. SAIDAN (ed. and trans.), *The Arithmetic of al-Uqlīdisī,* trans. from Arabic (1978). Islāmic mathematics is examined in J.L. BERGGREN, *Episodes in the Mathematics of Medieval Islam* (1986); E.S. KENNEDY *et al., Studies in the Islamic Exact Sciences* (1983); and ROSHDI RASHED, *Entre arithmétique et algèbre: recherches sur l'histoire des mathématiques arabes* (1984).

European mathematics during the Middle Ages and Renaissance: An overview is provided by MICHAEL S. MAHONEY, "Mathematics," in DAVID C. LINDBERG (ed.), *Science in the Middle Ages* (1978), pp. 145–178. A.P. JUSCHKEWITSCH, *Geschichte der Mathematik im Mittelalter* (1964; originally published in Russian, 1961), pp. 326–434, is the definitive modern work. Other sources include ALEXANDER MURRAY, *Reason and Society in the Middle Ages* (1978, reprinted 1985), ch. 6–8; GEORGE SARTON, *Introduction to the History of Science,* vol. 2, *From Rabbi Ben Ezra to Roger Bacon,* 2 parts (1931, reprinted 1975), and vol. 3, *Science and Learning in the Fourteenth Century,* 2 parts (1947–48, reprinted 1975); and, on a more advanced level, EDWARD GRANT and JOHN E. MURDOCH (eds.), *Mathematics and Its Applications to Science and Natural Philosophy in the Middle Ages* (1987). For the Renaissance, see PAUL LAWRENCE ROSE, *The Italian Renaissance of Mathematics: Studies on Humanists and Mathematicians from Petrarch to Galileo* (1975).

Mathematics in the 17th and 18th centuries: An overview of this period is contained in DEREK THOMAS WHITESIDE, "Patterns of Mathematical Thought in the Later Seventeenth Century," *Archive for History of Exact Sciences,* 1(3):179–388 (1961). Specific topics are examined in MARGARET E. BARON, *The Origins of the Infinitesimal Calculus* (1969, reprinted 1987); ROBERTO BONOLA, *Non-Euclidean Geometry: A Critical and Historical Study of Its Development* (1955; originally published in Italian, 1912); CARL B. BOYER, *The Concepts of the Calculus: A Critical and Historical Discussion of the Derivative and the Integral* (1930, reissued with the title *The History of the Calculus and Its Conceptual Development,* 1949, reprinted 1959); HERMAN GOLDSTINE, *A History of Numerical Analysis from the 16th Through the 19th Century* (1977); JUDITH V. GRABINER, *The Origins of Cauchy's Rigorous Calculus* (1981); I. GRATTAN-GUINNESS, *The Development of the Foundations of Mathematical Analysis from Euler to Riemann* (1970); ROGER HAHN, *The Anatomy of a Scientific Institution: The Paris Academy of Sciences, 1666–1803* (1971); and LUBOŠ NOVÝ, *Origins of Modern Algebra,* trans. from Czech (1973).

Mathematics in the 19th and 20th centuries: Surveys include HERBERT MEHRTENS, HENK BOS, and IVO SCHNEIDER (eds.), *Social History of Nineteenth Century Mathematics* (1981); WILLIAM ASPRAY and PHILIP KITCHER (eds.), *History and Philosophy of Modern Mathematics* (1988); and KEITH DEVLIN, *Mathematics: The New Golden Age* (1988). Special topics are examined in UMBERTO BOTTAZZINI, *The Higher Calculus: A History of Real and Complex Analysis from Euler to Weierstrass* (1986; originally published in Italian, 1981); JULIAN LOWELL COOLIDGE, *A History of Geometrical Methods* (1940, reissued 1963); JOSEPH WARREN DAUBEN, *Georg Cantor: His Mathematics and Philosophy of the Infinite* (1979); HAROLD M. EDWARDS, *Fermat's Last Theorem: A Genetic Introduction to Algebraic Number Theory* (1977); I. GRATTAN-GUINNESS (ed.), *From the Calculus to Set Theory, 1630–1910: An Introductory History* (1980); JEREMY GRAY, *Ideas of Space: Euclidian, Non-Euclidean, and Relativistic* (1979); THOMAS HAWKINS, *Lebesgue's Theory of Integration: Its Origins and Development,* 3rd ed. (1979); JESPER LÜTZEN, *The Prehistory of the Theory of Distributions* (1982); and MICHAEL MONASTYRSKY, *Riemann, Topology, and Physics,* trans. from Russian (1987).

Mathematics in China and Japan: Chinese mathematics is discussed in JOSEPH NEEDHAM, *Science and Civilisation in China,* vol. 3, *Mathematics and the Sciences of the Heavens and the Earth* (1959, reprinted 1970), pp. 1–168; *Ancient China's Technology and Science* (1983), a group of papers prepared by the Institute of the History of Natural Sciences, Chinese Academy of Science, in Peking; YAN LI (YEN LI) and SHIRAN DU (SHIH-JAN TU), *Chinese Mathematics: A Concise History,* trans. from Chinese (1987); ULRICH LIBBRECHT, *Chinese Mathematics in the Thirteenth Century* (1973); and LAY YONG LAM, *A Critical Study of the Yang Hui Suan Fa: A Thirteenth-Century Chinese Mathematical Treatise,* trans. from Chinese (1977). Useful journal articles include DONALD BLACKMORE WAGNER, "An Early Chinese Derivation of the Volume of a Pyramid: Liu Hui, Third Century A.D.," *Historia Mathematica,* 6(2):164–188 (May 1979), and "Liu Hui and Tsu Keng-Chih on the Volume of a Sphere," *Chinese Science,* 3:59–79 (1978). Besides *Historia Mathematica* (quarterly) and *Chinese Science* (irregular), many papers on Chinese mathematics may be found in *Archive for History of Exact Sciences* (8/yr.).

Overviews of Japanese mathematics include DAVID EUGENE SMITH and YOSHIO MIKAMI, *A History of Japanese Mathematics* (1914); YOSHIO MIKAMI, *The Development of Mathematics in China and Japan,* 2nd ed. (1974); and SHIGERU NAKAYAMA, "Japanese Scientific Thought," in CHARLES COULSTON GILLISPIE (ed.), *Dictionary of Scientific Biography,* vol. 15 (1978), pp. 728–758. *Historia Scientiarum: International Journal of the History of Science Society of Japan* (annual), contains many papers on Japanese mathematics.

(J.L.B./W.R.K./Me.F./C.G.F./J.J.G./K.C.C.)

Matrix Theory

The theory and solution of general systems of linear equations is facilitated by the introduction of entities called vectors and matrices. Vectors were originally introduced in order to interpret mathematically a physical quantity such as a velocity or force that has both a magnitude and an associated direction. A matrix is a rectangular array of numbers in a definite order, such as the array of coefficients of the unknowns in a system of linear equations. Rules of computation with vectors, which stem from their original physical interpretation, lead to closed systems of vectors called vector spaces, and matrices can be identified with special functions on these vector spaces called linear transformations. The theory of linear transformations of finite-dimensional vector spaces, which embraces the theory of matrices and that of systems of linear equations, constitutes the subject matter of linear algebra. (See also ALGEBRA: *Linear algebra;* ANALYSIS: *Vector and tensor analysis.*)

The article is divided into the following sections:

Historical and conceptual introduction 634
Matrices and arithmetic operations 634
  Addition
  Multiplication
  Transposition
Matrix transformations 635
  Row transformations
  Column transformations
Determinants 635
Solutions of systems of equations 635
  Homogeneous systems
  Nonhomogeneous systems
Matrices, polynomials, and eigenvalues 636
  Similar matrices
  Linear operators
  Diagonalizability
  Rational decomposition

Historical and conceptual introduction. Linear algebra originated in the study of systems of linear equations in several unknowns and in the attempt to find general methods for their solution. This attempt led quite early to the notion of a determinant, in 1693 by the German mathematician Gottfried Wilhelm Leibniz. In 1750 the Swiss physicist Gabriel Cramer gave the rule bearing his name, for expressing the solution of n linear equations in n unknowns in terms of determinants. The French mathematician Joseph-Louis Lagrange noted in 1773 the connection between a third-order determinant and the volume of a tetrahedron. Lagrange and Germany's greatest mathematician, Carl Friedrich Gauss, both observed that if the quadratic form $F = ax^2 + 2bxy + cy^2$ is subjected to a linear transformation (see Box, equations 1), expressed in two linear equations, to obtain the new form $f = a'u^2 + 2b'uv + c'v^2$ in the variables u and v, then the effect (see 2) of the transformation is to multiply the discriminant of the form (a function expressed in terms of the coefficients) by a factor depending on the transformation and not the forming.

The term matrix was first used by the British-born U.S. mathematician James Joseph Sylvester in 1850 to designate a rectangular array of numbers from which determinants may be formed. In 1855 Sir Arthur Cayley pointed out that a linear transformation (see 1) can be written, using matrix notation (see 3), and that, if two successive linear transformations (see 3, 4) are written this way, the composite (see 5) of the two transformations involves a third matrix, constructed by a definite rule from the first two, and having certain algebraic properties of a product. This discovery led him to consider matrices as separate entities and to define a matrix product in the way indicated by the composition of the linear transformations (see 6). He also gave the formula, using determinants, for the inverse of a matrix and made his often quoted, prophetic remark, "There would be many things to say about this theory of matrices which should, it seems to me, precede the theory of determinants." Cayley applied these ideas to the transformation of quadratic forms and, in 1858, formulated a complete algebra of matrices considered as mathematical entities in their own right. In 1853 William Rowan Hamilton, the noted Irish mathematician, had also used what was essentially the algebra of matrices under the name of linear and vector functions. Extensive development of the theory of matrices continued, and in the 20th century mathematicians began to focus their attention on the linear transformations rather than on the matrices that represent them. Linear algebra thus became the study of linear transformations of a vector space.

(1) $$\begin{cases} x = \alpha u + \beta v \\ y = \gamma u + \delta v \end{cases}$$

(2) $$a'c' - b'^2 = (\alpha\delta - \beta\gamma)^2(ac - b^2)$$

(3) $$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

(4) $$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \alpha' & \beta' \\ \gamma' & \delta' \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix}$$

(5) $$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha\alpha' + \beta\gamma' & \alpha\beta' + \beta\delta' \\ \gamma\alpha' + \delta\gamma' & \gamma\beta' + \delta\delta' \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix}$$

(6) $$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} \alpha' & \beta' \\ \gamma' & \delta' \end{bmatrix} = \begin{bmatrix} \alpha\alpha' + \beta\gamma' & \alpha\beta' + \beta\delta' \\ \gamma\alpha' + \delta\gamma' & \gamma\beta' + \delta\delta' \end{bmatrix}$$

Matrices and arithmetic operations. A matrix is a rectangular array (see 7) in which the individual entries a_{ij},

(7) $$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

called the elements of the matrix, belong to some fixed field $\mathfrak{F}$ of scalars or possibly to a commutative ring (see ALGEBRA: *Abstract algebra*) such as a ring of polynomials. If A has m rows and n columns, it is called an $m \times n$ matrix and is designated by $A = [a_{ij}]$ if m and n are assumed known. If $m = n$, A is a square matrix of order n and the elements $a_{11}, a_{22}, \cdots, a_{nn}$ are called the diagonal elements of A. If all elements of A other than the diagonal elements are 0, A is a diagonal matrix. If all elements below (or above) the main diagonal are 0, A is a triangular matrix.

Addition. If $A = [a_{ij}]$ and $B = [b_{ij}]$ are $m \times n$ matrices the sum $A + B$ is the matrix $[a_{ij} + b_{ij}]$ obtained by adding corresponding elements. A matrix all of whose elements are 0 is a zero matrix or null matrix and denoted by O. If O is an $m \times n$ zero matrix then $A + O = A$. If c is any scalar, then cA is the matrix $[ca_{ij}]$. With these definitions,

(8) $$d_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$$

(9) $$R = \begin{bmatrix} 0 \cdots 0 & 1 & c_{1j_1+1} \cdots 0 & \cdots & 0 \cdots c_{1n} \\ 0 \cdots 0 & 0 & 0 \cdots \cdots 1 & c_{2j+1} & 0 \cdots c_{2n} \\ & & & & 1 \cdots c_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 \cdots 0 & 0 & \cdots \cdots 0 & \cdots & 0 \cdots 0 \end{bmatrix}$$

(10) $$\det A = b_{i1}a_{i1} + b_{i2}a_{i2} + \cdots + b_{in}a_{in}$$

(11) $$A^{-1} = \frac{1}{\det A} C^T$$
in which $C = [c_{ij}]$ and $c_{ij} = (-1)^{i+j} \det A_{ij}$ in which A_{ij} is the matrix of order $n-1$ obtained from A by deleting the ith row and jth column.

the set $[\mathfrak{F}]_{m,n}$ of all $m \times n$ matrices, with elements in the field $\mathfrak{F}$ is a vector space over $\mathfrak{F}$. A basis for this vector space is the set of mn matrices $\{E_{ij}\}$ in which E_{ij} is the $m \times n$ matrix with 1 in the ith row, jth column, and 0 in every other place. Thus the dimension of $[\mathfrak{F}]_{m,n}$ is mn.

Multiplication. If $A = [a_{ij}]$ is an $m \times n$ matrix, and $B = [b_{ij}]$ is an $n \times r$ matrix, then the product AB is defined by the product rule (see 8) to be an $m \times r$ matrix $D = [d_{ij}]$. The product AB is defined only when the number of columns in A is equal to the number of rows in B. If $C = [c_{ij}]$ is an $r \times s$ matrix, then $(AB)C = A(BC)$. Thus matrix multiplication is associative, but not commutative since in general $AB \neq BA$. The $n \times n$ diagonal matrix I with diagonal elements all equal to 1 is called the identity or unit matrix of order n, since $IA = AI = A$ for any $n \times n$ matrix A. The distributive laws hold; that is, $A(B + C) = AB + AC$, and $(B+C)A = BA + CA$, whenever the indicated products and sums are defined. The set $[\mathfrak{F}]_n$ of all $n \times n$ matrices over $\mathfrak{F}$ is therefore a vector space over $\mathfrak{F}$, on which an associative multiplication is defined that is distributive with respect to addition. Such a system is called a linear associative algebra over $\mathfrak{F}$, or briefly, an algebra over $\mathfrak{F}$.

The transpose of a matrix

Transposition. If A is an $m \times n$ matrix, then the transpose of A is the $n \times m$ matrix A^T obtained by writing the rows of A as the columns of A^T. If B is an $n \times r$ matrix, $(AB)^T = B^T A^T$. A $1 \times n$ matrix is called a row vector and can be identified with a vector of $\mathfrak{U}_n(\mathfrak{F})$, because the vector space $[\mathfrak{F}]_{1,n}$ is isomorphic to $\mathfrak{U}_n(\mathfrak{F})$. Similarly, an $n \times 1$ matrix is called a column vector and its transpose is a vector of $\mathfrak{U}_n(\mathfrak{F})$. The row space of an $m \times n$ matrix A is the subspace of $\mathfrak{U}_n(\mathfrak{F})$ spanned by the row vectors of A. The column space of A is the subspace of $\mathfrak{U}_m(\mathfrak{F})$ spanned by the transposes of the column vectors of A. For any matrix, the row space and column space have the same dimension, called the rank of the matrix.

Matrix transformations. *Row transformations.* The solution of systems of linear equations and the determination of the rank of a matrix A are facilitated by the application to A of elementary row transformations. These are of three types: (1) interchange of two rows, (2) multiplication of one row vector by a nonzero scalar, (3) addition of a scalar multiple of one row vector to another. A matrix A is row-equivalent to B ($A \sim B$) if A can be transformed into B by a succession of elementary row transformations. Row-equivalence is reflexive ($A \sim A$), symmetric ($A \sim B$ implies $B \sim A$), and transitive ($A \sim B$ and $B \sim C$ imply $A \sim C$). Every matrix is row-equivalent to one and only one triangular matrix (see 9), characterized by four conditions: (1) all zero row vectors, if any, occur below all nonzero row vectors; (2) the first nonzero element in each row is 1; (3) if the first nonzero element in the ith row occurs in the j_ith column, every other element in the j_ith column is 0; and (4) $j_1 < j_2 < j_3 < \cdots < j_n$. The matrix R is called the reduced row-echelon form of A. Two matrices are row-equivalent if, and only if, they have the same reduced row-echelon form. Row equivalent matrices have the same row space and therefore the same rank. Because the nonzero row vectors of R are linearly independent, they constitute a basis of the row space of A, and the rank of A is equal to the number of nonzero rows in the reduced row-echelon form of A. An $n \times n$ matrix A is singular if its rank is less than n; nonsingular if its rank is equal to n. An $n \times n$ matrix is nonsingular if, and only if, its row vectors are linearly independent, and hence if, and only if, its reduced row-echelon form is the identity matrix.

An elementary matrix is a matrix obtained by applying an elementary row transformation to the identity matrix I. Every elementary row transformation of A can be achieved by multiplying A from the left by the elementary matrix obtained by applying the same elementary transformation to I. Every elementary matrix is nonsingular, and every nonsingular matrix is a product of elementary matrices. Thus, two $m \times n$ matrices A and B are row-equivalent if, and only if, there exists a nonsingular $m \times m$ matrix P such that $B = PA$. If A is a nonsingular $n \times n$ matrix, it is row-equivalent to I, and $PA = I$. In this case, the matrix P is unique and also satisfies $AP = I$. It is called the inverse of A, denoted by A^{-1}, and A is said to be invertible. An $n \times n$ matrix has an inverse if, and only if, it is nonsingular. The inverse, when it exists, is unique.

The matrix inverse

Column transformations. Elementary column transformations and column-equivalence of two matrices are defined analogously. Every elementary column transformation of A can be achieved by multiplying A from the right by an elementary matrix, and B is column equivalent to A if, and only if, there exists a nonsingular $n \times n$ matrix Q such that $B = AQ$. The matrix B is equivalent to A if A can be transformed into B by a succession of elementary row and/or column transformations; hence B is equivalent to A if, and only if, there exist nonsingular matrices P and Q such that $B = PAQ$. Equivalence is reflexive, symmetric, and transitive; and every matrix A of rank r is equivalent to a diagonal matrix in which the first r diagonal elements are 1, and all other elements are 0. This is called the canonical form of A. Two matrices are equivalent if, and only if, they have the same canonical form and hence if, and only if, they have the same rank.

Determinants. Associated with each $n \times n$ matrix A is a scalar called the determinant of A and denoted by $\det A$. The determinant has these basic properties: (1) If $A = [a_{ij}]$, $\det A$ is, for each i, a linear homogeneous function (see 10) of the elements of the ith row in which the coefficients depend only on the a_{kj} with $k \neq i$. (2) The interchange of two rows of A changes the sign of $\det A$. (3) $\det I = 1$. These properties actually determine $\det A$ uniquely. Other important properties of the determinant are: $\det A = 0$ if and only if A is singular; and for any two $n \times n$ matrices $\det AB = \det A \det B$. The determinant of a triangular matrix is equal to the product of the elements in its main diagonal. Elementary row transformations of type (3) leave the determinant of A unchanged, those of type (1) change the sign of $\det A$; and those of type (2) multiply $\det A$ by the same nonzero scalar by which one row of A is multiplied. Thus for any $n \times n$ matrix A, $\det A$ can be found by reducing A to its reduced row-echelon form R, keeping track of the changes in $\det A$ that result from the row transformations used, and then evaluating $\det R$ by finding the product of its diagonal elements. If $\det A \neq 0$, then A is invertible, and there is a formula (see 11) for the inverse of A in terms of determinants.

Solutions of systems of equations. Three distinct situations arise in attempting to solve general systems of linear equations, namely, a system may have a unique solution, infinitely many solutions, or no solution. These three possibilities can be illustrated by examples (see 12). A system of the first type is called an independent system; of the second type, a dependent system; and of the third type, an inconsistent system. Geometrically, in the case of two unknowns the first type represents two intersecting straight lines; the second type represents two coincident lines; and the third type represents two parallel lines.

A general system (see 13) of m equations with n un-

(12)
(a) $x-2y=6$, $2x+3y=5$. Solution: $x=4, y=-1$.
(b) $x-2y=6$, $2x-4y=12$. Solutions: Infinitely many.
(c) $x-2y=6$, $2x-4y=9$. Solutions: None.

(13)
$$\begin{cases} a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n=c_1 \\ a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n=c_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1+a_{m2}x_2+\cdots+a_{mn}x_n=c_m \end{cases}$$

(14)
$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & c_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & c_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & c_m \end{bmatrix}$$

(15) $x-5y+2z=0$ has independent solutions $S_1=[5,1,0]$ and $S_2=[-2,0,1]$ and every solution has the form $hS_1+kS_2=[5h-2k, h, k]$, in which h and k are arbitrary. Here $n=3$ and $r=1$. The nonhomogeneous equation $x-5y+2z=3$ has a solution $S_0=[2,1,3]$ and hence every solution of this equation is obtained from the formula $S_0+hS_1+kS_2=[2+5h-2k, 1+h, 3+k]$.

(16) For example, the augmented matrix of the system
$x-2y+2z+w=3$
$2x-2y+4z-2w=5$
$x-4y+2z+5w=4$
and its row-echelon form are
$$\begin{bmatrix} 1 & -2 & 2 & 1 & 3 \\ 2 & -2 & 4 & -2 & 5 \\ 1 & -4 & 2 & 5 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 & -3 & 2 \\ 0 & 1 & 0 & -2 & -\frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
and the general solution is therefore $x=-2z+3w+2$, $y=2w-\frac{1}{2}$, in which z and w are arbitrary. In vector form the general solution is $[2,-\frac{1}{2},0,0]+z[-2,0,1,0]+w[3,2,0,1]$, in which $z[-2,0,1,0]+w[3,2,0,1]$ is the general solution of the homogeneous system $AX=O$ and $[2,-\frac{1}{2},0,0]$ is the particular solution obtained by putting $z=w=0$.

knowns is said to be consistent if solutions exist. This system can be written in matrix form, $AX=C$, in which $A=[a_{ij}]$ is called the coefficient matrix of the system, and X and C are the column vectors $[x_1, x_2, \cdots, x_n]^T$ and $[c_1, c_2, \cdots, c_m]^T$, respectively. A vector X that satisfies this equation is called a solution vector of the system. A solution exists if, and only if, the vector C of constant terms is in the space spanned by the column vectors of A or if and only if the augmented matrix of the system (see 14), constructed from A by adjoining the column vector C, has the same rank as the coefficient matrix A. The system is said to be homogeneous if $C=O$, nonhomogeneous if $C\neq O$. Every solution of a consistent nonhomogeneous system $AX=C$ has the form S_0+S in which S_0 is any particular solution of $AX=C$, and S is a solution of $AX=O$. Conversely, if $AS_0=C$ and $AS=O$, then $A(S_0+S)=AS_0+AS=C$ and S_0+S is a solution of $AX=C$. If the number of unknowns is equal to the number of equations, then $AX=C$ has a unique solution if, and only if, A is non-singular, and this is given by $X=A^{-1}C$.

Homogeneous systems. A homogeneous system $AX=O$ is always consistent because it has the solution $X=O$. The set of all solutions of this system is a subspace of $\mathfrak{U}_n(\mathfrak{F})$, called the solution space of the system or the nullspace of the matrix A. The dimension of the solution space is $n-r$, in which r is the rank of A and n is the number of unknowns. This means that $AX=O$ has $n-r$ linearly independent solution vectors, and every solution is a linear combination of these. Examples (see 15) illustrate these facts and are easy to construct.

The general solution of a homogeneous system $AX=O$ can be obtained by reducing the matrix A to row-echelon form. This process is based on the principle that if two matrices A and B are row-equivalent then the systems $AX=O$ and $BX=O$ have the same solution space. If A has rank r, and R is the reduced row-echelon form of A, then R has a 1 in the ith row, j_ith column ($i=1, 2, \cdots, r$), and all other entries in the j_ith column are zero. Thus the system $RX=O$ can be solved immediately for the r unknowns $x_{j_1}, x_{j_2}, \cdots, x_{j_r}$ in terms of the remaining $n-r$ unknowns which may be assigned arbitrary values.

Nonhomogeneous systems. A nonhomogeneous system $AX=C$ can be solved in the same way. If R is the reduced row-echelon form of the augmented matrix, the first n columns of R provide the reduced row-echelon form R' of A. The system is consistent if, and only if, R and R' have the same number of nonzero rows. Again, provided the system is consistent its general solution can be read from the matrix R as in the homogeneous case. An example (see 16) illustrates these matters.

Matrices, polynomials, and eigenvalues. If A is an $n\times n$ matrix over $\mathfrak{F}$ and x is an indeterminate, the polynomial $f(x)$ (see 17) is called the characteristic polynomial of A. The roots in $\mathfrak{F}$ of the characteristic equation $f(x)=0$ are called the characteristic values or eigenvalues of A. The Cayley-Hamilton theorem states that every $n\times n$ matrix satisfies its characteristic equation (see 18). It follows that there exists a polynomial $m(x)$ of minimum degree $r\leq n$ that $m(A)=O$. If $m(x)$ is monic—*i.e.*, has leading coefficient 1—then it is unique and is called the minimal polynomial of A. It is necessarily a divisor of the characteristic polynomial $f(x)$. Moreover, every irreducible factor of $f(x)$ is also a factor of $m(x)$, although it may occur with a smaller multiplicity. If the field of scalars is algebraically closed—*e.g.*, if $\mathfrak{F}$ is the field of the complex numbers—then $f(x)$ has exactly n roots in $\mathfrak{F}$ if a root of multiplicity r is counted r times. Thus, every $n\times n$ matrix over the complex field has exactly n eigenvalues. If A has n distinct eigenvalues, then $m(x)=f(x)$.

Similar matrices. If A and B are $n\times n$ matrices, B is said to be similar to A if there exists a nonsingular matrix P such that $B=P^{-1}AP$. Like equivalence, similarity is a reflexive, symmetric, and transitive relation. Moreover, for a fixed matrix P, the mapping α defined by $\alpha(A)=P^{-1}AP$ is an isomorphism of the algebra $[\mathfrak{F}]_n$ with itself. Thus, for any polynomial $g(x)$ with coefficients in $\mathfrak{F}$, $P^{-1}g(A)P=g(P^{-1}AP)$. Similar matrices have the same characteristic polynomials, the same minimum polynomials, the same eigenvalues, and the same determinants.

If τ is a linear operator on an n-dimensional vector space $\mathfrak{U}$, and if $A_\tau=A_\tau^{(E)}$ is the matrix of τ relative to a fixed basis (E), then the matrix of τ relative to any other basis is similar to A_τ. Thus, the characteristic polynomial of A_τ can also be called the characteristic polynomial of τ. The minimal polynomial of τ—that is, the unique monic polynomial $m(x)$ of least degree such that $m(\tau)=O$—is also the minimal polynomial of $A_\tau^{(E)}$ and is therefore independent of the basis (E). The eigenvalues of τ are those scalars c for which the operator $\tau-cI$ has nonzero nullspace. These are precisely the eigenvalues of A_τ because

(17) $$f(x)=(-1)^n\det(A-xI)$$

(18) If $f(x)=x^n+a_1x^{n-1}+\cdots+a_0$ is the characteristic polynomial of A, then $f(A)=A^n+a_1A^{n-1}+\cdots+a_0I=O$

$A_{\tau-cI} = A_\tau - cI$. If c is an eigenvalue of τ, a nonzero vector V in the nullspace of $\tau - cI$ is called an eigenvector of τ belonging (or corresponding) to c. It has the property that $\tau V = cV$. A substance $\mathfrak{S}$ of $\mathfrak{U}$ is invariant under τ if τS is in $\mathfrak{S}$ whenever S is in $\mathfrak{S}$. Thus, an eigenvector of τ spans a one-dimensional subspace of $\mathfrak{U}$ that is invariant under τ. If V is an eigenvector of τ belonging to c, then the coordinate vector X of V relative to the basis $E_1, \cdots, E_n$ satisfies equation (19). Hence $A\tau^{(E)}X$ is equal to cX and X is called an eigenvector of the matrix $A\tau^{(E)}$ corresponding to c.

Linear operators. A vector space $\mathfrak{U}$ is a direct sum of subspaces $\mathfrak{S}_i$ (see 20) if $\mathfrak{S}$ is the sum of these subspaces and each of subspaces $\mathfrak{S}_i$ has zero intersection with the sum of the remaining subspaces. If $\mathfrak{U}$ is such a direct sum, then every vector V in $\mathfrak{U}$ has a unique representation as the sum (see 21) of vectors $\{S_i\}$, in which S_i belongs to $\mathfrak{S}_i$. If τ is a linear operator on a finite-dimensional space $\mathfrak{U}$ with minimal polynomial $m(x)$, and if the factorization (see 22) of $m(x)$ into irreducible factors over $\mathfrak{F}$ has the typical factor $[p_i(x)]^{e_i}$, then $\mathfrak{U}$ is the direct sum of the nullspaces $\{\mathfrak{S}_i\}$ of the operators (see 23) obtained from these factors by replacing x by τ. Moreover, each subspace $\mathfrak{S}_i$ is invariant under τ. In particular, if $\mathfrak{F}$ is algebraically closed, then $p_i(x) = x - c_i$ and $\mathfrak{S}_i$ is the nullspace of $(\tau - c_i I)^{e_i}$, in which $\{c_i\}$ are the distinct characteristic roots of τ, and e_i is the multiplicity with which c_i occurs as a root of $m(x)$. A basis (E) of $\mathfrak{U}$ can be chosen so that it is the ordered union of bases of the subspaces $\mathfrak{S}_i$. The matrix $A_\tau^{(E)}$ then takes the semidiagonal form (see 24) in which, for each i, the matrix A_i on the diagonal of $A_\tau^{(E)}$ has order equal to the dimension of $\mathfrak{S}_i$ and has minimal polynomial $[p_i(x)]^{e_i}$.

The minimal polynomial

Diagonalizability. A linear operator τ on $\mathfrak{U}$ is said to be diagonalizable if its eigenvectors span the space $\mathfrak{U}$; that is, if it has n linearly independent eigenvectors, and in which $n = \dim \mathfrak{U}$. A necessary and sufficient condition that τ be diagonalizable is that the minimal polynomial of τ factors in $\mathfrak{F}$ into distinct linear factors (see 25). Diagonalizability of τ implies that the matrix $A_\tau = A_\tau^{(E)}$ of τ, relative to any basis (E), has n linearly independent eigenvectors $X_1, X_2, \cdots, X_n$. If X_i is the eigenvector corresponding to the eigenvalue c_i, then $A_\tau X_i = c_i X_i$, and hence, if P is the matrix with column vectors $X_1, X_2, \cdots, X_n$, then $A_\tau P = PD$ in which D is the diagonal matrix (see 26) with the eigenvalues c_i along the diagonal. Since $X_1, \cdots, X_n$ are linearly independent P is nonsingular, and hence $P^{-1}A_\tau P = D$, and $A\tau$ is similar to a diagonal matrix. Conversely, the equation $P^{-1}A_\tau P = D$ implies that $A\tau$, and therefore τ, has n linearly independent eigenvalues.

Thus, diagonalizability of an operator means that its matrix, relative to any basis, is similar to a diagonal matrix, and that $\mathfrak{U}$ is the direct sum of n one-dimensional subspaces, each invariant under τ.

Rational decomposition. If τ is a linear operator in a nonzero n-dimensional space $\mathfrak{U}$ and V is a nonzero vector in $\mathfrak{U}$, then there is a unique monic polynomial $g(x)$ of minimum degree such that $g(\tau)V = O$, and $g(x)$ is called the τ-annihilator of V. If $g(x) = c_0 + c_1 x + \cdots + c_{r-1}x^{s-1} + x^s$, then the vectors $V, \tau V, \tau^2 V, \cdots, \tau^{s-1}V$ are linearly independent and span a subspace $\mathfrak{C}(V, \tau)$, called the τ-cyclic subspace generated by V. The subspace $\mathfrak{C}(V, \tau)$ is invariant under τ and the linear operator induced by τ in $\mathfrak{C}(V, \tau)$ has a matrix relative to the basis $V, \tau V, \cdots, \tau^{s-1}V$ of a simple form (see 27) which is called the companion matrix of the polynomial $g(x)$. The minimum polynomial of S is $g(x)$. The rational decomposition theorem states that there exist r nonzero vectors $\{V_i\}$ in $\mathfrak{U}$ with τ-annihilators $\{g_i(x)\}$ such that $g_{i+1}(x)$ divides $g_i(x)$ for $i = 1, 2, \cdots, r-1$, and $\mathfrak{U}$ is the direct sum (see 28) of subspaces $\{\mathfrak{C}(V_i, \tau)\}$. Moreover, the number r and the τ-annihilators $\{g_i(x)\}$ are uniquely determined by these conditions and do not depend on the choice of the V_i. A basis (B) of $\mathfrak{U}$ can therefore be chosen (see 29) of a form such that relative to this basis the matrix of τ (see 30) will have the matrices $\{C_i\}$ along the diagonal and zeros elsewhere. Each C_i here is the companion matrix of $g_i(x)$. Such a matrix is said to be in rational canonical form. Thus every $n \times n$ matrix over $\mathfrak{F}$ is similar to a unique matrix in rational canonical form and two matrices are similar over $\mathfrak{F}$ if and only if they have the same rational canonical form.

The rational canonical form of a matrix

If the field $\mathfrak{F}$ is algebraically closed, the minimal polynomial (see 31) of τ is the product of factors of the type $(x - c_i)^{e_i}$, in which $\{c_i\}$ are the distinct eigenvalues of τ, and $\mathfrak{U}$ is the direct sum of the nullspaces $\mathfrak{C}_i$ of the operators $(\tau - c_i I)^{e_i}$. If τ_i is the operator induced in $\mathfrak{C}_i$ by τ, and if $\sigma_i = \tau_i - c_i I$, then $\tau_i = c_i I + \sigma_i$, in which σ_i has the minimal polynomial x^{e_i}. Thus, if a basis for $\mathfrak{C}_i$ is chosen in accordance with the rational decomposition of $\mathfrak{C}_i$ into cyclic subspaces relative to the operator σ_i, then the matrix J_i of τ_i relative to this basis has a form (see 32) with matrices $J_k^{\,i}$ along the diagonal, zeros elsewhere, and in which each matrix $J_k^{\,i}$ has a simple triangular form (see 33) and the order of $J_k^{\,i}$ is greater than or equal to that of $J_{k+1}^{\,i}$ for each k. The union of these bases for the subspaces $\mathfrak{C}_i$ form a basis for $\mathfrak{U}$ relative to which the matrix of τ

(19) $$(A_\tau^{(E)} - cI)X = O$$

(20) $$\mathfrak{U} = \mathfrak{S}_1 \oplus \mathfrak{S}_2 \oplus \cdots \oplus \mathfrak{S}_r$$

(21) $$V = S_1 + S_r + \cdots + S_r$$

(22) $$m(x) = [p_1(x)]^{e_1}[p_2(x)]^{e_2} \cdots [p_r(x)]^{e_r}$$

(23) $$[p_i(\tau)]^{e_i}$$

(24) $$A_\tau^{(E)} = \begin{bmatrix} A_1 & O & \cdots & O \\ O & A_2 & & O \\ \cdots & \cdots & \cdots & \cdots \\ O & O & \cdots & A_r \end{bmatrix}$$

(25) $$\begin{cases} m(x) = (x - c_1)(x - c_2) \cdots (x - c_r), \text{ in which} \\ c_1, c_2, \cdots, c_r \text{ are the distinct eigenvalues of } \tau. \end{cases}$$

(26) $$\begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & & \\ \cdots & \cdots & \cdots & \\ 0 & 0 & & c_n \end{bmatrix}$$

(27) $$C = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & 0 & \cdots & 0 & -c_2 \\ & & & & & \\ 0 & 0 & 0 & \cdots & 1 & -c_{r-1} \end{bmatrix}$$

(28) $$\mathfrak{U} = \mathfrak{C}(V_1, \tau) \oplus \mathfrak{C}(V_2, \tau) \oplus \cdots \oplus \mathfrak{C}(V_r, \tau)$$

(29) $$V_1, \tau V_1, \cdots, \tau^{s_1 - 1}V_1, \cdots, V_r, \tau V_r, \cdots, \tau^{s_r - 1}V_r$$

(30) $$A_\tau^{(B)} = \begin{bmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ 0 & 0 & \cdots & C_r \end{bmatrix}$$

(31) $$m(x) = (x - c_1)^{e_1}(x - c_2)^{e_2} \cdots (x - c_r)^{e_r}$$

(32) $$J_i = \begin{bmatrix} J_1^{(i)} & 0 & \cdots & 0 \\ 0 & J_2^{(i)} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & J_{n_i}^{(i)} \end{bmatrix}$$

$$(33)\qquad \begin{bmatrix} c_i & 0 & 0 & \cdots & 0 & 0 \\ 1 & c_i & 0 & \cdots & 0 & 0 \\ 0 & 1 & c_i & \cdots & 0 & 0 \\ & & & & & \\ 0 & 0 & 0 & \cdots & c_i & 0 \\ 0 & 0 & 0 & \cdots & 1 & c_i \end{bmatrix}$$

$$(34)\qquad A_r = \begin{bmatrix} J_1 & 0 & 0 \\ 0 & J_2 & 0 \\ \cdots & \cdots & \cdots \\ 0 & 0 & J_r \end{bmatrix}$$

has the form (34) with the matrices J_i along the diagonal and zeros elsewhere. Such a matrix is said to be in Jordan canonical form. Every matrix A over an algebraically closed field is similar to a matrix in Jordan canonical form that is unique, if the diagonal blocks J_i are ordered so their eigenvalues increase with i. Two matrices are similar if and only if they have the same Jordan canonical form.

BIBLIOGRAPHY. RICHARD BELLMAN, *Introduction to Matrix Analysis* (1960), a comprehensive treatment of matrix theory and its applications to differential equations, probability theory, and mathematical economics; SIR ARTHUR CAYLEY, *The Collected Mathematical Papers* (1897), mainly of historical interest; F.R. GANTMACHER, *Theory of Matrices,* vol. 1 and 2 trans. from Russian by K.A. HIRSCH (1959), a standard textbook, advanced level; C.C. MACDUFFEE, *The Theory of Matrices,* 2nd ed. (1946), a concise but comprehensive account of matrix theory up to 1932 including copious references to original sources; and *Vectors and Matrices* (1943), a comparatively elementary discussion of similarity and the derivation of the rational and Jordan canonical forms of a matrix; THOMAS MUIR, *Theory of Determinants in the Historical Order of Development,* 2nd ed., 4 vol. (1906–23), an exhaustive history of the development of determinants to 1900; S. PERLIS, *Theory of Matrices* (1952), a standard textbook, intermediate level; O. SCHREIER and E. SPERNER, *Einführung in die analytische Geometrie und Algebra,* vol. 1 (1959; Eng. trans. of vol. 1, *Introduction to Modern Algebra and Matrix Theory,* 2nd ed., 1959), a standard textbook, intermediate level; J.H.M. WEDDERBURN, *Lectures on Matrices* (1934, reprinted 1964), a comprehensive account of matrix theory, including linear associative algebras, up to 1933—this book includes brief historical notes and also a bibliography from 1853 to 1933, which is arranged chronologically. Modern texts include HOWARD W. EVES, *Elementary Matrix Theory* (1980); KI H. KIM, *Boolean Matrix Theory and Applications* (1982); and ALAN JENNINGS, *Matrix Computation For Engineers and Scientists* (1980).

(Ed.)

Matter: Its Properties, States, Varieties, and Behaviour

All matter consists of atoms or molecules or a mixture of both. Matter can occur in various states or phases; the three most familiar are solid, liquid, and gas. Certain other states, though less clearly definable, are recognized; the most significant of these is plasma, an electrically conducting medium generally composed of negatively charged electrons, positively charged ions, and atoms or molecules or both.

Each state of matter has characteristics that distinguish it from all others and that are therefore independent of any given substance under study. The gaseous state, for instance, possesses properties simply as a gas, but if a substance in this state is condensed to the liquid state, it will have the distinctly different properties of a liquid. The most familiar example is water, whose properties as ice, liquid, and vapour are dramatically different within the temperature and pressure ranges of everyday experience.

This article focusses on the properties, structural features, and behaviour of the four major states of matter mentioned above. Due attention is given to the subdivisions of these general phases and to the variable forces that produce changes between the phases. The article also touches on colloidal systems and high-pressure and low-temperature phenomena, such as superconductivity and superfluidity. For additional information about the basic constituents and structure of matter, see the articles SUBATOMIC PARTICLES; ATOMS; MOLECULES; and CHEMICAL ELEMENTS. See also PHYSICAL PRINCIPLES AND CONCEPTS for a discussion of the relationship of matter and energy. (Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 124 and 125.

This article is divided into the following sections:

Phase changes and equilibria 640
  General considerations 640
    Phase number
    Component number
    The variables of a system
  Phase diagrams 641
    Unary systems
    Binary systems
    Ternary and higher order systems
Solid state 645
  Crystalline solids 646
    Classification based on the type of bonding
    Structure of crystalline solids
    Crystallization and crystal growth
  Amorphous solids 662
    General properties
    Structure of glass
    Specific properties of glass
  Mechanical properties of solids 666
    Elastic and plastic properties
    Thermal properties
  Electrical and magnetic characteristics 667
    Conductors
    Semiconductors and insulators
    Magnetism in solids
  Interaction of radiation with matter 679
    Behaviour of solids illuminated with radiation
    Generation of electromagnetic radiation
Liquid state 681
  The nature and structure of liquids at equilibrium 681
    Behaviour of liquids at equilibrium
    Molecular structure of liquids
    Pair distribution function
  Mechanical and thermal properties of liquids 684
    Transport properties
    Propagation of sound in liquids
    Electrical and magnetic properties
    Absorption and scattering of radiation
  Solutions and solubilities 685
    Classes of solutions
    Properties of solutions
    Thermodynamics and intermolecular forces in solutions
    Theories of solutions
    Solubility of solids and gases
Gaseous state 694
  General properties of gases 694
  Thermodynamic approach to gases 694
    The simple gas laws
    Thermal equation of state
    Empirical equations for specific gases
  The particle-description approach to gases 697
    Distribution functions
    Forces affecting distribution function
  Boltzmann transport equation 698
    Equilibrium solution for the Boltzmann equation
    Link between thermodynamic and microscopic properties
    Nonequilibrium solution for the Boltzmann equation
    Transport coefficients
    Distribution function for N particles
  Special conditions 701
    Hydrodynamic properties of a gas
    High temperatures
    Chemical reactions in gases
Plasma state 702
  General concepts 703
    Plasma properties and parameters
    Elastic and inelastic collisions
    Radiation from plasmas
  Basic plasma physics 704
    Plasma formation
    Effects of magnetic and electric fields
    Determination of plasma variables
    Fluid-like behaviour
    Applications of plasmas
  Natural plasmas 707
    Extraterrestrial forms
    Solar-terrestrial forms
    Terrestrial plasma phenomena
Colloids 708
  History of colloid studies 708
  Irreversible systems 709
    Lyophobic sols
    Emulsions
    Foams
    Pastes
    Gel structures
  Reversible systems 712
    Polymers
    Proteins and nucleic acids
    Gels
    Association colloids
Low-temperature phenomena 715
  Superconductivity 715
    Thermal properties of superconductors
    Magnetic and electromagnetic properties of superconductors
    Higher temperature superconductivity
  Superfluidity 719
  Special methods for obtaining low temperatures 719
    Adiabatic cooling
    Adiabatic dilution
High-pressure phenomena 720
  Physical and chemical effects of high pressure 720
    Effects on melting
    Effects on polymorphism
    Effects on minerals
    Effects on magnetic and electrical properties
  Applications of high pressure 722

PHASE CHANGES AND EQUILIBRIA

When matter in one form is transformed into another form, a phase change is said to take place. Whether the sample consists of a complicated mixture of many substances or of a single substance, the transformation may be a reversible one. To state the simplest case, in a reversible system, if form *A* is changed into form *B* by raising the temperature, form *B* will be transformed into form *A* by lowering the temperature, assuming that all other factors are constant. If both forms are present, the temperature will determine which form changes into the other. Because the energy of an atom is distributed over a wide spectrum, the whole of a sample of a solid substance, for example, will not be transformed instantly to liquid at the melting point. In fact, while some of the molecules or atoms are changing from solid to liquid, others will be changing from liquid to solid. In any time interval, the amount of phase *A* changing into phase *B* can be compared to the amount of phase *B* changing into phase *A*, and the proportions can be expressed in terms of the actual number of molecules involved—a quantity that can be calculated by using atomic weights of the elements and actual weights of the substances. It is possible, at a given temperature, to find that both *A* and *B* forms persist indefinitely in a fixed proportion. That is, the rate of change of *A* into *B* equals the rate of change of *B* into *A*, thus creating a system of phases that is in equilibrium. If even the slightest amount of heat is added to the system, the proportions will change until the rate in one direction again equals the rate in the other. Equilibrium frequently is an unwanted condition because it prevents reactions from going to completion and, for example, lowers the yields of desired products. Phase changes are of many kinds, and systems in equilibrium are also of various types. The few variable forces that bring about phase changes in any system include temperature, pressure, and concentration of substances.

General considerations

PHASE NUMBER

Concept of system

When one substance or a mixture of two or more substances in a given proportion is brought to equilibrium at a specified temperature and pressure, a phase or combination of such phases of matter will result. Such a system, as it is called, is described by enumerating the phases that are present, the composition but not the quantity of each phase, and the temperature and pressure of the system. A phase is defined as a homogeneous, mechanically separable, portion of a system, all parts of the phase being identical in intensive properties; *i.e.*, those qualities that characterize a phase, such as density and refractive index, but do not depend on the actual quantity of phase present. Although a phase must be homogeneous, it need not be continuous: crushed ice, for example, is a single phase since the intensive properties of one piece of ice are indistinguishable from those of every other piece. A system consisting of one substance throughout may have one, two, or three phases, but a complex system may have many liquid and solid phases, but normally only one gas phase because gases mix in all proportions. Thus water, held at 25° C under a pressure of one atmosphere, will consist entirely of liquid water, but if the temperature is raised above 100° C, or 212° F (the pressure still being maintained at one atmosphere), it will consist only of vapour. Liquid and vapour phases can coexist only at 100° C when the pressure is one atmosphere. If approximately equal weights of the compound diethyl ether and water are mixed and brought to equilibrium at 25° C and one atmosphere, they will not mix completely and two liquid phases will result, one of which is 1.3 percent water while the other is 94 percent water. Similarly, under appropriate conditions, one can form, by mixing common salt (formula NaCl) and water (formula H_2O), a system consisting of four phases: a vapour (H_2O), a liquid (NaCl dissolved in H_2O), and two solids, pure water as ice and a compound of salt and water with twice as many water molecules as salt (formula $NaCl \cdot 2H_2O$).

The equilibria in the aforementioned systems are essentially physical, but chemical equilibria are common, as in the system formed from solid calcium carbonate ($CaCO_3$) that, under suitable conditions, partly breaks down into gaseous carbon dioxide (CO_2) and solid calcium oxide (CaO); the system consists of two solid phases—one being calcium carbonate and the other calcium oxide—and one gas phase carbon dioxide, in proportions that remain constant until temperature, pressure, or volume are changed. All such equilibria are included in the term phase equilibria. The study of phase equilibria concerns the problem of determining what phases coexist at equilibrium under what conditions of temperature and pressure for a given combination of substances in given proportion. Equilibrium involving more than one phase is called heterogeneous equilibrium. The number of phases in a system is designated as beta (symbol β).

COMPONENT NUMBER

Distinction between components and substances present

The number of components, or the order, of the system, is designated as alpha (symbol a). Alpha is not necessarily the same as the number of substances used to form the system but is the smallest number of independently variable (explained directly below) substances required to account for the algebraic composition of each phase present. In this sense, a system containing only water has only one component, there being no recognition of the presence of molecular species other than water, such as the hydronium ion (H_3O^+) and the hydroxide ion (OH^-), because, although both ions are always present, their concentrations are not independently variable. In the abovementioned ether–water example, the system has two components, ether and water, because both are independently variable and both are needed to describe the composition of the two phases formed. In the system formed from calcium carbonate, however, even though there are three substances present, there are only two components, calcium oxide and carbon dioxide, because the composition of the $CaCO_3$ phase can be described as 50 percent CaO and 50 percent CO_2, that of the CaO phase as 100 percent CaO, 0 percent CO_2, and that of the gas phase as 0 percent CaO, 100 percent CO_2, these being mole percentages (one mole of a substance is its molecular weight in grams). Other selections of components could be made—*e.g.*, CaO and $CaCO_3$—but these would require negative percentages. In another example, barium sulfate ($BaSO_4$) and sodium carbonate (Na_2CO_3) react when melted together to produce barium carbonate ($BaCO_3$) and sodium sulfate (Na_2SO_4), all solids. This system involves four substances: $BaSO_4$, Na_2CO_3, $BaCO_3$, and Na_2SO_4, related through the reaction $BaSO_4 + Na_2CO_3 \rightleftharpoons BaCO_3 + Na_2SO_4$, but the system has only three components because the composition of all four can be described in terms of percentages of only three of them. It is, however, not the identity of the components themselves but the actual number of components that is of importance in the discussion immediately below.

The value of alpha corresponding to a given combination of substances can depend on the range of conditions over which it is studied. Thus, a system of oxygen (molecules containing two atoms, O_2) and ozone (molecules containing three atoms of oxygen, O_3) in any proportions has two components under conditions in which neither is converted to the other, but becomes a one-component system when they interact according to the equation $3O_2 \rightleftharpoons 2O_3$. When the reaction is at equilibrium, according to the laws of chemical reaction the concentrations of O_2 and O_3 are no longer independently variable; *i.e.*, one cannot change without the other also changing. A system consisting of solid ammonium chloride (NH_4Cl) in equilibrium with its vapour has one component, even if the vapour dissociates into ammonia (NH_3) and hydrochloric acid (HCl), provided the NH_3 and HCl are present in equal molar quantities. If, however, excess of either NH_3 or HCl is present, alpha will equal two. Systems with alpha (a) = one, two,

three, or four are said to be unary, binary, ternary, or quaternary, respectively.

THE VARIABLES OF A SYSTEM

Degrees of freedom. The description of any one system is facilitated by a statement of the number of degrees of freedom, or its variance, symbolized as F, that it possesses. This is the (minimum) number of variables of temperature (symbol T), pressure (symbol P), and concentration that must be fixed in order to define the system—*i.e.*, in order to fix the intensive properties of all the phases. At equilibrium, there is only one pressure variable because all the phases are at the same total pressure; there is only one temperature variable because all the phases are at the same temperature. The maximum number of independent concentration variables for each phase will be $a-1$ because fixing $a-1$ concentrations automatically fixes the concentration of the remaining component. In a system containing only water vapour, for example, the properties of the vapour will not be defined until both T and P are fixed, so that F equals two. There is no concentration variable when a equals one. For a system containing liquid water and water vapour in equilibrium, only one of the two variables, T and P, can be fixed arbitrarily—fixing one fixes the other—therefore F equals one. If ice, liquid water, and water vapour are all present, no variables can be fixed arbitrarily—the latter are beyond any control—and therefore $F=0$. For a system consisting of water vapour and liquid water saturated with solid salt (NaCl), fixing any one of the following will define the system: T, P, concentration of salt in the liquid solution, concentration of water in the liquid solution; F is therefore unity. (The concentrations of NaCl in the H_2O vapour and of H_2O in the solid NaCl are not determinative variables because experimentally none of them is present.)

Phase rule. The value of F is readily calculated by means of a rigorous thermodynamic equation known as the phase rule, which states that the number of degrees of freedom equals the number of components minus the number of phases, plus two, or $F=a-\beta+2$. Systems for which F equals zero, one, two, or three are invariant, univariant, bivariant, and trivariant, respectively.

The phase rule was derived by a U.S. physicist, J. Willard Gibbs, in some papers on thermodynamics published from 1873 to 1878, and represents an important milestone in the history of science. Its usefulness for chemists in the systematization of heterogeneous equilibria was brought to light two or three decades later. Its validity is restricted to systems in real equilibrium and assumes that the intensive properties of every phase of a system are determined only by its temperature, pressure, and composition. The influence of such factors as surface exposure, electric, magnetic, and gravitational fields is assumed to be negligible.

Phase diagrams

Derivation of phase diagrams

The subject of phase equilibria depends on graphical representation. With any system, the first step of experimental study is to establish what phases result under a variety of specific conditions by bringing together various proportions of the given substances. It is not yet possible to predict such information, at least in a quantitative sense. The results of the experiments are then plotted and embodied in phase diagrams that are graphical representations in two or more dimensions of the relationships among the variable factors.

Because phase changes are so much a part of the physical world, phase diagrams are pertinent to everyday existence: the melting of ice, the boiling of water, the formation of fog, the setting of cement, for example, are well-known phase changes, the details of which can be predicted from phase diagrams. In the laboratory, processes such as extraction, crystallization, distillation, precipitation, the use of freezing mixtures, and the identification of substances all involve phase changes describable by phase diagrams. Many industrial operations—such as the manufacture of ceramics, alloys, glass, and salts—depend on knowledge of phase diagrams. Finally, geological studies of the history of minerals in igneous, sedimentary, and metamorphic rocks in Earth and Moon surfaces are based on previously established phase equilibria.

While phase diagrams ideally describe only systems at equilibrium, the attainment of equilibrium is often a very slow process. In the interpretation of phase behaviour, therefore, the application of phase diagrams must be made with caution and an awareness that time may be an additional important factor.

UNARY SYSTEMS

When only one component is present, there are considered to be no concentration variables, every possible phase being 100 percent of that component. The two variables of pressure and temperature suffice to determine the state of the system. Thus all possible phase relations can be shown by a graph in which pressure is plotted against temperature (see Figure 1). Since the number of components alpha (a) equals one, and the number of degrees of freedom (F) equals two, one, and zero, by the phase rule, when the number of phases, beta (β) equals one, two, and three, respectively, systems with only one phase are bivariant (have two independent variables) and are represented by areas and defined by the two variables; two-phase systems are univariant (have one independent variable) and are represented by lines; three-phase systems are invariant (have no variables) and are represented by points. The lines represent the limits of the condition of bivariance. Three such lines (*OA, OB, OC*) meet at an invariant point, such as point *O*, known as a triple point. The substance thus will consist of a single liquid phase if pressure and temperature are such that the plotted point lies in the area partly enclosed by *COA*, designated liquid; the substance will consist of a single solid phase if the plotted point lies in the area partly enclosed by *BOC;* similarly, if the plotted point lies in the area *AOB* the substance will exist as a single vapour phase. Liquid and vapour phases can coexist only when pressure and temperature are such as to lie on *OA*, the vaporization curve. Solid and vapour phases can coexist only along *OB*, the sublimation curve (*i.e.*, the solid phase transforms reversibly into the gas without passing through a liquid phase), and solid and liquid can coexist only along *OC*, the fusion curve. Solid, liquid, and vapour can coexist only at the unique pressure and temperature of point *O*. *AO* can be extended for a short distance into the region for solid and, as such, represents a metastable (less stable) liquid–vapour equilibrium that, on being disturbed, yields the more stable solid–vapour equilibrium. The vaporization curve terminates at *A*, the critical point, beyond which liquid is no longer distinguishable from vapour, a condition called the supercritical fluid state. The sublimation and fusion lines, on the other hand, are often found to terminate in other triple points. *OA* gives the variation of the vapour pressure of the liquid with temperature, *OB* the variation of the sublimation pressure of the solid with temperature, and *OC* the variation of the melting point of the solid with pressure. The one-atmosphere dotted line intersects *OC* at temperature T_m, and *OA* at temperature T_b, T_m and T_b being, respectively, the normal melting and boiling points. (If, as with carbon dioxide, point *O* is above the one-atmosphere line, there are no such points—only a normal sublimation point.) *OC* is nearly vertical, because the melting point is

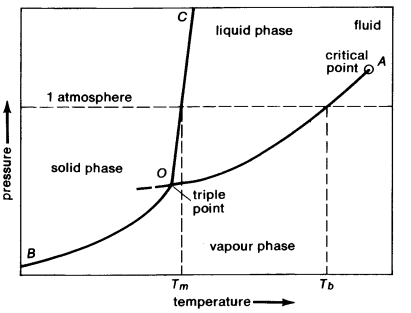


Figure 1: Phase equilibria for the solid, liquid, and vapour states of a substance (see text).

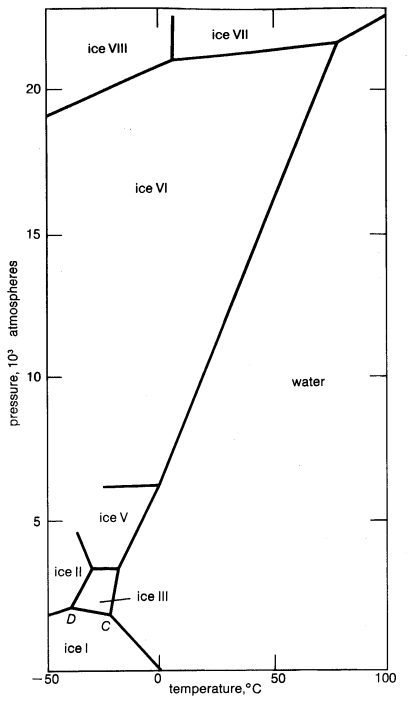


Figure 2: Equilibrium relations for high-pressure forms of ice.

Thermodynamics of univariant transition lines

only slightly sensitive to changes in pressure. The slopes of any of these univariant transition lines at any point thereon are governed by thermodynamic principles embodied in the Clausius–Clapeyron relation, which states that if ΔH is the heat absorbed when a given mass of one of the two phases is transformed to the other phase (under equilibrium conditions at that point), and ΔV is the corresponding increase in volume, then the ratio of a small change in pressure (dP), to a small change in temperature (dT) will equal the heat absorbed divided by the product of the temperature and the volume change, thus: $dP/dT = \Delta H/T\Delta V$. For transitions along OA and OB, ΔH and ΔV have the same sign, so the lines have a positive slope. This is also true usually along OC, but for exceptional cases, including water, ΔH and ΔV have opposite signs and, therefore, dP/dT is negative. Thus the melting point of ice is lowered by an increase in pressure. Point O for water is at 4.58 torr (one torr equals the pressure of 1/760 of the atmosphere at standard conditions) and 0.0100° C (32.0180° F). (The melting point of ice, measured in the presence of air at ambient or normal atmospheric pressure, is 0.00° C [32.00° F], because the pressure is about 760 torr and because air is dissolved in the water.)

Polymorphic substances. Many substances are polymorphic—*i.e.*, they exist in more than one crystalline form—a property well illustrated by water. When such crystalline forms are stable under a given pressure and temperature, each will be represented by an area, as in Figure 2, that shows several polymorphic forms of ice (the various forms being identified by roman numerals) and is an extension to higher pressures of Figure 1 as applied to the substance water. Two such forms occupying adjacent areas may be said to be enantiotropic; *i.e.*, they are directly interconvertible and can coexist stably at pressures and temperatures anywhere on the line separating the two areas. In contrast, some substances, such as phosphorus, exist in solid forms (*e.g.*, white phosphorus), that are not known to be stable under any conditions; on standing, they change spontaneously to a more stable form (*e.g.*, white phosphorus changes to red phosphorus). Such metastable forms are said to be monotropic, but the terms enantiotropic and monotropic are both chiefly of historic interest. Figure 2 shows the variety of ice forms existing under appropriate conditions of temperature and pressure. Ice IV, known only for heavy water, is metastable and is not shown. Ice VII has a density of nearly 1.7 grams per cubic centimetre and requires pressures of at least 21,700 atmospheres to be stable.

Other well-known substances exhibiting polymorphism are sulfur, carbon, iron, tin, silica, carbon tetrachloride, *p*-dichlorobenzene, ammonium chloride, and potassium nitrate. Rhombic and monoclinic sulfur are stable below and above 95.5° C (203.9° F), respectively, at one atmosphere pressure, but this temperature (at which they coexist) increases with pressure along a line similar to a fusion line and called a transition line. Crystalline carbon is known in both graphite and diamond forms, each having its own area of stability and bearing to each other a relation similar to that of ice I and ice III (Figure 2), in which ice I is replaced by graphite and ice III by diamond. With carbon, however, DC slopes upward (instead of down as with water) to the right, rising from approximately 22,000 atmospheres at room temperature to 200,000 atmospheres at 4,000° C (7,200° F). Thus diamonds can be formed from graphite only under immense pressures. (The temperature must also be high enough for an appreciable transition rate.) Diamonds remain unchanged at ambient pressure and temperature only because their conversion to graphite is extremely slow.

Liquid crystals. Finally, mention must be made of certain compounds, such as *p*-azoxyanisole, the solid form of which melts sharply at one temperature to give a milky, "liquid crystal" or mesomorphic phase that is fluid but anisotropic (with different properties in different directions) and, at a higher temperature, becomes clear and isotropic (uniform in all directions). Such substances, known since 1888, have only recently become of wide interest because of their unique properties.

BINARY SYSTEMS

Variables of binary systems

When two components are present, there are, in addition to the pressure and temperature variables, two concentration variables, only one of which is independent because fixing the percent of one component in a phase fixes that of the other. Thus, the complete description of the phase behaviour requires a three-dimensional model, referred to three axes for P, T, and composition, respectively. A typical portion of such a model for a system of two volatile liquid components, A and B, that are mutually soluble (miscible) in all proportions, is shown in Figure 3. Along

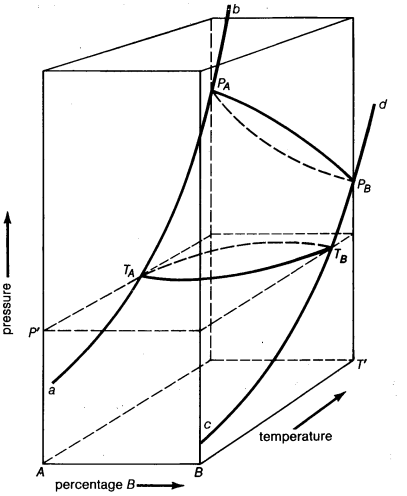


Figure 3: Portion of a typical solid model for a binary system.

AB is plotted the concentration or composition in weight or mole percent (points A and B representing pure A and B, respectively). Because alpha equals two, then F equals three, two, one, or zero according as beta equals one, two, three, or four. Systems consisting of a single phase are thus represented in the model by volumes. If one regards Figure 3 as depicting a system that, within a limited range of pressure and temperature, shows a liquid phase and vapour phase or both, then ab on the A face of the model is part of the line OA of Figure 1 for component A, and cd is part of the line OA of Figure 1 for component B. Within the model lie two curved surfaces that meet along ab and cd and that, between ab and cd, envelop a region of space

such that a section (slice) through the model at any temperature T', called an isothermal (uniform temperature) section, gives a lenticular area (P_A to P_B). The space (volume) below the lower surface represents systems consisting entirely of vapour and containing both A and B, and the volume above the upper surface systems consisting entirely of liquid containing both A and B. The space between the surfaces represents systems consisting of both liquid and vapour phases. Every point on the upper surface is joined by an imaginary line to a corresponding point on the lower surface, these lines—called tie lines—being parallel to the AB axis. The two points that terminate each tie line give the values of pressure, temperature, and composition of the two coexisting phases. Because beta equals two in this region, F equals two. Regions of space crossed by such tie lines may be called ruled volumes. If Figure 3 covered a wider range of temperature and pressure, then other such pairs of curved surfaces could appear. Intersection of such pairs give ruled surfaces, which, in turn, intersect along lines parallel to AB, the details of which are available in extended works on the subject.

It is usually more convenient to consider sections of the solid model rather than the model itself. Figure 4, left, is an isobaric section of Figure 3 taken at pressure P', implying

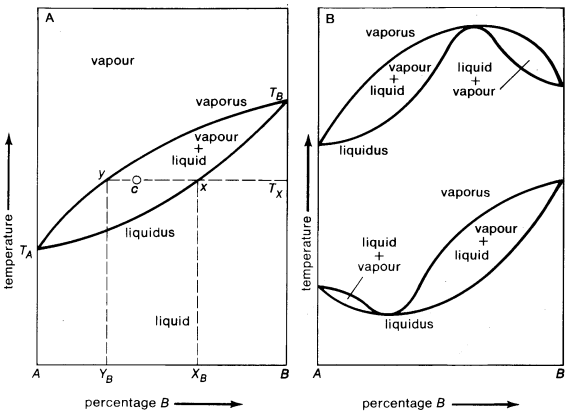


Figure 4: Three types of binary liquid-vapour relations at constant pressure (boiling-point diagrams); the curves in B have maxima and minima, respectively (see text).

that the experimenter has fixed the pressure (throughout the system) at the value P', conveniently one atmosphere. What was a volume in Figure 3 becomes an area in Figure 4, left, and what was a surface becomes a line. Thus, what were trivariant volumes are now areas, labelled "vapour" and "liquid"; what was a ruled volume is now a ruled area, marked "vapour + liquid," crossed by tie lines such as yx joining every point on the upper line (vaporus) to a corresponding point on the lower line (liquidus). With pressure fixed arbitrarily (isobaric) in this way, and composition and temperature variable, one-phase systems are said to be isobarically bivariant, two-phase systems isobarically univariant, etc. An isobarically invariant system, obtained by taking an isobaric section of the appropriate solid model (not shown) is represented by a (three-phase) tie line to be illustrated later (Figure 6 [left]). In like manner, one may draw isothermal sections of Figure 3, such as Figure 5 taken at temperature T'. P_A and P_B are the vapour pressures of A and B, respectively, at T'. The similarity of Figures 4, left, and 5 is evident. Such qualitative relationships between isotherms and isobars are common.

Simplest binary system. Figure 3 represents the simplest possible type of binary system and gives only the liquid–vapour relations. The labelling of the areas of the typical isobaric section, Figure 4, is self-explanatory. (T_A and T_B are boiling points of A and B.) In a closed system, an overall composition such as that of c, held at temperature T_X and pressure P', will yield a liquid of composition X_B and a vapour of composition Y_B. The ratio of the intercepts yc/cx on the tie line yx, gives the ratio of the quantity of liquid to the quantity of vapour phase. This principle is valid for all two-phase tie lines. If P' is the barometric pressure, the lower curve gives the boiling points of all solutions at that pressure. A solution of composition X_B boils at temperature T_X, and the vapour it gives has composition Y_B. Thus, the vapour is richer in A than the liquid. Continued removal of vapour from the boiling liquid causes the temperature and liquid and vapour compositions to change following the liquidus and vaporus lines, respectively, to the right. Such boiling-point diagrams are pertinent to fractional distillation of liquids. In general terms, most liquid solutions do not continue to boil at a fixed temperature but, because the vapour is richer in the more volatile components than the liquid, the temperature rises toward the boiling point of the least volatile component. Simultaneously both the remaining liquid and the vapour being formed become richer in the less volatile component. By proper arrangement of cooling apparatus, the vapours can be separated into samples or fractions such that each fraction has a different ratio of components. Usually the aim is to isolate one of the components, and, therefore, a specific fraction richest in that component will be distilled again, further reducing the unwanted components. This process of separating liquids by distilling them and trapping fractions for further distillation is fractional distillation.

Separation of liquids

Frequently, the solid model is such that its isobaric sections have the appearance of Figure 4, right (top and bottom), which show maxima or minima in the curves. At such unique points the two curves touch tangentially, and the liquid solutions distill unchanged in composition. This phenomenon is called azeotropy. Examples of liquid pairs belonging to this type, shown in Figure 4, left and right (top and bottom), are, respectively, methyl alcohol–water, acetone–chloroform, ethyl alcohol–water. Azeotropy prevents complete separation of a solution into its pure components by simple fractional distillation.

The comparative simplicity of Figures 3, 4, and 5 results from the assumed miscibility of liquids A and B and of their vapours. Similar diagrams result in considering equilibria between solid and liquid phases if solid A and solid B are miscible (continuous series of solid solutions) and if liquid A and liquid B are miscible. The previous discussion can be made pertinent by replacing the words liquid by solid, vapour by liquid, boiling by melting, and distillation by crystallization wherever they occur. Figure 4 then shows possible melting-point diagrams at constant pressure. The constant-pressure requirement is usually unimportant in melting-point diagrams.

More complex systems. A common solid–liquid relationship is one in which two components, A and B, have no miscibility as solids but are completely miscible as liquids. If the solids are not very volatile, an isobaric section at ambient pressure often has the appearance of Figure 6 (left), in which T_A and T_B are the respective melting points of A and B. The labelling of the various areas indicates, as before, what phases result when known mixtures are brought to equilibrium at a given temperature. For example, a mixture of composition X_B at temperature T_X will yield a liquid of composition l and solid B, lT_X being a tie line. The line T_BE is thus the solubility curve of solid B. A solution of composition l begins to freeze at

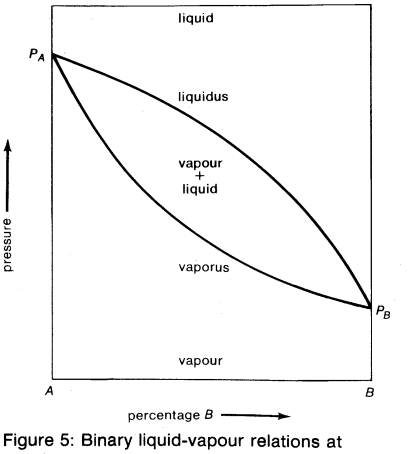


Figure 5: Binary liquid-vapour relations at constant temperature.

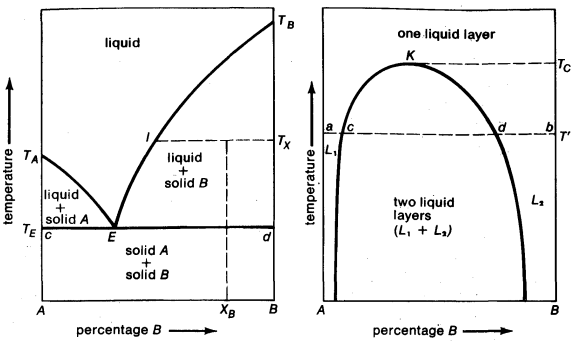


Figure 6: (Left) Binary system with eutectic and pure components as solid phases. (Right) System showing partial miscibility with upper critical solution point; L_1 = solution of B in A; L_2 = solution of A in B.

temperature T_X. With further cooling, solid B continues to separate, and the liquid composition follows the line lE to E, whereupon the liquid solidifies without change in temperature to give an intimate mixture of solids A and B. Thus the line T_BE also shows the lowering of the freezing point of B by A. T_AE has an analogous meaning. The line cEd is a (three-phase) tie line where three two-phase areas come together and represents the coexistence of the three phases solid A, solid B, and liquid E. Because alpha equals two and beta equals three, and because pressure is constant, the system is isobarically invariant and therefore defines itself as long as all three phases are present. Hence the temperature is fixed at T_E and the liquid composition at E. Any attempt to dislodge the system from this state will be resisted if equilibrium is maintained. For example, if heat is added in an attempt to raise the temperature, solid A and solid B will melt and dissolve in each other to give more liquid of composition E, and this process, which is accompanied by the absorption of heat, will nullify the external heat input. Such a reversible phase reaction (which can be represented by the equation solid A + solid $B \rightleftharpoons$ liquid E) at an invariant point in a binary system, characterizes a system said to be at a (binary) eutectic point. (This statement applies even when solids A and B are solid solutions instead of pure substances.) The mixture of solids A and B in the proportions required by this equation is called the eutectic mixture, and it melts as though it consisted of one substance. Temperature T_E is the lowest temperature at which a liquid phase can form. Liquid, and solids A and B, can coexist at, and only at, temperature T_E. With vapour present, the system is at a quadruple point. The creation of a constant low-temperature environment by mixing the solids ice and salt is based on this eutectic phenomenon, as is also the use of low-melting alloys. Thus, the temperature of a mixture of solid potassium iodide and ice cannot rise above −23° C (−9° F), which is the T_E for this system, as long as both solids are present. The use of salt to melt ice on the streets, the purification of water by partial freezing, the use of sharpness of melting as a partial criterion of purity of a substance, the process of zone refining, and the determination of mixed melting points are all related to diagrams of the type represented by Figure 6 (left).

Ice–salt mixture

Between the extremes of complete miscibility and complete immiscibility lies partial miscibility, as exemplified by the liquids ether and water mentioned earlier. Liquids A and B of Figure 6 (right) are in this category. Addition of B to A at temperature T' means following the line $acdb$. With small additions, only one layer is present (a to c); with larger additions, two layers (points c and d) result. Total compositions very rich in B again give only one layer (d to b). With rise in temperature, the miscibility gap (cd) often closes so that above T_C complete miscibility exists. Point K is an upper critical solution point. Less commonly, the gap closes with falling temperature, and in some examples it closes in both directions. Analogously, two solids may form solid solutions showing complete miscibility at one temperature and partial miscibility (discontinuous series of solid solutions) at another. The solids sodium chloride and potassium chloride, for example, are completely miscible above 500° C (900° F), yet only partially so below this temperature.

Complex variations. Many systems in their solid–liquid or liquid–vapour relations combine the features just described. Thus a boiling-point diagram can involve partial miscibility of the liquid components with or without azeotropy. A melting-point diagram can combine partial miscibility of the liquids with immiscibility or with partial miscibility of the solids. Melting in systems with partial miscibility of the solid components may show eutectic behaviour, similar to that discussed above, or peritectic behaviour, outlined below. One or both of the components can be polymorphic. Solid, liquid, and vapour phases may appear in the same diagram. Supercritical phenomena may be encountered. Intercomponent compounds may form. Some of these are referred to below.

Components A and B, themselves frequently compounds, may form addition compounds (C in Figure 7). In Figure 7 (left), which resembles two figures in juxtaposition—each

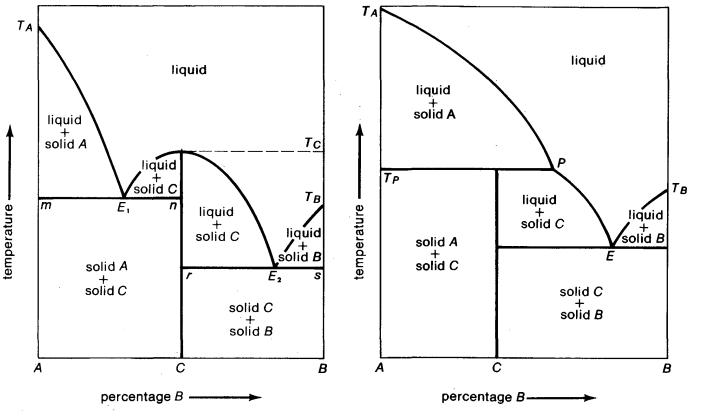


Figure 7: *Systems with intercomponent compounds* (C). (Left) Congruently melting compound. (Right) Incongruently melting compound.

similar to Figure 6 (left)—compound C melts congruently at T_C; *i.e.*, it melts to a liquid of the same composition. Lines mE_1n and rE_2s both represent eutectic points. In Figure 7 (right), C melts at T_P to a liquid, P, of a different composition and solid A; *i.e.*, C melts incongruently at T_P. The coexistence of solid A, solid C, and liquid P at temperature T_P, represented by the horizontal line at P, is an invariant system at constant pressure. Under these conditions, the system is said to be at a (binary) peritectic point (*cf.* eutectic point), characterized by the reversible phase reaction that can be written in equation form: solid $C \rightleftharpoons$ solid A + liquid P, which occurs on the addition or removal of heat and maintains the temperature at a constant value. The system consisting of copper and zinc, which includes ordinary brass, shows a series of intermetallic compounds and solid solutions; the system iron–carbon, which includes steel, shows the polymorphism of iron, compound formation, and solid solution.

TERNARY AND HIGHER ORDER SYSTEMS

With a third component present, there are a maximum of four independent variables. It is thus necessary, for graphical representation, either to keep one variable constant or to ignore the effect of one of the variables (the one to which the system is least sensitive, commonly the pressure). This leaves the temperature and two independent concentration variables to be plotted as a solid model. In the absence of displacement reactions (in which parts of two components change places with each other in chemical reaction), the latter usually consists of a triangular prism as in Figure 8 (left), the base of which is equilateral. Composition is plotted on triangular coordinates as illustrated in Figure 8 (right), in which point P represents the composition 50 percent A, 30 percent B, 20 percent C (either in weight or in mole percent). Temperature is measured vertically upward. Figure 8 (left) shows a typical solid–liquid relationship—that of three immiscible solids that form no compounds and that melt at T_A, T_B, T_C to give liquids that are completely miscible. Each of the

Variables in ternary systems

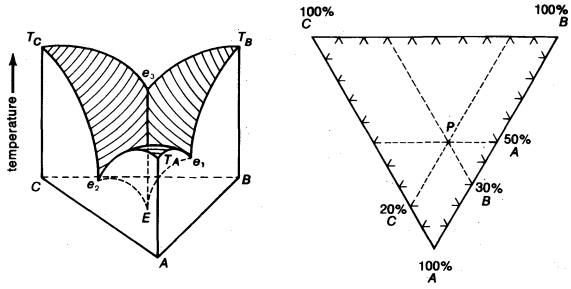


Figure 8: (Left) Freezing-point relations for a ternary system with pure components as solid phases. (Right) Method of plotting ternary composition with triangular coordinates.

three binary systems (*A–B, A–C, B–C*), which comprise the ternary, then shows the phase behaviour of Figure 6 (left) as seen on the vertical faces of Figure 8 (left). There are three curved surfaces sloping downward and inward, which intersect along the lines e_1E, e_2E, and e_3E, where e_1, e_2, and e_3 are the eutectic liquids of the respective binary systems. These surfaces give the (initial) freezing points of all solutions, the order in which the solids separate on cooling, and other details. Point *E*, a ternary eutectic liquid, gives the composition and temperature of the first liquid to form on heating, or last to remain on cooling. The cumulative effect on the lowering of the minimum temperature of melting through the addition of more components is evident. The curved surfaces may also be regarded as solubility surfaces. Figure 8 (left) is purposely simple, but other systems can be complicated by the presence of phenomena such as formation of compounds, partial or complete miscibility in the solid state, partial miscibility in the liquid state, ternary peritectic behaviour, presence of a vapour phase, etc. An extensively studied example is the complex system calcium oxide–aluminum oxide–silicon dioxide, $CaO–Al_2O_3–SiO_2$, because of its relation to the setting of portland cement.

When both pressure and temperature are held constant—*e.g.*, conveniently at one atmosphere and 25° C (77° F)—the phase relationships can be portrayed by a horizontal (isothermal) section of the triangular prism. The resulting plane triangular diagrams are widely used and illustrated in Figure 9. In this figure, for the system

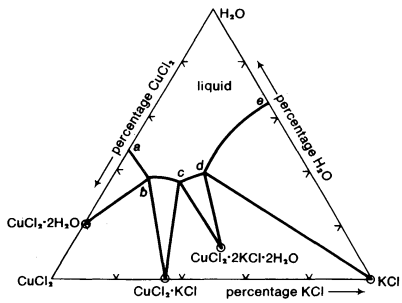


Figure 9: Phase diagram of $CuCl_2 · KCl · H_2O$ at 75° C (167° F).

copper(II) chloride–potassium chloride–water; *i.e.*, $CuCl_2–KCl–H_2O$, at 75° C (167° F), the lines *ab, bc, cd,* and *de* are sections of curved surfaces of the prism (not shown) and give the compositions of liquids saturated with solid $CuCl_2 · 2H_2O$, $CuCl_2 · KCl$, $CuCl_2, · 2KCl · 2H_2O$ and KCl, respectively. Because tie lines join points representing coexisting phases, every point on the curved edge of each wedge-shaped region is joined by an imaginary tie line to the corresponding opposite corner of the wedge. Such diagrams are useful in determining, for example, the conditions under which these solids can be prepared at 75° C. Diagrams similar to Figure 8 (left) form the basis for purification by recrystallization where, for example, *C* is the substance to be purified, *B* is the impurity to be removed, and *A* is the solvent.

Figure 10 is for three nonreacting liquids, *A, B,* and *C,* in which the pairs *A–B* and *A–C* are miscible but the pair *B–C* is only partially so. Total compositions anywhere above the "binodal curve," *cK'b,* give two layers. For example, composition *P* gives layers of compositions *m* and *n*, respectively, *mn* being a tie line, several of which are shown.

Three nonreacting liquids

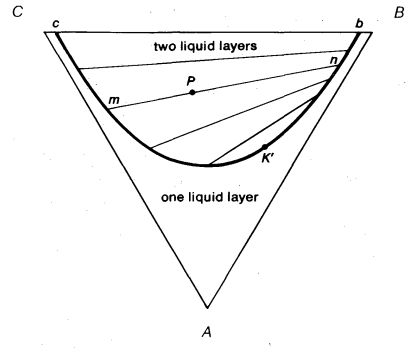


Figure 10: Typical ternary liquid system at constant temperature.

Below *cK'b,* only one phase is possible. As drawn, the slope of the tie lines indicates that there is more *A* in the *C*-rich layer than in the *B*-rich one. *K'*, where the tie lines have shrunk to a point, is called a plait point. Addition of sufficient *A* to the partially miscible *B–C* pair produces complete miscibility. Such systems are important in extraction processes.

Isothermal ternary diagrams often combine the phenomena typified by Figures 9 and 10. They may also involve the complications mentioned earlier. Systems with more than one binodal curve and systems with solid solutions have also been widely studied. When double-displacement reactions occur among the components a triangular diagram is not adequate.

Systems of four, five, and more components have been investigated but none completely. A historic study by a Dutch-born chemist, J.H. van't Hoff, of the solubility relations of the chlorides and sulfates of sodium, potassium, magnesium, and calcium in water was undertaken in connection with the utilization of the famous Stassfurt (now in East Germany) salt deposits and the determination of their geological history. (N.O.S.)

SOLID STATE

Solids may be divided into crystalline (*e.g.*, sodium chloride) or amorphous (*e.g.*, glass), into metallic (*e.g.*, copper), ionic (*e.g.*, sodium chloride), covalent (*e.g.*, germanium), or molecular (*e.g.*, solid argon), etc. These classifications are not exact. The properties of the solids depend on the properties of the atoms that form the solid, on the arrangement of atoms in the solid, and on the conditions (such as temperature and pressure) under which the solid is examined.

The wide variety of behaviour exhibited by solids is a direct result of the type of bonding that holds the atoms of the solid together and of the structural arrangement of the atoms. Classification of solids based on their structures can be developed at various levels of magnification, ranging from the scale of centimetres to angstroms (10^{-8} centimetre). As might be expected, the earliest structural studies were based on features that could be seen by the unaided eye. Thus, speculations on atomism by the ancient Greeks (*e.g.*, Democritus, at about 400 BC) were based on observations of the textures of ceramic materials and visual observations of the fractured surfaces of ceramics and bronzes. In the 19th century the etching of metals with acids and their subsequent examination with low-magnification optical microscopes was extensively carried out and revealed microstructural features of solids on the scale of what is presently known as the grain structure

(about 10^{-3} centimetre). This work was part of a progression of much earlier work by metalsmiths in many parts of the world, who revealed the macrostructure of objects such as swords and armour by differentially polishing, scraping, or etching with naturally acidic liquids. These processes revealed the distribution of hard and soft constituents of the metals and gave rise to the beautiful and remarkable patterns on the Japanese and Damascus swords.

Early recognition of solid structure

Parallel to the developing understanding of the microstructure of solids, mineralogists had been examining and classifying crystals of naturally occurring minerals. By the 19th century many of the crystallographic features of crystals and their symmetry properties were known as a result of such studies, since the shapes of natural crystals often reflect the atomic arrangements. It was not, however, until the discovery of X-ray diffraction from crystalline solids in 1912, following a suggestion by Max von Laue, a German physicist, that the atomic structure of crystals began to be understood. By this date the atomic nature of matter was accepted, and it was the patterns produced by the diffraction of X-rays from the regular arrangement of atoms in crystals that resulted in an understanding of the structure of solids on an atomic scale (about 10^{-8} centimetre). Resolution by different methods of examination has improved, until today, with the use of the electron microscope and the field ion microscope, individual atoms in a solid can be resolved and studied. As understanding of the structure of solids has progressed, it has become clear that imperfections in the arrangements of atoms in solids have a major effect on the properties of solids, and a wide variety of these imperfections have been studied.

The solid state of matter cannot be understood without knowledge of the atomic structure of matter. A brief summary of this knowledge is provided below.

General view of atomic structure

In most general terms, each of the chemical elements consists of a characteristic type of atom that is composed of subatomic particles arranged in a specific fashion. The nucleus of every atom consists of neutrons, particles with no electrical charge, and protons, particles with a single positive charge; this positively charged nucleus is surrounded by electrons, particles with a single negative charge, arranged in a structure according to the force of attraction between opposite charges and the force of repulsion between similar charges. The lightest element, hydrogen, has only one proton in the nucleus and a single electron; the heaviest naturally occurring element, uranium, has 92 protons and 92 electrons; between these are 90 other elements. The number of neutrons also increases, but not in such an orderly manner. The electrons around a nucleus may be thought of as occupying positions in a series of concentric shells. Some of these configurations are less stable than others, and the distribution of energy is such that the electrons tend to arrange their positions to attain the most stable arrangements. Thus, some atoms tend to donate electrons and others to accept them in chemical reactions; frequently, an exchange produces ions, which may then form one of the chemical bonds that hold together compounds. Some atoms may share electrons in a chemical reaction, a bonding process in the formation of compounds. The electrons involved in chemical reactions are called the valence electrons, and they may occupy one or more of the outer shells. Changes in the states of matter also involve bonds among atoms through their electron structures.

Crystalline solids

CLASSIFICATION BASED ON THE TYPE OF BONDING

Role of energy in formation of solids

As previously noted, solids form from either the liquid or the vapour state, because the energy of the atoms (more properly, the thermodynamic variable called free energy) decreases when the atoms take up a relatively ordered, three-dimensional structure. The decrease in energy on forming the solid from the vapour phase is termed the cohesive energy and results from two types of interactions: a repulsive action (or term) that prevents the atoms from moving too close together and an attractive, or bonding, term that stabilizes the crystal against separation of the atoms. The origin of these energy terms serves as one basis for classifying solids into five groups: ionic, covalent, metallic, molecular, and hydrogen-bonded solids. In discussing real materials, it is important to realize that many cases occur in which the type of bonding falls into two or more of the above ideal classes. Thus, the compounds gallium arsenide, indium antimonide, and aluminum phosphide are examples of materials that have both covalent and ionic contributions to the atomic bonding, and the elements bismuth and antimony are examples of materials that have a partially metallic and partially covalent type of atomic bonding.

Ionic solids. Ionic bonding occurs between strongly electropositive (with weak affinity for electrons) elements, such as the alkali metals (lithium, sodium, etc., which form positive ions, and strongly electronegative (with strong affinity for electrons) elements, such as the halogens (fluorine, chlorine, etc.), which form negative ions. In this type of solid, electron transfer occurs between atoms, from electropositive to electronegative, and the solid consists of the resulting positively and negatively charged atoms or ions. For simple ionic solids such as sodium chloride, lithium fluoride, and cesium chloride, this electron transfer results in each ion having a closed electronic shell. The charged ions arrange themselves in space so that the coulombic attraction (a term for the attraction or repulsion between charges) between unlike charged ions equals the coulombic repulsion of similarly charged ions. While many different crystal structures can result from ionic bonding, three commonly occurring crystal structures are cesium chloride, sodium chloride, diamond, and zinc sulfide, types shown in Figure 11. It should be noted that in these crystal structures each ion has oppositely charged ions as its nearest neighbours, as a result of the coulombic forces acting between the ions.

From C.A. Wert and R. Thomson, *Physics of Solids*, 2nd ed. 1970 (copyright 1964); used with permission of McGraw-Hill Book Co.

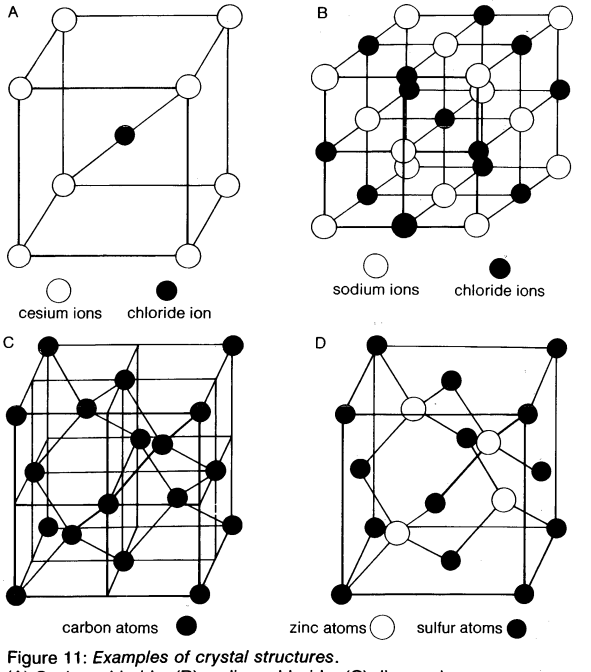


Figure 11: *Examples of crystal structures.*
(A) Cesium chloride; (B) sodium chloride; (C) diamond; (D) zincblende.

Calculating electrostatic interaction

The major contributions to the cohesive energy are electrostatic in origin and are called the Madelung energy. In calculating this energy, coulombic interactions between all the charged ions are summed. For a stable atomic arrangement, this energy must be attractive. The electrostatic, or coulomb-interaction, energy is calculated as the product of the charges on each ion, $\pm q$, divided by r, the distance between the ions, this quantity having the form q^2/r. Thus, as the separation becomes less, the interaction increases, and the closely packed crystal structures in which each ion is surrounded by nearest neighbours having opposite

charges lead to a total electrostatic energy that is negative; *i.e.,* energy decreases on forming the crystal from dissociated ions. As the spacing between ions decreases, the crystal energy becomes increasingly negative, due to the $1/r$ dependence of the coulomb interaction. At small ionic separations, however, the closed electron shells of the adjacent ions begin to overlap in space, resulting in a rapid increase in energy as the ionic separation decreases. (The basic cause of this repulsion is accounted for by the Pauli exclusion principle, which states that no two electrons can have the same quantum numbers.) Equilibrium separation of ions of the crystal corresponds to the minimum energy position and is equal to about three angstroms (3×10^{-8} centimetre). The properties of ionic crystals are largely determined by the charge transfer that takes place. For example, when an electric field is applied, the current is carried by the motion of the charged ions and therefore depends on the diffusion of negatively charged ions toward the electrode carrying a positive charge or the positive ions toward a negative electrode or both. The electrical conductivity of a crystal therefore increases exponentially with temperature (since the increased thermal energy causes an increased rate of atomic jumping) at a rate that reflects the rate of diffusion of the ions.

Covalent solids. Covalent bonding (also known as homopolar bonding in organic chemistry) occurs when the valence (or outer) electrons of neutral atoms are shared between nearest neighbours in the crystal. This atomic arrangement allows each of the atoms to achieve a closed outer electron shell by sharing electrons with its neighbours. Covalent bonds generally contain two electrons, one from each neighbouring atom, and, since eight electrons are required for a closed electron shell, the arrangement of the atoms is such that each has four nearest neighbours. This results in a tetrahedral (having four faces) arrangement of atoms in space, and covalently bonded solids often have the diamond structure shown in Figure 11. The covalent bonds are highly directional, with the bonding electrons being predominantly located along the directions between nearest neighbour atoms; *i.e.,* the atoms that provide the two electrons forming the bond. One of the parameters describing the energy state of the electron is the direction of its own spin, and, when the electrons of the bonding pair have opposite spin, the interaction between them is attractive. The energy of the associated atoms of the solid system is decreased from the energy of the atoms when they are free by the sharing of electrons between nearest neighbours (to allow a completed electron shell for each atom), and the decrease provides the cohesive energy for the system.

The role of electron spin

This ideal picture of the covalent solid is applicable to the bonding of the Group IV elements in the periodic table, such as carbon (when it occurs in the form of diamond), silicon, and germanium. Each of these has four valence electrons, which are shared in four tetrahedral covalent bonds. Other elements, however, also can exhibit covalent bonding. The two atoms of molecular hydrogen (H_2) are covalently bonded. Compounds of Group III and Group V elements such as gallium arsenide (GaAs) or indium antimonide (InSb) have at least partial covalent bonding, with the Group III element, gallium or indium (Ga, In), contributing three and the Group V element, arsenic or antimony (As, Sb), contributing five valence electrons to the covalent bonds. In these compounds the atomic structure is termed the zincblende structure, which differs from the diamond structure only in that each atom (*e.g.,* gallium) is surrounded by nearest neighbours of unlike atoms (*e.g.,* arsenic). Some of the elements of Groups V, VI, VII of the periodic table (such as arsenic, antimony, tellurium, iodine, and bromine) form complex crystal structures in which the bonding is at least partially covalent. In these less simple solids, covalent bonding occurs between atoms with fewer than four nearest neighbours arranged in sheets, chains, or molecules. The bonding between these sheets, etc., is the result of the induced displacement of the electrons and nuclei of neighbouring atoms and is called the van der Waals force; it is much weaker than covalent bonding.

The properties of crystals in which the bonding is predominantly covalent are determined largely by the localized nature of the electrons forming the bonds. These materials include electrical insulators such as diamond (room-temperature conductivity of about 10^{-14} [ohm-centimetre]$^{-1}$) and semiconductors such as germanium (room-temperature conductivity of about 40 [ohm-centimetre]$^{-1}$). In both cases, the electric current is carried by electrons that have been thermally excited (*i.e.,* their energies have been raised by heat) from their localized configuration (*i.e.,* basic position in the normal atom) to a more energetic state, or state of higher energy, in which they are nonlocalized and can move in response to an outside electric field. This excitation of electrons from the valence band (*i.e.,* electrons involved with bonding forces) to the conduction band (*i.e.,* electrons with enough energy to overcome local bonds and move through the spaces among the atoms if an electrical force is applied) leads to an electrical conductivity that increases exponentially with temperature. The various covalently bonded crystals differ in the amount of energy required to excite an electron to the conduction band, the value being about 160,000 calories per mole (*i.e.,* per 6.02×10^{23} electrons) in diamond and 28,000 calories per mole in silicon.

Valence and conduction bonds

Metallic solids. Bonding of solids formed from elements such as copper, aluminum, beryllium, sodium, and the many others classified as metals arises from the nonlocalized nature of the valence electrons in these solids. When an atom is brought into a metallic crystal, those properties of the valence electrons that are called its wave functions are altered because of the proximity of the valence electrons and the positively charged ionic cores (*i.e.,* the electrons that do not function in the chemical bond because they are held so firmly by the nucleus) of the other atoms in the structure. The resulting valence-electron wave functions are associated with the entire volume of the metal, and therefore each valence electron can no longer be associated with a particular ion core. The average energy of these nonlocalized valence electrons, which is partly kinetic energy of motion through the solid and partially potential energy due to interactions with the ionic cores and the other valence electrons, is reduced on forming the solid, thus providing the cohesive energy. It is the nonlocalized or free-electron nature of the valence electrons in metals that results in their characteristic properties. In some cases, however, the electrons can still retain much of their localized nature. Thus, many of the properties of the transition elements in the periodic table (*i.e.,* those with unfilled *d* shells in the structure of orbitals arranged according to energy distribution among the electrons; *e.g.,* iron, vanadium, chromium, zirconium, molybdenum) are determined by the behaviour of these *d* electrons in the solid. The *d* electrons remain largely localized around the ionic core; *i.e.,* they are not free to move through the solid but still contribute to properties such as the cohesive energy and magnetic properties.

Metallic crystals have characteristically high electrical conductivity (about 10^6 [ohm-centimetre]$^{-1}$ for a metal such as copper) due to the free-electron character of the valence electrons. They can move away when an electrical field is applied without requiring any thermal activation; *i.e.,* without having to be thermally excited (heated) into a conduction band, as in covalent crystals. In these cases, however, the motion of the electrons is limited; they are scattered when they meet imperfections in the crystal and when they approach atoms that are vibrating due to heat energy. The electrical conductivity, therefore, decreases as the temperature is increased, because an increase in the magnitude of thermal vibrations of the atoms obstructs the movement of free electrons (see below *Conductors*).

Molecular solids. Rare-gas solids (such as argon, krypton, and xenon) and some molecular solids (*e.g.,* methane, CH_4; hydrogen, H_2; chlorine, Cl_2) exhibit weak van der Waals or molecular bonding in the solid state. Since the bonding occurs between atoms that have completely filled outer electron shells or non-ionized molecules, the crystals are only weakly bonded and are stable only at low temperatures. The van der Waals bonding results from the displacement of the atomic nuclei from the centres of the electron charge distribution in each atom, the

displacement resulting when neutral atoms or molecules are brought close to each other. This induced charge displacement creates an electric dipole (*i.e.*, a body with two poles oppositely charged) that interacts with the adjacent induced dipoles to reduce the energy of the crystal and provide a small cohesive energy. The attractive interaction energy between the induced electric dipoles is of very short range and varies as the distance between atoms to the −6 power (or 1/distance^6). It is balanced by the repulsive interaction that occurs as the filled electron shells of the atoms begin to overlap on bringing the atoms closer together, a consequence of the increase in energy due to electrons having the same energy; *i.e.*, the same four quantum numbers, which account for all the energy of an electron relevant to that electronic structure, occupying the same volume. (According to the Pauli exclusion principle, no two electrons of an atom may have the same four quantum numbers.) The nature of this type of molecular bonding leads to electrons that are localized at each atom and therefore to molecular solids that do not conduct electricity. They are insulators. Very large amounts of energy are required to excite an electron in such a solid from a localized state to a nonlocalized one. Due to the low cohesive energy of these crystals, significant electron excitation does not occur in the temperature range in which they are solid.

Hydrogen-bonded solids. Hydrogen bonding is of great significance in organic solids and in certain molecular solids such as ice, hydrogen fluoride, and potassium hydrogen phosphate. The single valence electron of hydrogen allows it to form a covalent bond with only one other atom, as with another hydrogen atom in the hydrogen molecule, H_2. In some solids containing strongly electronegative atoms, such as nitrogen, oxygen, or fluorine, hydrogen can form a bond between two of the electronegative atoms. In these cases, the hydrogen is at least partially positively ionized (*i.e.*, its electron has been partially removed) and serves to bond two negatively charged ions. Such bonding is of significance in many polymeric materials and in organic solids such as DNA (deoxyribonucleic acid), where it acts to bond the strands of the DNA molecule.

Comparison of bond strengths

As was previously stated, solids rarely exhibit a single idealized type of atomic bonding. In general, ionic and covalent bonding is very strong, leading to high heats of vaporization and melting points; molecular bonding is very weak; and metallic bonding occupies an intermediate position. The type of bonding also places some constraints on the structures in which the solids crystallize; however, the variety of crystal structures observed is much greater than might be suspected from the few simple types of atomic bonding described above. (C.A.W./H.K.B.)

STRUCTURE OF CRYSTALLINE SOLIDS

Crystalline solids reflect a highly symmetric and periodic internal arrangement of atoms. This is characteristic of all such solids—metals, alloys, minerals, ceramics, and some organic materials. Many rows and planes of atoms occur within a crystal and some rows (and planes) exhibit equal spacing of atoms in different directions, whereas others do not. Because the atomic arrangement is generally different in different directions, properties such as elasticity or light transmission commonly depend on direction.

Interfacial angles

Patterns of atoms in crystals. One of the first activities undertaken by those interested in natural crystals was an attempt to classify their shapes. This was not easy because even crystals of the same compound may fail to resemble one another; the difference in shape results from different degrees of development of the several faces. In 1671 the Danish physician Nicolaus Steno took the first grand step in this field. He cut sections from a quartz crystal and discovered that, even when the shapes were vastly different, there was a constancy of interfacial angles (the angles formed by normals or perpendiculars from crystal faces). For this reason it is common practice to use "normals" drawn perpendicular to crystal faces to represent the relationship between faces. This can be accomplished by stereographic projection as shown in Figure 12. The intersection of normals with an imaginary sphere surrounding the crystal (Figure 12A) is represented on the two-dimensional map (Figure 12B) of the shaded plane. The high degree of symmetry present is evident.

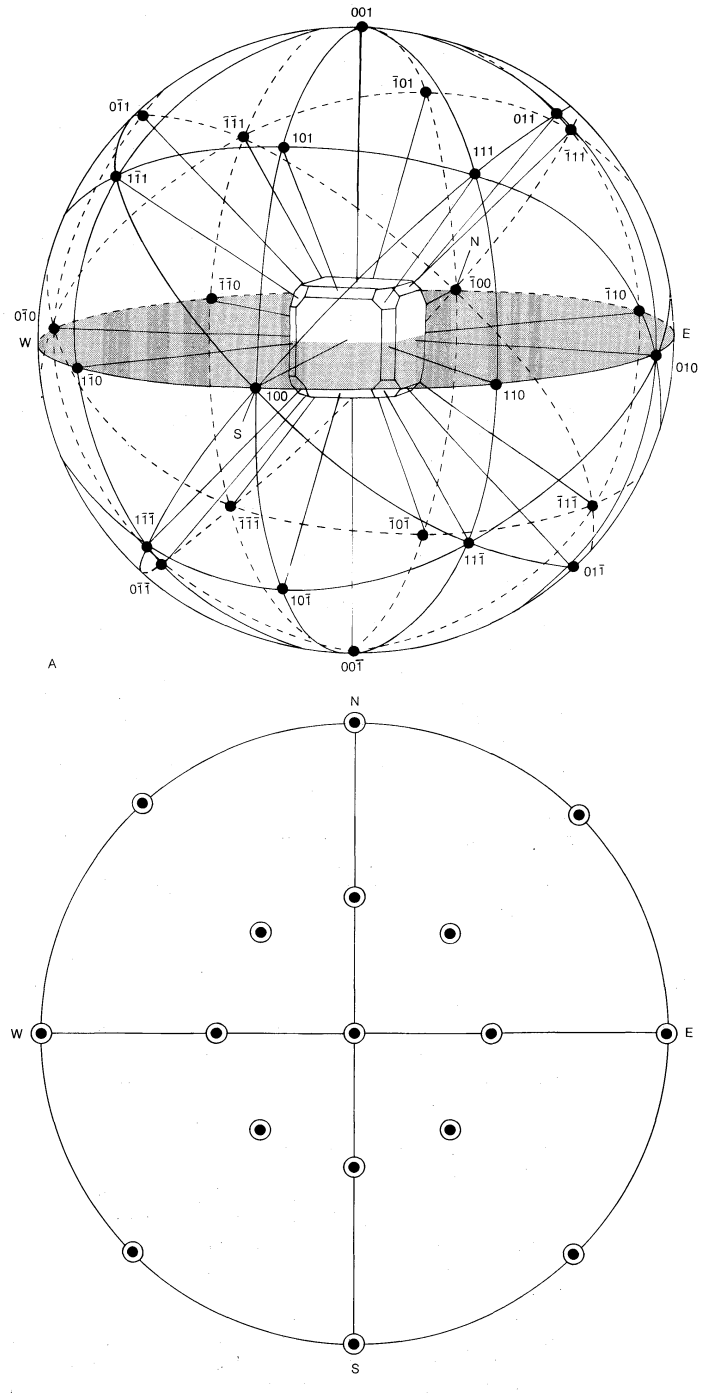


Figure 12: A two-dimensional representation (stereographic projection) of a crystal and its symmetry that can be obtained by surrounding the crystal with a sphere, erecting perpendiculars to the planes until they intersect the sphere as in A, and then projecting these intersections to 001 or $00\bar{1}$ positions through the shaded circle of the sphere. The appearance is then as in B. This process eliminates the shapes of the planes and their relative sizes, which depend on growth conditions and can confuse the similarity of crystals.
From E.E. Wahlstrom, *Optical Crystallography* (1951); John Wiley and Sons, Inc.

Crystal planes and their notation. In the late 18th century the idea that a large crystal could be built of a basic unit or building block took hold. This block is now known to be the unit cell—the smallest group of atoms within a crystal that will yield the crystal structure and properties by simple repetition of pattern. This arrangement is indicated in two dimensions in Figure 13, in which two planes of atoms perpendicular to the page are shown by

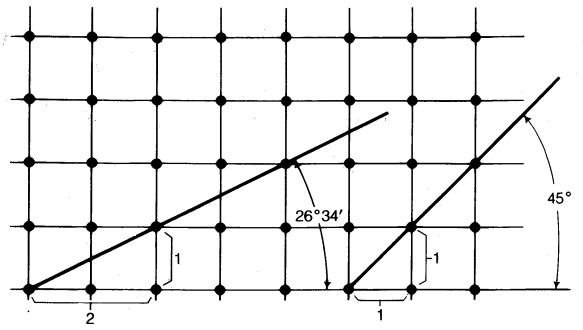


Figure 13: Two planes of atoms in a crystal represented by their edges (dark lines). Note that these make intercepts that are whole numbers along the edges of the cells or building blocks.

From F.C. Phillips, *An Introduction to Crystallography*, 2nd ed. (1956); Longman Group Ltd.

the dark lines. Such planes will appear perfectly flat to the eye (rather than composed of a series of steps) because the building blocks are very small, about 10^{-9} to 10^{-10} metre on an edge. The angle formed by the plane on the right is 45° with the horizontal plane. Note that it is not possible to have just any angle; a plane must contain atoms in the structure, and the intercepts on the edges of the crystal are ratios of whole numbers of blocks. The intercepts of the plane on the left and the horizontal plane have a ratio 1:2 and the angle involved is 26°34′.

A simple notation for naming planes has developed based on the fact that the intercepts of planes on the edges of the crystal's unit blocks are ratios of whole numbers. A basic unit cell or block is assumed, and three noncoplanar edges (a_1, a_2, a_3) are defined as unit lengths in Figure 14.

Reprinted with permission of the Macmillan Company from J.B. Cohen, *Diffraction Methods in Materials Science*; Copyright © by Macmillan Company (1966)

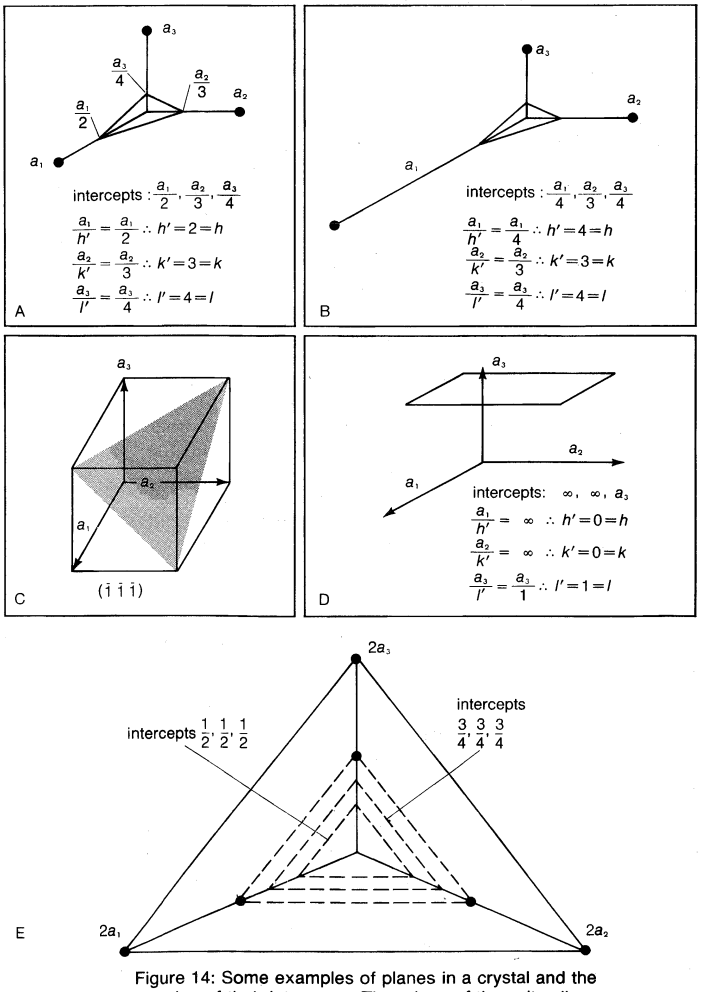


Figure 14: Some examples of planes in a crystal and the notation of their intercepts. The edges of the unit cell are defined by a_1, a_2, and a_3 (see text).

These need not be equal in all directions. The intercepts of planes on these edges are written as

$$\frac{a_1}{h}, \frac{a_2}{k}, \frac{a_3}{l}.$$

Thus, if a plane intercepts the axes at their tips,

$$\frac{a_1}{h} = \frac{a_1}{1}, \qquad \frac{a_2}{k} = \frac{a_2}{1}, \qquad \frac{a_3}{l} = \frac{a_3}{1},$$

and (111) is the designation. If the intercepts are ½, ½, ½, then *h,k,l* are 2,2,2. Common factors are cleared and the designation is still (111)—because all such planes are parallel and have identical atomic arrangements in the structure (Figure 14E). If a plane intercepts one axis at unit distance and is parallel to the other axes, as in Figure 14D, then two integers will be zero—for example, (001) in this case.

Symmetry considerations. Rotation of 90° around the centre of the map of normals to planes in Figure 12B leads to the same figure as before the rotation, as does rotation of 180°, 270°, or 360°. This rotation is generally described by the number *n*, by which 360° (a complete turn) is divided. In this case $360°/n = 90°$, so that *n* is 4, and then this apparent rotational symmetry axis is thus called "4-fold." To return to the starting point after *n* rotations, it is then necessary that *n* be an integer. In examining many natural crystals, *n* has been found to have values of 1,2,3,4, or 6; sometimes several of these values are found to be present in one crystal. But the values are never 5 or 7, although certain flowers and starfish are excellent examples of these symmetries. In order to understand the reason such values are excluded in crystals, it is necessary to ask more precisely what results from periodic arrangements of atoms. One-dimensional arrays are easiest to consider, and the different ways a one-dimensional molecule can be periodically arranged in one dimension provides a suitable example.

Rotational axes of symmetry

A periodic set of crosses, as shown below, is first set down as a frame of reference; this imaginary collection is called a lattice.

\+ \+ \+ \+ \+

The actual spacing of the crosses is not important because it is the kinds of arrangements that need to be considered, not their sizes. The entire array can be constructed after one point and the spacing between the points are designated. This basic framework—a point and the spacing to the next point—is called a unit cell, and in one dimension there is obviously just one such cell. If a molecule (represented by a dot–dash) is placed within this framework, two arrangements may be considered, as follows:

(1) +·— +·— +·— +·—

(2) +·— —·+·— —·+·— —·+·—

In the second lattice the "molecules" are arranged at each lattice point, so that one molecule leads to another after rotation of 180° around an axis normal to the paper (at the crosses or halfway between them). This is called 2-fold rotation because repetition occurs twice in one complete (360°) rotation. There are two molecules per lattice point. A mirror can also be imagined at these locations, with the molecules related by reflection through the mirror. Two reflections yield the original molecule. This indicates that the interaction of symmetry leads to some redundancy—one symmetry element produces another; rotational symmetry at lattice points produces rotational symmetry between lattice points, as well as the mirrors, but only one symmetry element or type is required to describe the figure. Only the unit cell and its contents are needed to produce the entire structure by translation, although the unit cell in lattice (2) contains two molecules.

Another possible pattern is the following:

(3) +·— —··—+·— —··—+·— —··— +
c a b d

But this is really not different from pattern (1); the unit in each cell is just longer—three molecules long. It might

appear that there is rotational symmetry about an axis between a and b, but c would not be carried into d by the rotation. This is not a symmetry element of the whole structure. Hence the requirement of periodicity leads to only two patterns in one dimension. In these periodic structures, examination of some point within the structure, a lattice point, or any point on a molecule, reveals the same surroundings at every such point in every unit cell.

Symmetry in two and three dimensions

The entire array is called a space group, and the symmetry about a point is called a point group. There are one unit cell, two point groups, and two space groups in one dimension. Compact notations have been developed for these in one, two, or three dimensions and will be considered shortly. Symmetry in two and three dimensions will be treated first, however. Symmetry at a point in two dimensions involves the question: How many ways can a molecule be arranged periodically about a lattice point? Ways without any symmetry are also included because a 1-fold rotation axis produces this effect—*i.e.*, the entire crystal will exhibit no symmetry at all. Arrangements that do include some symmetry are shown in Figure 15, in which a stick hand has been chosen as a general figure for two dimensions because it has no symmetry of its own. Hands A and C are related by 180°, or 2-fold rotation, and A and B are related by reflection through the imaginary mirror plane m_1 perpendicular to the page. But A and C could also be related by reflections through m_1 and then through m_2. A rotation of 180° and reflection still retain some equivalence, as in one dimension, but there is no way to get B from A except by reflection.

Another symmetry element has arisen at the intersection of the two mirror planes; figures A and C are related by inversion through this point of intersection. That is, C can be obtained from A by drawing lines from every point on A, through the intersection, over an equal distance on the other side of the intersection. In two dimensions then, C can be produced by starting with A and rotating 180°, or by inversion. The presence of one of these symmetry elements leads to the other. Inserting m_1 and 180° rotation leads to all four hands and, in turn, to two other symmetry elements (m_2 and an inversion centre) that were not needed in the construction. The mirror plane m_2 leads to nothing new in terms of the pattern produced by m_1 and 180° rotational symmetry, for example.

A periodic array, or space group, is shown in two dimensions in Figure 16A. A flag is employed to represent a two-dimensional molecule; one unit cell is shaded and contains four such molecules. To create the entire periodic array, the lengths of the two sides of the unit cell, a_1 and a_2, and the angle (a) between them are required. One cell can then be constructed and translated in the directions of a_1 and a_2 to fill all space. If one molecule is placed near each lattice point and if each lattice point has two perpendicular mirrors, then the four molecules at each lattice point result,

From J. Cohen, *Diffraction Methods in Materials Science*

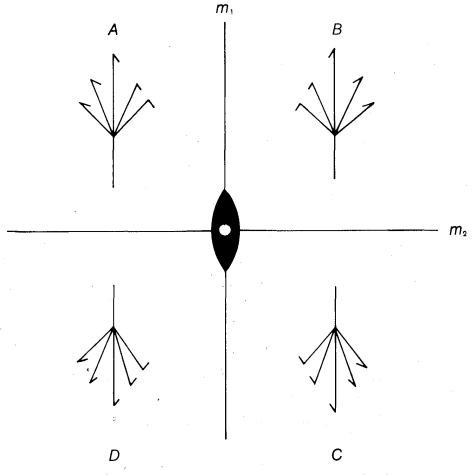


Figure 15: A set of stick hands arranged symmetrically around a point. Both mirror and rotational symmetries are present. The first relates A and B through m_1. The second relates A and C; the axis is normal to the page and is represented by the oval (see text).

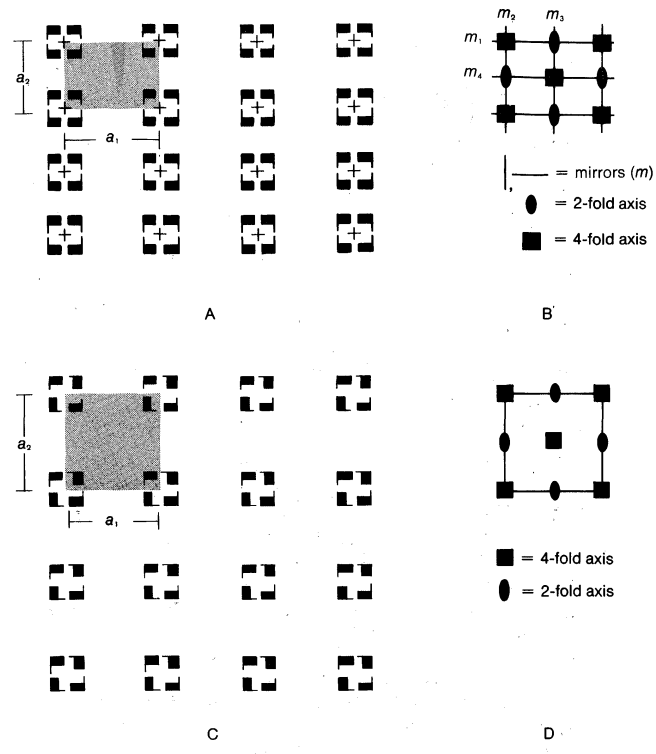


Figure 16: Two-dimensional space groups in which a unit cell is shaded (A and C). At each lattice point there are four molecules (flags). These are related by two perpendicular mirror planes, m_1, m_2 in A but additional symmetry arises, such as m_3, m_4 and 2-fold and 4-fold rotational axes normal to the paper (B). Similarly D shows the symmetry in C.

together with the additional symmetry elements shown in Figure 16B. Another arrangement, involving a 90° rotational axis without mirrors, is shown in Figure 16C.

Impossibility of 5-fold or 7-fold symmetry

It is clear that any axis of rotational symmetry in a crystal must be $360°/n$, in which n is an integer, because to have periodicity at a point there must be a return to the original figure after n rotations. The reason that n is restricted to 1, 2, 3, 4, 6 in a crystal, excluding 5 or 7, becomes clear with the aid of Figure 17. Assume that there is a rotational axis, with rotation $a°$ at each lattice point. Not only must all molecules or atoms around any lattice point be related by rotation of $a°$ around this axis, but also this must be true for all points in the entire structure; otherwise, the symmetry would relate only to molecules at one point and not to the entire crystal. The lattice point itself must be carried by rotation to another lattice point; this preserves the internal periodicity and the external symmetry.

From Figure 17, if m is an integer, then the relation between the distances a_1, b, and the angle of rotation, a, can be expressed by the equation $b = ma_1 = a_1 + 2a_1 \cos a$, or

$$\cos a = \frac{m-1}{2} = \frac{M}{2}, \qquad (1)$$

in which M must be an integer if m is an integer. The cosine of an angle cannot exceed plus or minus unity; and with M an integer the possible values for the cosine are 0, plus or minus $\frac{1}{2}$, plus or minus 1. For cos a to be zero, a must be 90° ($n = 4$). Furthermore, for this 4-fold rotation axis, any side a_2 must equal a_1, because after a 90° rotation a_2 must coincide with a_1. The cell must be square—it cannot be a rectangle. The other allowable values for the cosine correspond to 1-fold, 2-fold, 3-fold, and 6-fold rotations and to two other shapes, or unit cells, a rhombus and a parallelogram. It is possible to have a molecule with 5-fold symmetry—like a 5-pointed star—at each lattice point, but the symmetry of the entire structure will not be 5-fold; similarly, 7-fold symmetry is prohibited.

Combined symmetry operations

Combined operations—other than just rotation, reflection, inversion, and translation—are also periodic in two dimensions. A glide, for example, is a combined operation

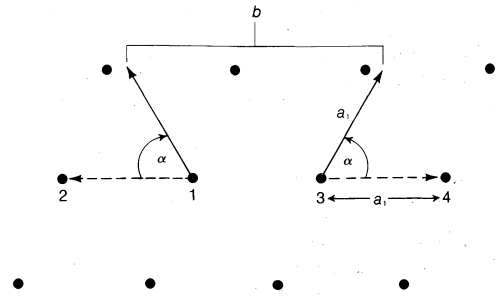


Figure 17: Rotational symmetry in crystals: rotation of α around lattice point 1 must bring lattice point 2 into coincidence with some other point; the same is true for 4, around 3. If not, as shown here, the periodicity, and hence the external symmetry, is destroyed (see text).

in which the molecule is related to another by translation followed by reflection; the translation is half the spacing between lattice points, and thus two glides produce the same appearance at each lattice point. Proceeding from one lattice point to another, a molecule is seen in left-handed orientation, one halfway between lattice points is in a right-handed orientation, and at the next lattice point the original left-handed orientation appears. Other possible combined operations in two dimensions, such as rotation followed by reflection or rotation followed by translation, give rise to figures which are equivalent to rotation alone or to combinations of rotation plus reflection at each lattice point. It is possible to show that there are only 17 two-dimensional patterns or space groups and only 10 point groups that are consistent with periodicity.

In three dimensions, two new kinds of arrangements of symmetry become possible. It is possible to have three rotation axes intersecting at a point in certain combinations, and axes can be combinations of rotation and translation called screws, as shown in Figure 18. As with glide, this operation in a crystal must bring a molecule to a definite location and orientation near a lattice point in every cell. This serves to limit the possible types or pitches of the screws.

Displacements of planes of atoms or molecules that result from screw axes in natural crystals are approximately equal to the magnitude of atomic spacing—*i.e.*, 10^{-10} metre or one angstrom (Å). The effect, therefore, is indistinguishable from pure rotation to the naked eye, as shown in Figure 19.

The seven crystal systems

Seven crystal systems and 14 unique unit cells arise, as shown in Figure 20. There are 32 point groups and 230 possible structures based on atomic position. Other arrangements can arise if such properties as the magnetic moment are considered, but the number of unique periodic arrangements is still finite.

Optical determinations. Optical techniques can be quite useful in the study of transparent crystals, because only in cubic crystals does light travel at the same velocity in all directions. In a tetragonal structure, for example, light is broken into two rays for any direction of propagation other than that of the 4-fold axis. These rays travel at different velocities and vibrate in different planes. If light is transmitted through a sheet of Polaroid, which allows passage of light vibrating in only one plane (direction), then rotation of the crystal will lead to elimination of one ray when its plane of vibration is perpendicular to the vibrations passing through the Polaroid sheet. If another piece of Polaroid is placed over the crystal and turned 90°, the light passing through the crystal cannot pass this second sheet. Every turn of the crystal by 90° produces this extinction, but between these positions the crystal is visible. If a thin slice of the crystal is examined between crossed Polaroids along the 4-fold axis, this splitting does not occur, and the crystal is always dark. This procedure can help distinguish the presence of a unique axis of symmetry. If a screw axis is present the position of the two Polaroids for extinction varies with the thickness of the crystal. Glide cannot be detected in this way, however. (J.B.Co.)

Atomic arrangements in metals and alloys. In metals, atomic arrangements are characterized by dense packing, similar to packing billiard balls in a box. Two of the most common metallic crystalline structures can in fact be understood in just this way. The first layer of atoms is visualized as on a flat surface as shown in Figure 21, each atom touching its six nearest neighbours. The point at the centre of each atom is labelled *A*. The next layer of atoms is placed with each atom nestled into position *B* (position *C* could also be selected) between three atoms of the first layer. Again, for the third layer there are two available positions: *A*, directly over the first layer, or *C*. If *A* is used, and the pattern repeated, *ABABAB*, the hexagonal close-packed structure characteristic of magnesium results. In the figure, the position *C* is used for the third layer. Repeating this pattern, *ABCABCABC*, produces the face-centred cubic structure characteristic of aluminum. The term face-centred derives from the fact that this arrangement, viewed from a different angle, can be seen to be made up of cubes with one atom at each corner and one at the centre of each cube face.

From J. Cohen, *Diffraction Methods in Materials Science*

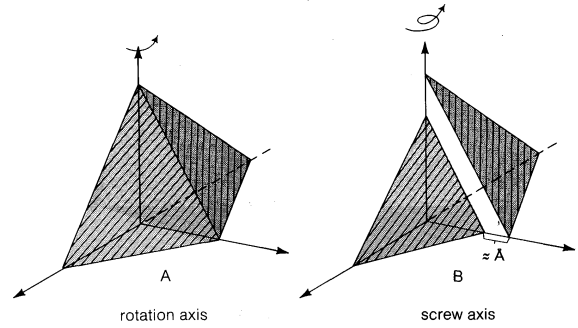


Figure 19: *Distinction between rotational and screw axes.* (A) Rotation axis relates the shaded faces. (B) A screw axis relates the shaded faces, but there is also a translation. This is so small (about 10^{-10} metre, or one angstrom), however, that the differences in these two cases cannot be seen on a real crystal (see text).

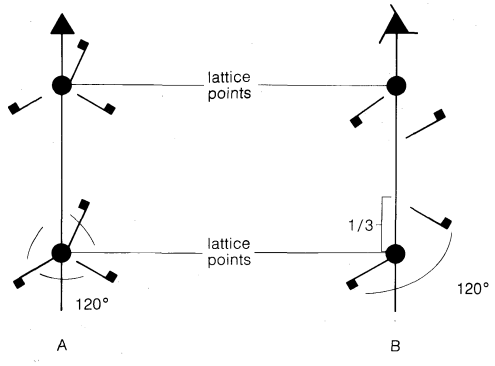
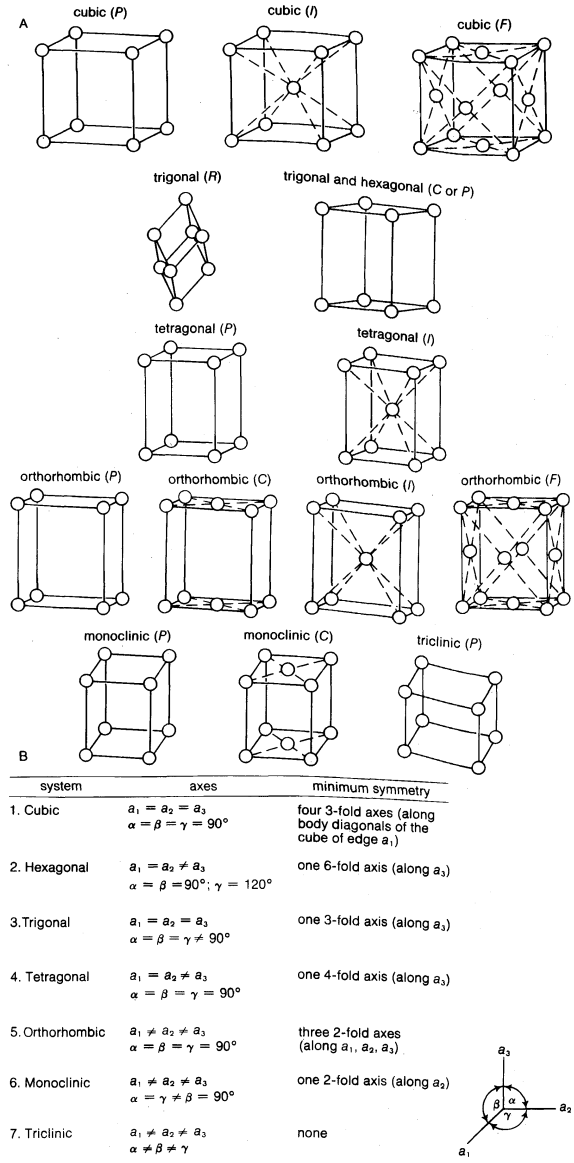


Figure 18: *Symmetry associated with screw axes.* (A) Two lattice points in a structure with three molecules rotated by 120° around each; the whole group at the bottom is translated to every other lattice point (like the one at the top) by the periodicity of the crystal. (B) Periodicity can also be obtained with screwlike motions (rotation plus translation along the rotation axis) where there is only one molecule at each lattice point; each molecule is related to the other molecules by a 120° rotation and a translation of ⅓ the distance between lattice points. Triangular symbols at the top represent rotational and screw axes in A and B, respectively.

The hexagonal close-packed and face-centred cubic structures correspond to the densest possible packing of spheres. The other common structure is obtained by beginning on the bottom layer with a square arrangement of atoms and then nestling a similar square array on top. This arrangement is the body-centred cubic structure and is only slightly less densely packed than the face-centred structure.

Polycrystalline form

Ordinarily, metals occur in polycrystalline form; that is, consisting of a large number of small crystals, or grains, separated by thin grain boundaries (see below *Crystallization and crystal growth*). In some cases, such as the zinc coating on a galvanized pail, the grains are large enough



system	axes	minimum symmetry
1. Cubic	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma = 90°$	four 3-fold axes (along body diagonals of the cube of edge a_1)
2. Hexagonal	$a_1 = a_2 \neq a_3$ $\alpha = \beta = 90°; \gamma = 120°$	one 6-fold axis (along a_3)
3. Trigonal	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma \neq 90°$	one 3-fold axis (along a_3)
4. Tetragonal	$a_1 = a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$	one 4-fold axis (along a_3)
5. Orthorhombic	$a_1 \neq a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$	three 2-fold axes (along a_1, a_2, a_3)
6. Monoclinic	$a_1 \neq a_2 \neq a_3$ $\alpha = \gamma \neq \beta = 90°$	one 2-fold axis (along a_2)
7. Triclinic	$a_1 \neq a_2 \neq a_3$ $\alpha \neq \beta \neq \gamma$	none

Figure 20: *Classification of crystal systems.*
(A) The 14 unique three-dimensional unit cells. Balls represent lattice points around which, or on which, atoms or molecules can be arranged. (B) Symmetry characteristics of each crystal system.
From J. Cohen, *Diffraction Methods in Materials Science*

to be easily discerned with the naked eye. In ordinary iron or steel, the grains are so small that they can be seen only under magnification. Large, single crystals of metal can be grown from a melt or from a metallic vapour under special controlled conditions.

When a metal is melted, the regularity of the crystalline lattice is destroyed, just as if the box of closely packed balls described above was shaken, and the density is correspondingly reduced (*i.e.,* the sample expands). Semiconductors, such as silicon and germanium, which are not classed as true metals because of their electrical properties, become metallic when melted; but, in contrast to metals, their density increases with melting. In liquids, the atoms are constantly rearranging themselves. Solids that do not have crystalline structure, such as glass, are called amorphous (see below *Amorphous solids*). It is possible to obtain such noncrystalline, amorphous structures by depositing metals upon very cold surfaces, but this amorphous form of metal is unstable and will crystallize when warmed. Grain boundaries in polycrystalline metals may be considered amorphous.

Alloys

Metallic alloys are obtained by first dissolving one metallic element in another under heat and then solidifying the melt, the solid alloy assuming a crystalline form. Most alloys are substitutional; when the atoms are of similar size, the atoms of one element replace those of another in the lattice. A binary alloy consists of two metallic elements. In some cases, as when carbon or hydrogen are added to iron, the atoms of the added element are small enough to fit between the larger atoms of the host crystal to form what are called interstitial alloys. In both the substitutional and interstitial alloys, there is some misfit of the constituent atoms and a distortion of the crystalline lattice from that which occurs in the pure metal.

Many alloys become ordered at low temperatures; that is, the different kinds of atoms are arranged in a regular sequence throughout the crystal volume. An example is beta (β-) brass, which consists of equal numbers of copper and zinc atoms. This alloy occurs in a body-centred cubic structure, atoms of copper and zinc alternating in the ordered manner shown in Figure 22. The copper atoms form a cubic superlattice as do the zinc atoms. (In some alloys, this tendency to order is so strong that the properties are much the same as those of a compound; in these instances an intermetallic compound is said to have been formed.)

If the temperature of this brass is raised, it becomes disordered at 470° C (880° F). This phase transition is comparable to that of the disappearance of crystalline order of a pure metal at the melting point. In both cases, the transition leads to higher entropy (condition of unavailable energy), corresponding to greater disorder in the high temperature phase.

Many pure metals and alloys undergo phase transitions to different crystalline structures as the temperature is changed. Sodium makes such a transition from body-centred cubic to hexagonal close-packed with decrease in temperature. Again, the entropy of the higher temperature phase is larger; in this case, the additional disorder arises because the body-centred cubic phase is softer, and therefore the thermal vibrations of the lattice, another kind of disorder, are larger.

In all such cases, the disordering of the material provides an additional mechanism for the storage of heat energy and therefore an increase in the heat content. The disorder also disrupts the flow of electrons so that ordinarily, though not always, the higher temperature phase will have higher electrical resistance. (W.A.H./Ed.)

Scattering of radiation from crystals. In 1912 Max von Laue became intrigued with the question of what would happen when radiation with a wavelength equal to the spacing of the atoms in a solid was scattered from the atoms. Light scattering could not reveal atomic structure because atomic spacings are 10^{-10} metre, whereas light has a wavelength 10^{-7} metre. Laue thought, however, that scattering of radiation from a row of atoms in a crystal (if the wavelength of the radiation was the same size as the atomic spacing) would produce interference patterns. In a three-dimensional crystal, there are rows in three directions and, hence, such patterns in three directions. Only at their points of intersection will all the waves be in phase, and a set of dark spots should occur on a film placed in front of a small crystal when an X-ray beam impinges on it. It was not known whether X-rays had the proper wavelength, but Laue persisted in his experiment, and he succeeded on his second try. Laue and his associates demonstrated the existence of a new tool for examining the internal arrangements of atoms in crystals and proved that X-rays behave as electromagnetic waves (like light) but with a shorter wavelength.

Bragg's law and crystal structure. Laue's group had difficulties in understanding which planes of atoms and wavelengths produced the spots on the film. In England, Sir Lawrence Bragg was in his mid-twenties and finishing his studies in physics. His father was involved in research on the ionization of gases by X-rays, and it was quite natural for them to discuss Laue's exciting new result. The young Sir Lawrence was able to formulate a simpler and more physical understanding of Laue's discovery. The scattered waves from successive planes of atoms reinforce one another at certain angles if the path difference between rays from the two planes is a multiple (n) of the wavelength. As shown in Figure 23, this construction leads to a simple equation for diffractions or constructive interference. For

Relation of wavelength and angle of incidence

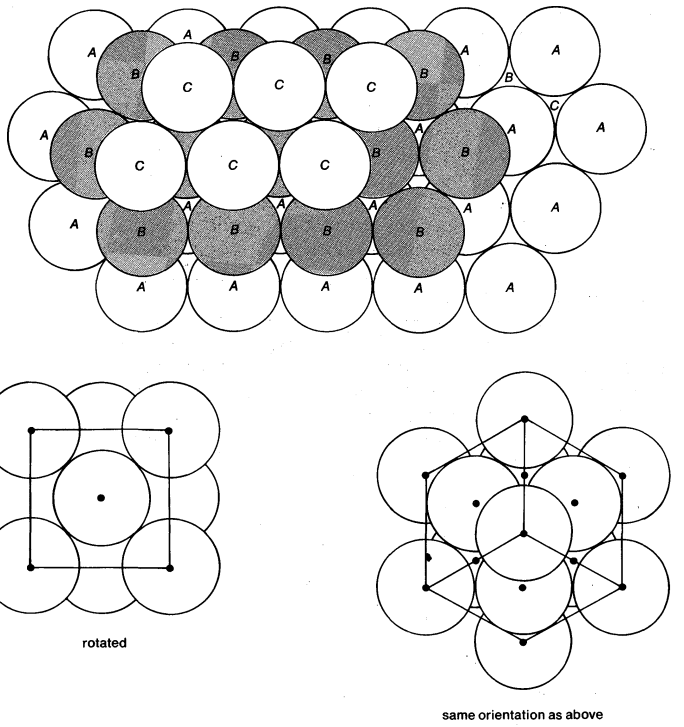


Figure 21: Derivation of face-centred cubic structure from planes of hexagonally close-packed atoms (see text).

diffraction the path difference, AB + BC, is equal to an integer times the wavelength, or $n\lambda$. But by simple trigonometry, AB = BC = $d \sin \theta$, and thus Bragg's law is derived: $2d \sin \theta = n\lambda$. As n increases from 1 to 2 to 3, etc., so does the scattering angle θ. Each plane, for each n, selects in effect the appropriate wavelength from the X-ray beam (which must consist of many wavelengths). This selection depends on the angle of the plane of atoms to the X-ray beam. With this formulation, Bragg was able to explain successfully all the spots from Laue's patterns from different planes of atoms. He then tried bouncing an X-ray beam from a crystal in reflection, as from a mirror, catching the scattering on a film. This was more like his formulation, and sharp increases in intensity were recorded at certain angles.

Bragg's law permits the atomic structure in crystals to be unravelled. If another set of planes is inserted at half the d spacing of the planes in Figure 23 and if these planes contain the same kind of atoms, then the prevailing conditions are as shown in Figure 24. If the angle employed is that for first-order diffraction without these additional planes—*i.e.*, $n = 1$ in Bragg's law—then the path difference between rays 1 and 2 is one wavelength (Figure 24A). But the path difference between rays 1 and 3 (to the inserted plane) is half a wavelength. The scattered waves 1 and 3 are out of phase (one wave is at a crest, the other is at a minimum) and their sum is zero. When $n = 2$ (Figure 24B) the scattering from the inserted plane differs in path length from plane 1 by one wavelength, and all planes are in phase. The second reflection is present and is even stronger than if there were no inserted plane because there are more planes (atoms) contributing. Reflections are labelled by multiplying the Miller (hkl) indices by n. Thus, a second order ($n = 2$) reflection from (100) planes is labelled (200). Bragg's law is rewritten:

$$\lambda = \frac{2d}{n} \sin \theta, \qquad (2)$$

and these (200) planes have half the spacing of the (100) planes.

Determination of a cubic structure

These simple ideas provided the solution of many inorganic structures almost immediately. Consider a cubic crystal with one of the three possible arrangements of atoms shown in Figure 20A and assume that there is only a single wavelength for the incident radiation. The spacing of planes (d) in terms of the length of the edge of the unit cell (a) and the indices of the reflection planes (hkl) can be expressed:

$$d = \frac{a}{\sqrt{h^2 + k^2 + l^2}}. \qquad (3)$$

Thus, Bragg's law can be restated by substituting this value of d as:

$$\lambda = \frac{2a}{\sqrt{h^2 + k^2 + l^2}} \sin \theta. \qquad (4)$$

Now λ is a known, fixed quantity, and the larger the values of h, k, l, the larger the angle θ. As shown in Figure 25, a face-centred structure contains (100) planes halfway between the faces of the cube. Thus, there will not be a (100) reflection as there would be for a cube with atoms only at corners; instead there will be a (200) reflection of larger intensity. The (100) reflection would also be missing for a cube with atoms at the corners and in the centre of the cube, however. How can the proper choice of structure be made? The (110) planes for the face-centred cubic crystal, simple cube, and body-centred cube also are shown in Figure 25. Because planes halfway between these occur only in the first case, a (110) reflection will not occur from the face-centred structure but, rather, from a simple cube or a body-centred cube. If the first reflection from the cube faces of a natural crystal is obtained, then a can be calculated from equation (4) above and the angle of the observed reflection, θ. Then θ can be calculated for the (110) planes from the same equation. The crystal can then be rotated 45° to get a reflection from these (110) planes, because the (110) planes are 45° from the (100) planes. If the crystal actually is face-centred cubic, then the first reflection from the cube face was really a (200), not a (100). If this were not known, then the calculated value of a from Bragg's law would be half as large as that of the real unit cell. If this incorrect value of a is used to calculate the θ position for the (110) reflection, the value obtained would actually be where a (220) reflection occurred, but this would not be known. If, however, the crystal were body-centred, the lowest angle reflection would be for a (110) reflection, which would be at a lower angle than calculated. Recalculation assuming that the first peak is a (200) reflection would give a correct value for a, from which the correct angle of the (110) reflection could be calculated. If this does not occur, then the crystal must be face-centred or simple cubic. Considering the (111)

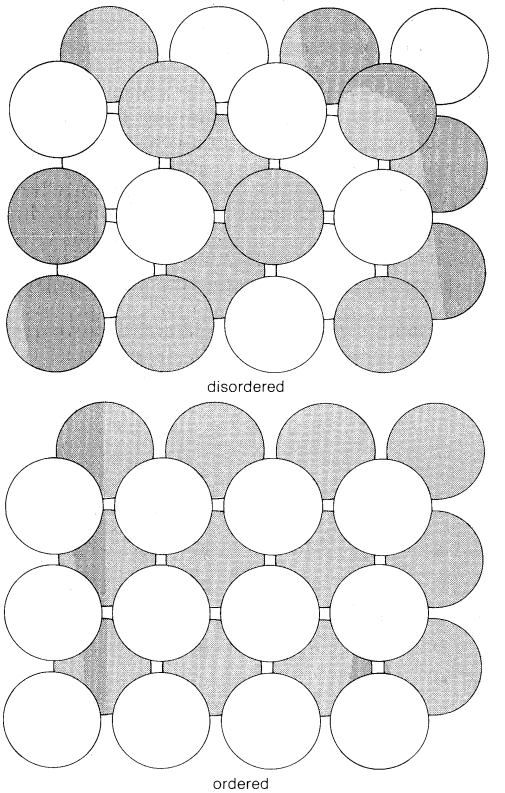


Figure 22: The body-centred cubic structure of beta brass, composed of zinc and copper atoms.

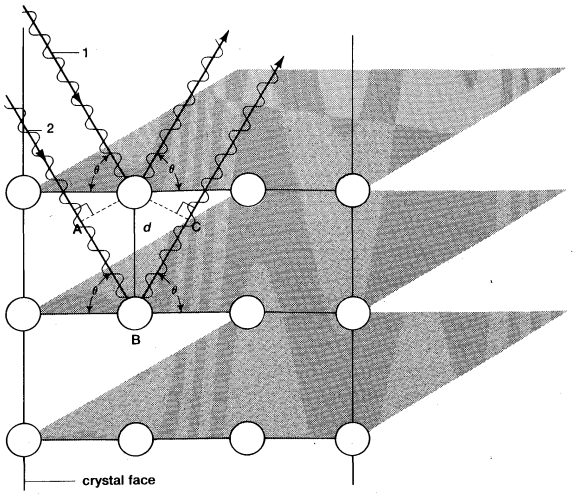


Figure 23: Incident rays (1 and 2) at angle θ on planes of atoms in a crystal. Rays will reinforce if their difference in path (AB + BC) is an integer times the wavelength (see text).

planes in the same way would distinguish these cases. If the atoms on these "in-between" planes are of a different species, a reflection will not be missing, only weakened. Scattering from different kinds of atoms is different, and the scattered waves do not exactly have the same heights, thus these waves cannot completely cancel the way the equal waves did in Figure 24.

The general method involved in such crystallographic detective work is repeated comparison of experiment and calculation for the angles and intensities of the various reflections. In this manner, it is possible to sort out some of the simpler structures. As the structures examined become more complex—involving in some cases thousands of atoms per unit cell—these simpler procedures provide only limited information. The progress of crystallography from the 1920s consisted of developing in detail systematic methods of sorting out the diffraction pattern. The process divides itself naturally into two parts. First, it is necessary to find all the diffraction spots (and to identify which ones are missing); and, second, it is necessary to record the intensity of these spots and devise procedures for interpreting these.

Missing reflections

Certain general information can be obtained from the indices of planes reflecting and not reflecting. If atoms are arranged with a face-centred cubic unit cell, for example, the only reflections that will appear are those with h, k, l all even or all odd. The (111), (200), and (220) reflections occur, but not the (100) and (110). Furthermore, the positions of reflections can be calculated with a single parameter, the edge of the unit cell. Other unit cells express themselves differently. Certain symmetry elements give rise to specific absences of reflections. A glide plane halves the spacing of planes perpendicular to it, and this requires h to be even for a reflection; for h odd, the molecules on the plane at half the spacing cancel the reflection. Such rules are tabulated for different symmetries and the different structures for easy reference. Those symmetries that are hard to detect from visual examination of a crystal—glides and screws—are revealed easily at this stage of a study. One way to determine which reflections are present and which are absent is to systematically tilt a crystal and hunt for these. From the external appearance of the crystal it is often possible to decide if it is a cubic structure or one of the other crystal systems. It is then possible to choose dimensions for the cell edges that would give the measured reflection angles and to decide on the indices of reflections present and those that are extinct. If the crystal is orthorhombic (see Figure 20), for example, then the a_1, a_2, a_3 axes (but not their lengths) can generally be assigned from the shape of the crystal. Then it is possible to mount the crystal to get the (100) or (010) or (001) reflections. It is known that the relation between the spacing of planes (d), the indices of the reflections (h,k,l), and the lengths of the edges of the unit cell (a_1, a_2, a_3) for orthorhombic crystals can be expressed:

$$\frac{1}{d^2}=\frac{h^2}{a_1^2}+\frac{k^2}{a_2^2}+\frac{l^2}{a_3^2}. \qquad (5)$$

Substituting for d in Bragg's law then yields:

$$\lambda = 2\sqrt{\left(\frac{h^2}{a_1^2}+\frac{k^2}{a_2^2}+\frac{l^2}{a_3^2}\right)}\ \sin\theta. \qquad (6)$$

The (100), (200), (300), or (400) reflection will permit determination of a_1; a_2 is obtained from (010), (020), (030), or (040); and a_3 from (001), (002), (003), or (004). These calculated values can then be tested in computing θ for other reflections, to decide whether a reflection was (100), or (200), etc., and to decide on absences. A more common procedure, though, is to surround the crystal with film, to catch many reflections at once while continually moving the crystal, and then to examine the pattern for dimensions and shape of the unit cell and missing reflections.

The powder method. It is not always possible to obtain single crystals of suitable size and perfection for structural studies with X-rays or neutrons. This is the major problem in extending structural studies to enzymes and other biological molecules. It is far easier to obtain powders. Also, most engineering materials, such as steel and ceramics, are polycrystalline. The way in which a diffraction pattern can be obtained from a random array of such grains or particles is illustrated in Figure 26. Uniform cones of diffracted intensity are produced instead of the spots characteristic of single crystals. An adequate specimen can be prepared by coating a hair with an adhesive and dipping it into the powder, which is, in effect, a randomly oriented array of crystals. A camera lens about 60 millimetres (2.4 inches) in diameter yields patterns of lines in a few minutes; they are recorded on a thin film wrapped in a cylinder around the sample to intersect the diffracted rays. Such patterns are quite useful in solving simple structures. Also, the spacing and intensity of the lines are characteristic of the material doing the scattering—just like a fingerprint. A file of more than 21,000 patterns is available, and the routine use of these for identification of unknown specimens constitutes one of the major uses of X-ray diffraction. One interesting field for this method stems from the fact that paint pigments contain crystalline materials, as do glazes. The composition of these has changed over the years with the art of paint manufacture, and it is sometimes possible to detect a forged art object from the diffraction patterns of the sample or to trace an automobile from a streak of its paint left after an accident.

Commercial importance

By exploring the entire cones of scattering, it is possible to learn how random a specimen actually is. This can be very important in controlling properties, because these depend on direction in a single crystal. Processing of a polycrystalline material is often adjusted in order to provide a certain deviation from complete randomness as an aid in obtaining desired properties in certain directions.

(J.B.Co.)

Imperfections in crystalline materials. Most crystalline materials are not perfectly regular. Accidents in growth

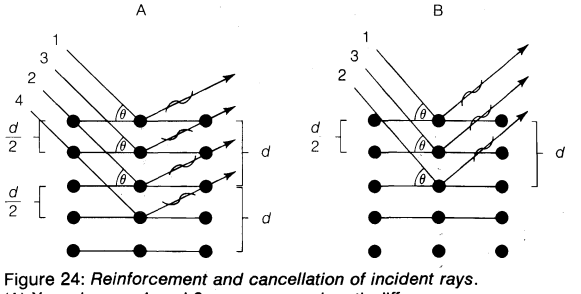


Figure 24: *Reinforcement and cancellation of incident rays.* (A) X-ray beams 1 and 2 are one wavelength difference in path; beam 3 to a plane at half the spacing is half a wavelength different in path with respect to 1, is thus out of phase, and cancels it. Similarly, beam 4 cancels beam 2, etc. (B) The angle of X-ray beam to crystal has been changed so that all rays are in phase and reinforce one another.

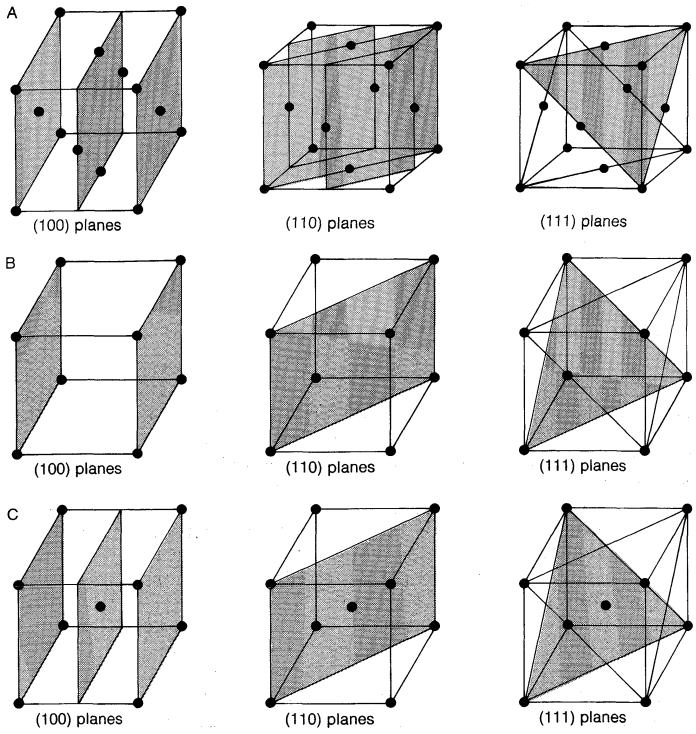


Figure 25: Various planes in (A) face-centred cube, (B) simple cube, and (C) body-centred cube. Note that there are planes of atoms halfway between cube faces in A and C but not in B. Similarly there are (110) planes in A halfway between those in B and C, and in C there is a (111) plane containing the atom at the cube's centre halfway between those in A and B.

lead to imperfections (see below *Crystallization and crystal growth*). Common among these are vacant lattice sites (vacancies) and missing parts of planes of atoms, as shown in Figure 28B. Mechanical and electrical properties can be highly dependent on the number and arrangement of such imperfections. Other types of defects can affect the optical properties of a crystalline solid. For example, among the most important defects in ionic crystals are colour centres, so called because, in many cases, the defect imparts a particular colour to the crystal. There are many kinds of colour centres that have been extensively studied, of which only a few appropriate to the alkali halides will be discussed.

Effects on optical properties

The simplest of these, called the F centre, consists of an electron trapped at a halogen vacancy. If the F centre traps a second electron, it becomes an F' centre. Two adjacent F centres form an M centre, while three F centres form an R centre. The F and M centres in an alkali halide are shown in Figure 27A.

From L.V. Azaroff, *Elements of X-Ray Crystallography*, copyright 1968; used with permission of McGraw-Hill Book Co.

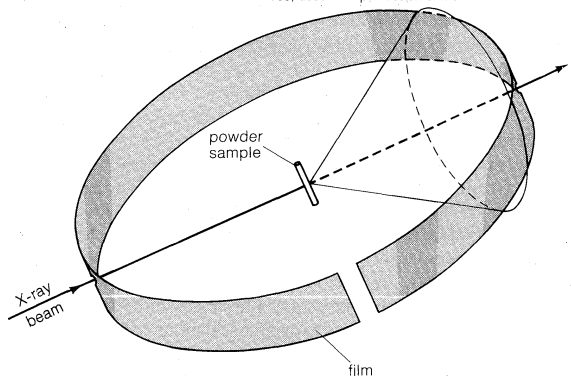


Figure 26: X-rays directed on a powder sample diffracting from the same planes in different grains. Because these planes are oriented differently in each grain, the result is a continuous cone of scattering at each diffraction angle. The continuous cones of diffracted radiation produce patterns of characteristic lines on a strip of film in a cylinder wrapped around the specimen.

A second class of colour centres is electron deficient. One of these, the so-called V_k centre (Figure 27B) involves the covalently bonded halogen molecule–ion mentioned above. An H centre is similar, except that one of the halogens is in an interstitial position. In the terminology discussed below, these involve self-trapped holes.

Colour centres can be produced in various ways. One method, involving ionizing radiation, is an extension of the vacancy formation mentioned above. The electrons that are freed in the ionization process are trapped at the halogen vacancies to form F centres. At the same time, self-trapped hole centres can be formed, since halogen molecule–ions are involved in formation of the vacancy. A second technique for producing F centres in alkali halides is additive coloration. In this method, the crystal is heated in an alkali vapour. Alkali atoms stick to the surface and combine with halogen ions that migrate from the interior to form new atomic layers of crystal but, in the process, leave behind halogen vacancies which can trap electrons to become F centres. A third class of defects involves impurity ions. In this case, there are many possibilities; positive silver and thallium ions are among the most widely studied. The commonest way of adding impurities is addition while the crystals are molten. If the molten crystal is cooled properly, the resulting crystal will contain a random distribution of impurities. (W.B.F./Ed.)

Observation by electron microscopy and X-ray methods. Many materials have a sufficiently low density of defects to make attractive an exciting new expression of diffraction. With the electrons in modern electron microscopes operating at 50,000 to 1,000,000 volts, it is possible to penetrate foils of materials about 1000 angstroms and make pictures at magnifications of 1,000,000 times. Such foils can be prepared by electropolishing. The wavelengths of the electrons at these voltages are about 0.01 angstrom (the value depends on the voltage of the microscope). The diffraction angles, from Bragg's law, are very small, approximating one degree. Small tilts near an imperfection, such as the dislocation in Figure 28, can bring a group of planes into the proper orientation for diffraction. The diffracted beam is blocked, but the unscattered beam passes through an aperture. A dark line appears in the position of the dislocation on a fluorescent screen which is illuminated by the electrons to form the image. It is also possible to heat, cool, or deform the specimen in the microscope and observe the change in number and arrangement of defects, as well as their interaction.

X-ray techniques

With X-rays, quite similar pictures can be taken. One advantage of this approach is that it can be employed in reflection on thick specimens rather than on the foils needed for an electron microscope. One major disadvantage, however, is that the pictures are not taken at any enlargement. Photographic enlargement of the best of films is possible up to only about 200 times. Because this is far removed from the magnification of an electron microscope, the density of defects must be much lower to resolve them. The principle of the method for the longer wavelength X-rays is also somewhat different. On scattering from a plane, there is a 90° phase change of the scattered wave compared to the incident wave. If a region is nearly perfect, the diffracted beam will rescatter back in the direction of the incident beam before leaving the crystal, but the 90° change in phase a second time leads to a total change of 180°; this twice-scattered beam is thus out of phase with the incident beam and reduces it, lessening its penetration into the specimen. This effect reduces the diffracted intensity in two ways. First, less material "sees" the incident beam; and, second, some of the diffracted beam is rescattered. In the vicinity of an imperfection, less of this occurs, and the scattered intensity is larger than for the perfect regions, making a darker region on a film placed above the specimen. (J.B.Co.)

CRYSTALLIZATION AND CRYSTAL GROWTH

Crystallization is the process that results in the formation of crystals. Crystal growth is the enlargement of crystals at the expense of materials in contact with them.

Crystals and crystallites

In most crystalline solids, small, individually ordered regions, which are called crystallites, are oriented differ-

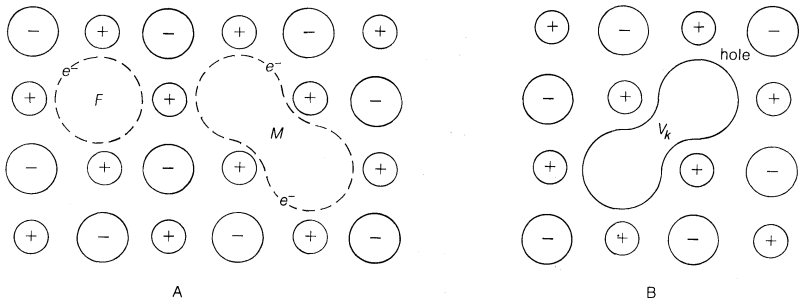


Figure 27: *Colour centre defects.*
(A) The *F* centre, consisting of an electron trapped at a halogen vacancy, and the *M* centre, which is two adjacent *F* centres. (B) The electron-deficient V_k centre produced by formation of a covalent bond by halogen ions.

ently from one another, with the result that the crystalline pattern is repeatedly disarranged; such materials are said to be polycrystalline. In certain rare crystals, such as the gemstones, and in crystals man takes great care to prepare, however, a single ordered region makes up the entire solid object; in such cases a single crystal may range in size from fractions of a centimetre to many centimetres. One of the identifying features of single crystals is that they usually have flat planes or faces that are set at fixed angles to one another. Quartz crystals, and other crystals displaying such regular faces, are most commonly seen in mineralogical collections.

Crystals are formed in many natural processes. In the atmosphere, for example, ice crystals are formed by the direct freezing of cloud droplets or by the freezing of water vapour onto particles of dust. When the cooling is slow, large, single crystals of ice occur in frozen lakes. When cooling is rapid, however, many small crystals, or crystallites, form, and the solid blocks of ice are polycrystalline. Almost all crystalline solids of nature and commerce are polycrystalline rather than single crystals.

Comparatively slow rates of formation are required to produce large single crystals free of crystallite grains. Other more subtle imperfections occur to some extent in all crystals; these may be missing atoms or vacancies in the crystal structure, or foreign atoms present at sites that should be occupied by the atoms of the material located between the atoms of the material.

From (B—D) C.A. Wert and R. Thompson, *Physics of Solids*, 2nd ed. 1970 (copyright 1964); used with permission of McGraw-Hill Book Co., (E) J. Ziman, "The Thermal Properties of Materials." Copyright © 1967 by Scientific American, Inc. All rights reserved.

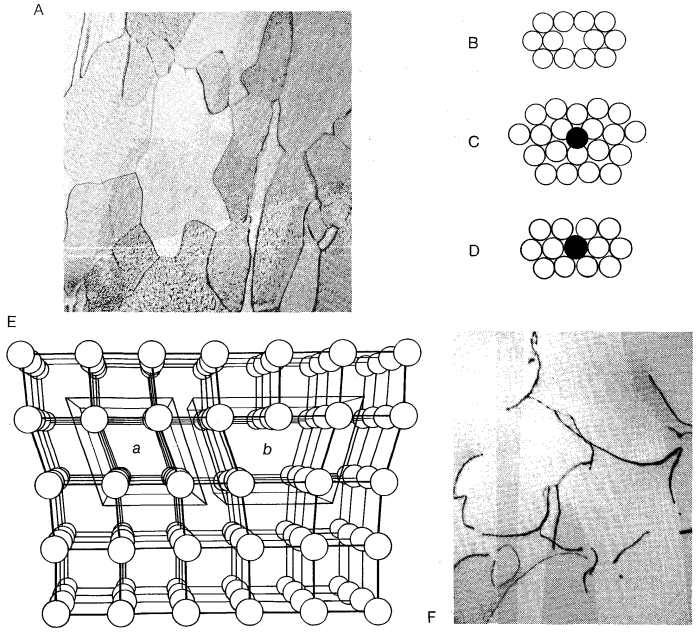


Figure 28: *Crystalline lattice defects.*
(A) Grain structure in polycrystalline iron. The orientation of the grains changes across the grain boundaries (magnified 100 ×). (B) Lattice vacancy. (C) Interstitial atom. (D) Substitutional solute atom. (E) Edge dislocation in simple cubic lattice. The extra half plane is shown at position *b*, which indicates the distortion and disruption of bonding at the dislocation. At position *a* the lattice is perfect except for elastic distortions. (F) Electron microscope photograph of dislocations in vanadium (magnified 100,000 ×).

Dislocations

In addition to the above atomic or "point" defects, crystals may have more extensive "line" defects, in which the disorder or dislocation extends along a line or plane of atoms. These are of two principal kinds: edge dislocations and spiral, or screw, dislocations. Figure 29, top, is a representation of an edge dislocation. The imperfection may be thought of as resulting from the addition of an extra plane of atoms, ABB′A′, in the upper half of the crystal. Because of the added atoms in the upper half of the crystal, this portion is strained by compression; correspondingly, the lower half of the crystal is strained by tension as the fewer number of atoms seek to fill the same volume as that filled in the upper part of the crystal. If the dislocation consists of a gradually increasing shift of alignment between planes (as illustrated in Figure 29, bottom), the imperfection is called a screw dislocation. Dislocations are created when existing crystals are deformed or when new crystals are growing. In fact, dislocations are important in growth mechanisms. Under careful control several materials, including silicon (Si), germanium (Ge), gallium arsenide (GaAs), and quartz (silica, SiO_2), have been grown in the form of crystals free of dislocation, but theoretical considerations require that all crystals have vacancies except at absolute zero temperature (−273.15° C [−459.67° F]). Indeed, the more rapidly a crystal is grown and the less pure its growth environment, the greater the number and variety of imperfections it acquires. In relatively perfect crystals the misorientation between adjacent grains may be undetectable—that is, the only imperfections are point defects and dislocations—but as the conditions of formation become less favourable, misorientations increase to the point where a polycrystalline mass is the product.

Single crystals

Single crystals are important in science because measurements of the fundamental properties of matter in such a highly ordered state are interpreted more easily than when irregularities in structure are present. Impurities that often segregate at the grain boundaries in polycrystalline materials may mask such subtle properties as semiconductivity of electricity. Properties of crystals that depend on the direction through which the crystal measurement is made (such as refraction of light) either are not exhibited in polycrystalline material or are much harder to interpret. Partly for this reason, single crystals are of especial interest in electronics. Boundaries between grains in polycrystals scatter light to produce opacity, thereby generally making polycrystalline materials unsuitable for applications requiring optical transmission. Similarly, the laser and related devices have created new demands for single crystals for research and applications. Devices that depend on semiconductivity, such as transistors, utilize single crystals, as do devices such as piezoelectric oscillators and filters, which depend on unsymmetrical distribution of properties (anisotropy) with regard to the crystal structure. Indeed, it is fair to say that the electronics industry could not exist without single crystals. In addition, of course, the intrinsic beauty of natural crystals has fascinated man from prehistoric times to the present and the manufacture of synthetic gemstones is an expanding area of crystal technology.

Disturbing of equilibrium

All crystals are grown by disturbing a balance, or equilibrium, that exists between the crystalline material and the gas, liquid, or amorphous solid that exists in contact with it. Under the conditions of the equilibrium, the relative amounts of the crystal and the material in contact with it do not change. Disturbing the equilibrium by altering the conditions, such as lowering the temperature in the case of an equilibrium between water and ice, may result in an increase in amount of crystal with respect to the noncrystalline material.

The material that is crystallizing may be the only substance present in the growth system or other substances may be added to lower the temperature of the system or otherwise alter its properties.

Crystal growth from pure, or monocomponent, systems presents several important advantages and is widely practiced. The purity of the crystals obtained is high, because no additional components are present to be incorporated during the growth process. Furthermore, the speed of growth can be relatively rapid because, in the absence of additional components, slow movement of the additional components away from the growing crystal surface is not required.

The most common monocomponent growth process is growth by the freezing of a pure liquid, or melt, of the material in question. In contrast to most other growth processes, freezing of a pure melt is uncomplicated and readily controllable; it is probably the most widely practiced commercial growth process. Freezing is also the easiest process in which to consider the distribution of undesired impurities and impurities deliberately added to impart desired properties (dopants) to the crystal. The most important consideration regarding distribution of impurities is whether the crystallization process is conservative or nonconservative. In a conservative process no material is added to or taken from the melt except, of course, for the crystalline material that separates as a result of the freezing process. In a nonconservative process, however, additional material may be melted into the liquid during crystallization, or material may be lost from the melt as, for instance, by evaporation during the process. (For discussion of actual growth methods, see below *Preparation of crystals*.)

Conservative and nonconservative systems

From (Top) R. Laudise, *The Growth of Single Crystals* (© 1970) Prentice Hall, Inc.; (Bottom) W. Read, Jr., *Dislocation of Crystals* (Copyright 1953), used with permission of McGraw-Hill Book Company

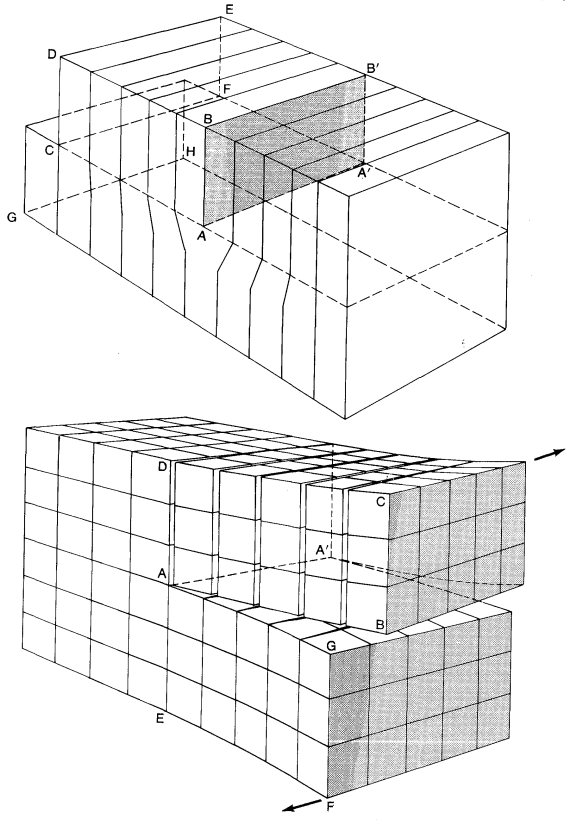


Figure 29: (Top) Edge dislocation. The planes of atoms in a single crystal are represented as vertical layers. AA′ is the dislocation line and the dislocation can be produced by pushing (shearing) on CDEF while pulling on GCFH. (Bottom) Screw or spiral dislocations. AA′ is the dislocation line. The dislocation can be thought of as having been created by pushing ABCD while pulling AEFG.

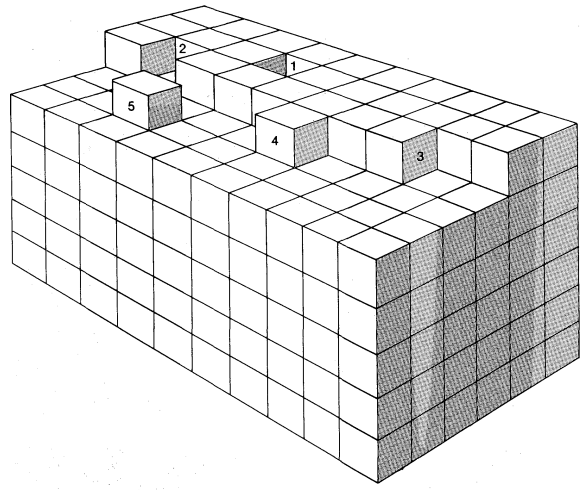


Figure 30: Positions for the addition of new atoms at the surface of a growing crystal (see text).

Theoretical aspects of crystal growth. Extensive theoretical work on crystal growth has been conducted, chiefly in the 20th century. One of the most important considerations seems to be the energy released by the attachment of atoms at different sites on a crystal face.

Energy considerations. When a new atom is added to a growing crystal, the most probable position for addition will be one that is energetically most favourable. Since energy generally is liberated during bond formation, the most favourable position for adding a new atom will be at the point where the greatest number of bonds to adjacent atoms can be formed, thus providing the most potential energy. Figure 30 may be considered as an example of a developing crystal structure in which the atoms are identical and are represented as cubes. In this structure, the atoms are closely packed so that each has six nearest neighbours, which touch it on the faces of the cube. In this example, an atom is considered to bond only to its nearest neighbours. The most favourable position for the addition of a new atom, therefore, is at a "hole" in the crystal structure, such as position (1), at which the atom will be surrounded by neighbouring atoms on all sides but one. The next most favourable position is at a "notch" in one step of the crystal structure, such as position (2), in which location the new atom will find four neighbours. Growth at either of these positions, of course, cannot be repeated. The next most favourable position for the addition of a new atom is at a "kink" in a step, such as position (3), at which the new atom has three neighbours. In this case, the addition process can be repeated until the step in the structure is completed. The next most favourable position for addition is at a jog at the front of a step, position (4), and the least favourable position for addition, based on nearest neighbour considerations, is at a lone site on the surface, position (5).

Ionic crystals

Complications in the above picture arise in the case of ionic crystals, which are composed of ions carrying opposite electrical charges, because when an ion of a given charge approaches the face, it is attracted by ions of opposite charge while at the same time it is repelled by ions of like charge. Similar considerations apply to crystals composed of molecules bearing separate regions of positive and negative charge.

The problem of how crystallization begins on an already formed atomically smooth crystal surface is one that has received great attention. As seen in the discussion of Figure 30, the implantation of the first unit of a crystal, a process called nucleation, is energetically less favourable at a new

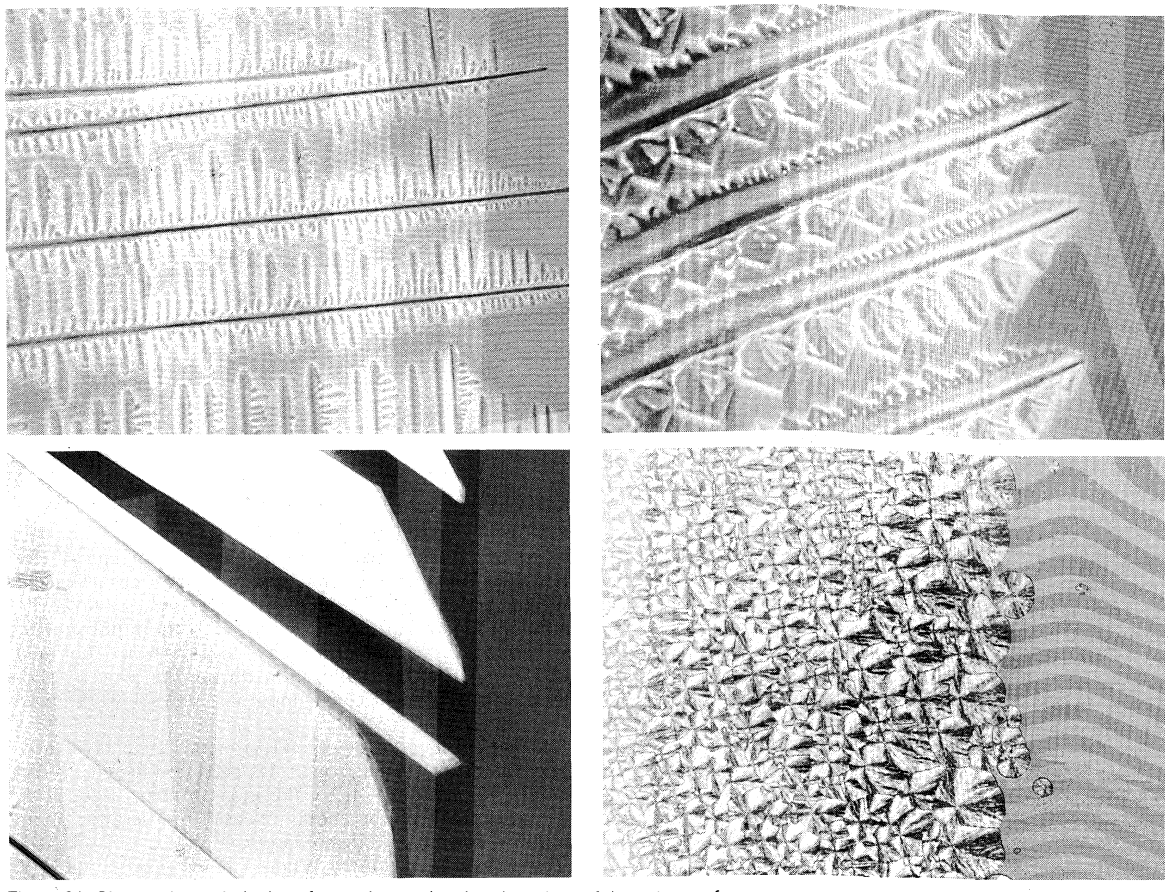


Figure 31: Observed morphologies of crystals as related to the values of the entropy of fusion. As the entropy of fusion increases, it is possible to observe (top left) dendrites (carbon tetrabromide), (top right) facetted dendrites (*tert*-butyl alcohol), (bottom left) facets (benzil), and (bottom right) spherulites (tristearin).

By courtesy of Bell Telephone Laboratories

step than at a kink in a step already partially formed. Once a step has grown across the surface, however, nucleation of a new step is required. Quantitative calculations indicate that either large supercooling (cooling far below the melting point) or supersaturation (addition of more of the crystallizable solid to a solution than it can hold) is required for the nucleation of a new step on an atomically smooth surface. Nevertheless, crystals are observed to grow at conditions of low supercooling or low supersaturation, contrary to what might be expected. What is obviously required in such cases is a step in the crystal structure that grows without coming to an end; that is, indefinitely. It was only recently realized that spiral dislocations, of the type pictured in Figure 29 (bottom), supplied locations at which growth could continue without completing the regular crystal structure. Careful observation of crystals grown under different conditions provided an understanding of how such growth took place.

The above energy considerations generally apply to circumstances in which a crystal is grown extremely slowly, so that new material is deposited on the crystal surface strictly in terms of the energy requirements of the system. In most cases, however, the growth of crystals is too rapid to achieve equilibrium, or balance, between the rate of addition to the crystal and rate of removal from it, so that the observed shape of crystals usually depends on the growth rate of the various individual faces. In the 1960s it was shown that the shape, or morphology, of the crystal could be correlated with a property known as the entropy of fusion, entropy being a measure of the loss of energy from an organized system, resulting in an increase of disorder. In particular, it was shown that the entropy of fusion could be used to distinguish four classes of materials: (1) in which growth is equivalent in all directions (isotropic) and cells and dendrites (see below) occur when impurities are added; (2) in which growth occurs as facets, or as facetted dendrites, when sufficient impurity is added; (3) in which facets are always observed; (4) in which growth of spherulites is usual. Cells, or bumps on the surface, result from an instability caused by supercooling. These produce a fairly regular pattern of segregation of impurities within the crystal. Dendrites are crystals with a branched, treelike morphology resulting from growth that is controlled by the rate of diffusion. Facets are flat faces parallel to the crystal planes. Spherulites are oriented spherical masses of crystals. Figure 31 illustrates some of these structures.

Classification by entropy of fusion

Constitutional supercooling. Supercooling that results from concentration changes within the crystallizing system is called constitutional supercooling. Constitutional supercooling is important in the crystallization of all polycomponent systems and has been much studied, especially in melt systems. When growth takes place in the presence

From R. Laudise, *The Growth of Single Crystals* (© 1970); Prentice Hall, Inc.

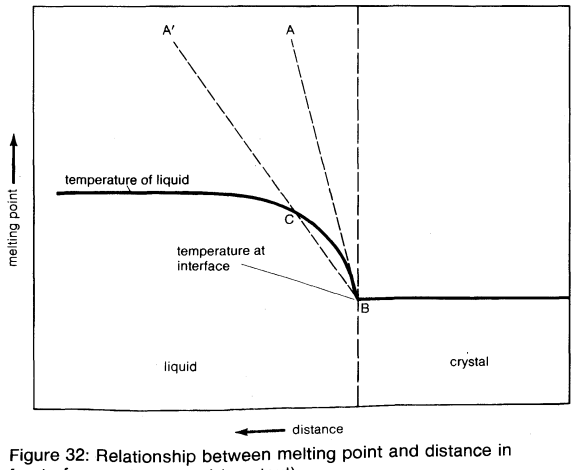


Figure 32: Relationship between melting point and distance in front of a growing crystal (see text).

of a very high concentration of impurities, or in a polycomponent system, there is a buildup of the components that are rejected from the growing crystal interface. Crystallization, therefore, depends on the rate of diffusion of the impurities away from the growing crystal surface. If crystal growth is so slow as to be almost at equilibrium, diffusion has ample time to remove impurities, and the effect of diffusion, therefore, is negligible. In many cases, however, the diffusion problem is severe. Figure 32 shows the melting points of the compositions existing near a growing crystal as a function of distance (in a system in which diffusion is important). As can be seen, the melting point of the solution decreases as the interface is approached. The broken lines AB and A′B are two different possible temperature gradients in the solution. For the larger gradient, AB, at all positions in front of the interface the temperature is above the melting point (*i.e.*, there is no supercooling); for the smaller gradient A′B, the region CB is below the melting point; it is supercooled, and crystallization tends to take place in front of the solid interface. The region CB is referred to as constitutionally supercooled, and large temperature gradients, such as can be obtained easily in crystal pulling (see below *Preparation of crystals*), lessen the tendency toward constitutional supercooling. In a crystal growing from a solution in which the liquid is maintained at uniform temperature, supercooling is least near the crystal and increases farther from it. Supercooling should lead to instability of the crystal surface.

A crystal face that has a very slow intrinsic growth rate is not as susceptible to instability as are more rapidly growing faces of the same crystal. The crystal faces that become unstable under these conditions are the fastest growing faces and, as a result, they break up into small segments of more slowly growing faces. Under such conditions, the peaks of some of the facets will be in regions where the supercooling, or supersaturation, is greater than the valleys. The peaks then grow ever more rapidly, with a tendency toward extreme dendritic growth.

The behaviour of eutectics. A eutectic is a mixture of two phases that freeze together in a fixed ratio and at a constant temperature. A great variety of structures can be found among eutectics; the common feature, however, is that two phases can always be seen. These two phases usually form a structure in which they alternate by growing around and over each other, starting from a single pair of nuclei. The structures can be broadly classified as regular or irregular. The former includes both layered, or lamellar, structures and rod types, in which one phase is rod-shaped and is imbedded in a continuous matrix of the other. The irregular structures include a variety of complicated intertwinings of the two phases, as well as discontinuous eutectics, in which one phase consists of isolated crystals imbedded in a continuous matrix of the other. If the intrinsic growth rate of the two phases is isotropic, eutectics can crystallize in the lamellar form with regular spacing between the lamellae. The boundaries between the lamellae are parallel to the growth direction; that is, lamellae of both phases crystallize at the same time. The lamellar widths are determined by the interaction of two opposing considerations: (1) diffusion of solids must occur over great distances to form thick lamellae, a factor that favours narrow lamellae, and (2) the narrower the lamellae, the more surface energy that is associated with boundaries between the two phases, a factor that favours wide lamellae. Rods occur whenever the volume fraction of one of the phases is small. Many complex and irregular eutectic structures result from unevenly directed growth rates of the two phases.

Lamellar growth

Nucleation and impurities. The formation of a crystal in the absence of a seed involves nucleation, or initiation of a crystalline centre. When nucleation occurs on a surface, such as a wall of the container, or on a foreign body, such as a dust particle, it is called heterogeneous nucleation. If the structure and interatomic spacing of the surface on which heterogeneous nucleation takes place approximate that of the crystal, growth on the surface can resemble growth on a normal seed crystal, a phenomenon called epitaxial growth. When nucleation occurs in the absence of a surface—*i.e.*, in the bulk of a liquid—it is said to be homogeneous.

In some cases, nucleation can be looked upon as the aggregation of particles, and precursors of nuclei large enough to grow are formed by association of particles in the system. If the number of particles in the aggregate and the equilibrium constant for the crystallization reaction are large enough, the nucleus will tend to grow instead of redissolving. Aggregates too small to grow in the average supersaturation in the medium are sometimes called embryos. Most embryos never survive to grow. A few do so, however, because of statistical fluctuations of supersaturation or because energy is added locally to assist their growth (as, for example, when a beaker containing a supersaturated solution is scratched with a stirring rod). The higher the supersaturation, the larger the average embryo size and the greater the chance that statistical fluctuations will allow nucleation. Given enough time, there is a statistical chance that nucleation will occur in any supersaturated medium. In the presence of dust, of course, nucleation is heterogeneous and comparatively easy. Some form of particulate contamination appears to be inevitable in most actual nucleation. The deliberate, controlled addition of impurities is very important in controlling the properties of commercial crystals. In semiconductor crystals the type and degree of conductivity are controlled by the nature and concentration of impurities present.

Crystal embryos

Preparation of crystals. *Monocomponent methods.* Growing crystals in systems consisting of only one chemical component present in both the solid and liquid phases can be considered to be conservative or nonconservative depending upon whether the amount of material in the system is or is not altered during the process. Figure 33 illustrates the important conservative and Figure 35 the important nonconservative crystal processes. In a typi-

From R. Laudise, *The Growth of Single Crystals* (© 1970); Prentice Hall, Inc.

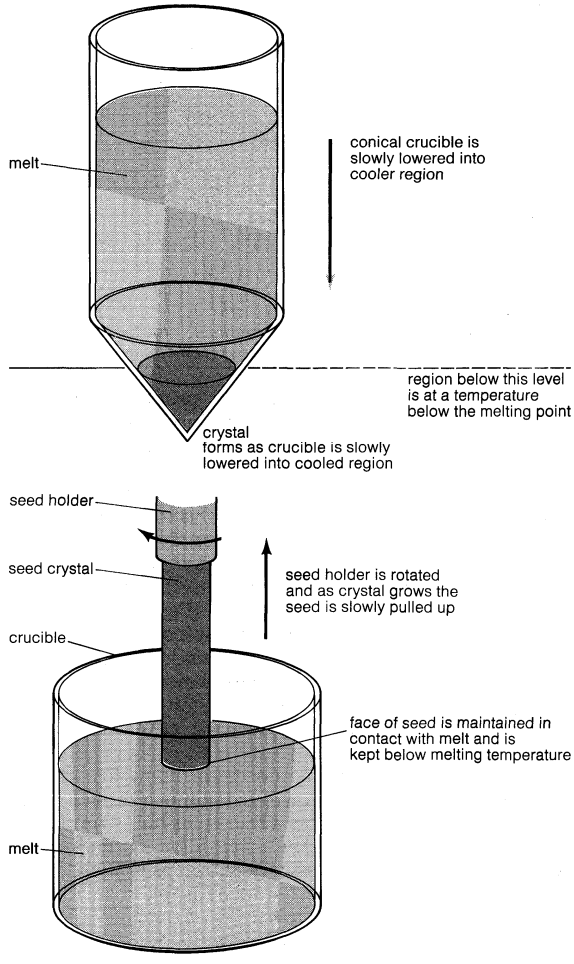


Figure 33: *Conservative growth processes.*
(Top) Bridgman-Stockbarger method, (bottom) Czochralski method.

cal example of one of the conservative (the Bridgman-Stockbarger) methods (Figure 33, top), the container is a cylindrical crucible with a conical bottom. The material to be grown is melted in the crucible and lowered through a furnace with a temperature gradient such that the first freezing takes place in the conical tip of the crucible. In the ideal case, one crystallite comes to dominate the liquid–solid interface and this interface moves through the melt as the crucible is lowered.

Another conservative method is the crystal-pulling (Czochralski) method (Figure 33, bottom), in which a seed is introduced into a melt contained in a crucible arranged so that a temperature level corresponding to the melting point lies just above the liquid surface. The seed is then slowly withdrawn from the melt and growth takes place on it. In a third conservative (Kyropoulos) method a cooled seed is immersed in a melt and the temperature of the melt is lowered to cause growth.

Zone melting

Zone melting is an example of an especially powerful nonconservative process, which is used for purification and for the growth of single crystals. In horizontal zone melting (Figure 34A), the material is held in a boat-shaped crucible and the temperature profile is arranged so that a narrow molten zone is produced. This zone is moved down the crucible rather slowly in order to effect a purification or to grow a crystal. If further purification is required, the zone is reestablished at the front end of the crucible and the process is repeated. Seeding may be effected by placing a seed at the front end of the crucible and beginning the zone so as to avoid completely melting the seed. Alternatively, the ingot to be melted can be arranged vertically (Figure 34B). This technique, called float zoning when no container is used for holding the melt, allows purification, homogenization, or growth without crucible contamination. The molten zone is held in place by surface tension. The diameter of the melting ingot and the growing crystal need not be the same.

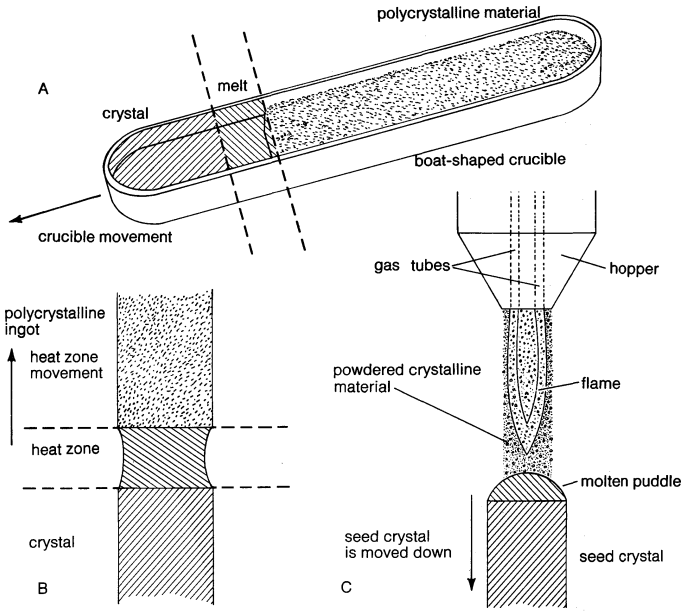


Figure 34: *Nonconservative growth processes.*
(A) Horizontal zone melting using a boat-shaped crucible, (B) vertical zone melting, (C) Verneuil method or flame-fusion.

Flame-fusion process

Another nonconservative method is the flame-fusion process (Figure 34C). In this growth technique a molten puddle, which is held in place by surface tension, is produced on the surface of the seed by a flame, or plasma, or by radiant heating from a focussed high-intensity light. Solid particles of the material or molten droplets are added to the puddle.

Gas–solid (or vapour transport) growth in a monocomponent system, a process called sublimation, is of relatively limited applicability. Gas–solid growth in a polycomponent system (growth by reaction) is, however, a very powerful technique. Both reversible and irreversible reactions may be used.

Growth by means of solid–solid transformation is difficult to control because the density of sites (nuclei) for growth in a solid is so high that growth tends to occur at many places in a transforming solid, with the result that good quality single crystals usually are not obtained. The most generally useful method is strain annealing, a process in which a polycrystalline material is deformed plastically by stretching it and then heating (annealing) the specimen to promote crystal growth as strain is relieved. Transforming an amorphous material to a single crystal by direct conversion from a glassy to a crystalline state (devitrification) is possible, but the high density of nuclei in amorphous materials has so far prevented the use of this method for the growth of crystals.

If a particular material can exist in several different structural modifications (polymorphs), it is often possible to convert a single crystal of one form to the other form by carefully moving the specimen from a region of temperature stability for the one polymorph to that of temperature stability for the other. The geometry and technique is analogous to that used, for instance, in lowering a liquid–solid sample through a temperature gradient (Figure 33, top). Such polymorphic transitions result in twins in the newly formed polymorph, or even in polycrystallinity, when the structures of the two polymorphs are too dissimilar, twins being composite crystals in which the individual parts are related to one another in a definite crystallographic manner. Sometimes appropriately applied stresses can be used to "de-twin" samples or convert them to single crystals.

Polycomponent methods. Polycomponent methods are particularly important in growth from liquids, in which the addition of a second component can lower the melting point of refractory material to a point where advantages in ease to growth are obtained. Such growth is ordinarily called growth from solution, and the driving force that makes it possible is supersaturation.

A useful classification of the growth methods can be made on the basis of the procedure used for bringing about supersaturation. It includes both isothermal and non-isothermal methods. The isothermal methods are: (1) solvent evaporation or change in solvent concentration, (2) temperature differential, and (3) chemical or electrochemical reaction. A non-isothermal method is slow cooling. Isothermal methods have an advantage over non-isothermal methods in that the concentration of impurities (or dopants) in the grown crystal is uniform. Non-isothermal methods, however, usually require less elaborate equipment and generally do not utilize such long diffusion paths, so that growth rates can be higher. Temperature-differential methods are capable of producing the largest crystals because, in this case, crystal size is not limited by the solubility of the material being grown.

Isothermal methods

Another convenient method of classifying solution-growth methods is based on the nature of the solvent used. This may be (1) water at ambient conditions, (2) water at high temperature and pressure (hydrothermal conditions), (3) molten inorganic salts at high temperature (flux) conditions, or (4) other nonaqueous solvents.

Figure 35 shows the various solution growth methods. The most generally used aqueous-solution growth apparatus is the glass-jar crystallizer (Figure 35A). In this apparatus a saturated solution is exposed to suitably mounted seeds and the solution is cooled a few degrees every day to produce growth. The seeds are rotated at the rate of a few revolutions per minute, first in a clockwise direction and then counterclockwise, to aid in mixing and to hasten diffusion.

If the solubility of a material is so low (less than 5–10 percent) that it makes growth by other methods impractical, growth by chemical reaction can be used; for instance, although the salt calcium tartrate is virtually insoluble in water, tartaric acid and calcium chloride are appreciably soluble, and the reaction between them can be used to form crystals of calcium tartrate. Direct mixing of solutions of tartaric acid and calcium chloride, however, inevitably results in such a large local supersaturation that many centres for growth form, and the product will

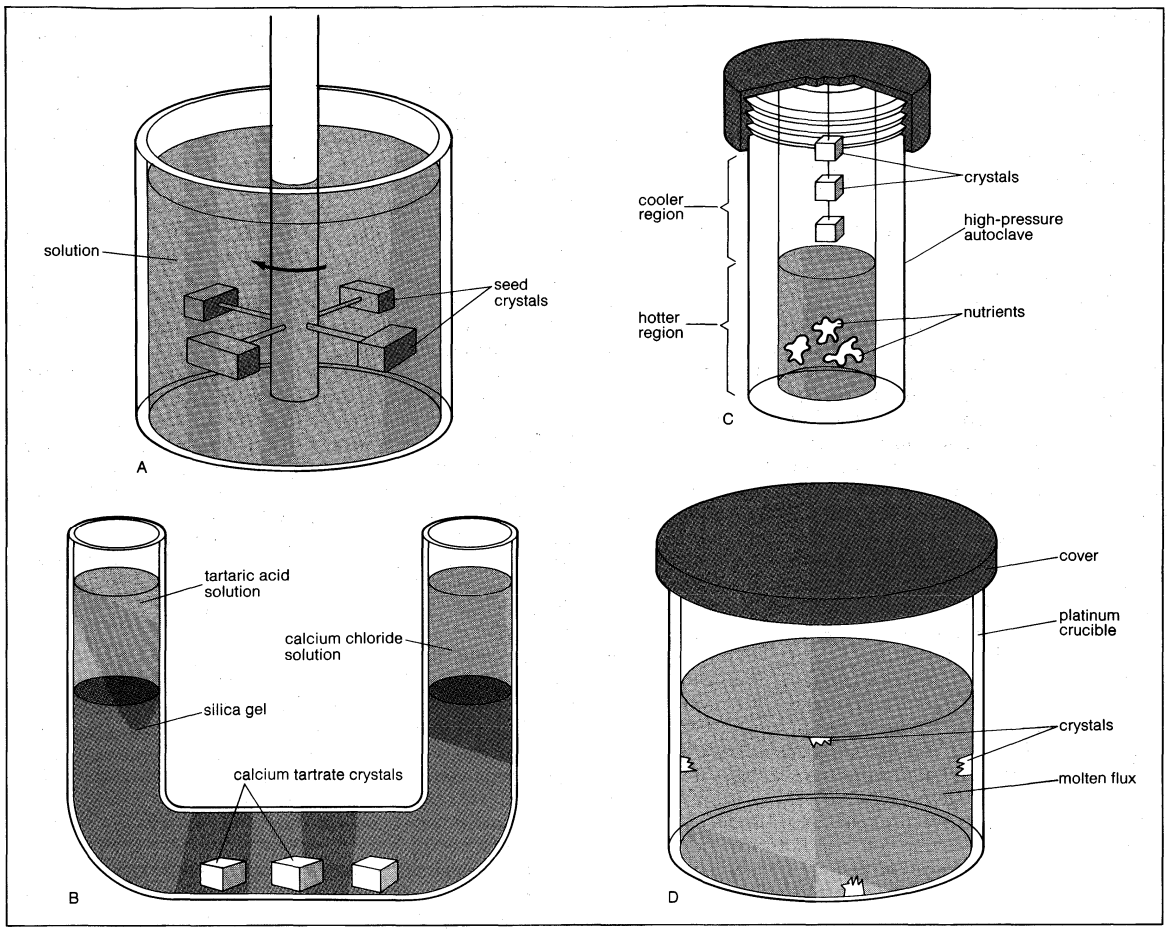


Figure 35: *Solution growth methods.*
(A) Glass-jar crystallizer; growth is produced by gradually lowering the temperature of the aqueous solution. (B) Gel growth; growth is by the reaction of the two solutions that diffuse toward one another through the gel medium. (C) Hydrothermal growth; the high-temperature–high-pressure aqueous solution moves by convection to the cooler region where it is supersaturated and growth takes place. (D) Flux growth; growth is produced by gradually lowering the temperature of the molten salt solvent.

consist of many small crystallites. In a similar reaction, however, solutions of calcium chloride and tartaric acid are allowed to diffuse through a substance called silica gel from the opposite ends of a U-tube (Figure 35B); the region of high supersaturation is localized near the centre of the tube, and a few comparatively large crystals are formed. This method is called gel growth and is quite useful for the growth of insoluble materials by means of a chemical reaction.

Single crystal growth by an electrochemical reaction, although little used, in principle also could be an important means of preparing certain crystals. Careful control of voltage and current, for example, could be used to grow single crystals of metals on an appropriate single-crystal seed electrode.

Hydrothermal method

Under hydrothermal conditions the increased solvent power of water at high temperatures is used to dissolve refractory materials. The water is heated in a high-pressure autoclave where it expands and fills the vessel (Figure 35C). Depending on the temperature and on the percent of the free volume of the autoclave initially filled with water, the pressure can range from a few hundred to many thousand atmospheres. The solvent power of water can be increased further by the use of "complexers"—substances that react with the material to be crystallized in such a way as to form soluble complexes.

Flux method

In flux growth a molten salt is used as the solvent (Figure 35D). Generally, flux growth is carried out in the absence of deliberately added seeds. Initial nucleation occurs spontaneously when an appropriate saturation is reached in the cooling cycle, and these nuclei provide the centres for further growth.

Gas–solid polycomponent growth has been referred to above (see *Monocomponent growth*). Solid–solid polycomponent growth (by solid-state precipitation) is difficult to control because of the high density of nuclei in solids, and it is slow because solid-state diffusion is slow. As a result, the process has not been used.

Vapour–liquid–solid growth makes use of a gas-phase reaction for the transport of material to a liquid solvent. The material dissolves in the liquid solvent, supersaturates it, and deposition takes place by a transformation from liquid to the crystal.

The choice of growth method. An important problem in crystal growth is the choice of the most suitable growth method. Generally, monocomponent methods are to be preferred because of their comparative rapidity and simplicity. In them, growth rate is limited essentially only by the rapidity with which the heat produced by crystallization can be dissipated. The purity of crystals grown in monocomponent systems is high, because the growth system does not contain additional components to contaminate the grown crystal. If the temperature required for melting or sublimation is too high, however, additional components must be added to produce growth from solution or by vapour-phase reaction. The reasons for growing crystals at lower temperatures include: (1) avoidance of undesired high-temperature polymorphs, (2) elimination of decomposition or irregular melting behaviour, (3) elimination of high vapour pressure, which would be present at the melting point, and (4) convenience of operation. When an additional component is added to make possible lower operating temperature, however, the rate of growth usually must be reduced, in order to accommodate diffusion of the growth component to, and the additional component from, the growing interface. If the additional

Limiting temperatures

component has appreciable solubility in the growing crystal, its inclusion may become a problem. (R.A.L.)

Amorphous solids

Glasses and also certain semiconductors (see below) and many plastics are amorphous. As noted earlier, such solids do not have a definite lattice pattern of constituent atoms and molecules.

The amorphous state of matter, also called the glassy state, is perhaps best defined by the process that produces it: when a substance in the liquid state is cooled to a rigid condition without any crystallization having taken place, the substance is said to be amorphous. As a liquid cools, the vibration of its atoms due to heat decreases and their spatial arrangement becomes less open. A steady contraction in the volume of the liquid results. At the freezing point, a liquid that crystallizes releases heat (*i.e.*, heat of fusion) with a sharp reduction in total volume as the atoms are rearranged into the crystal state. On further cooling, the crystal contracts much less rapidly because no more changes take place in the configuration of the atoms, even though the atomic, thermal, or heat, vibrations still continue. If the liquid can be cooled below its freezing point without crystallization, the supercooled liquid will be in internal thermodynamic equilibrium, in contrast to the condition of a crystallizing liquid at the freezing point. This means that the available free energy of the supercooled liquid is less than the available free energy in any other structure into which the atoms could rearrange themselves and, therefore, the supercooled structure will not by itself change. It needs energy to do so and it can remain in a precariously stable, or what is called a metastable, state until the energy is supplied. If complete nuclei form, or if crystals are dropped into the supercooled liquid, freezing will begin at those points with consequent release of heat of fusion. The system will no longer be in internal equilibrium and crystallization will spread through it.

(Ed.)

GENERAL PROPERTIES

Volume and temperature relationship. Figure 36 shows typical volume changes plotted against temperature changes, and the resulting graphs give a general description of the glassy state. A liquid at A is cooled to the freezing point at B. The volume shrinks as indicated from A to B and then contracts drastically from B to C without change in temperature (heat must be constantly removed from the system), as the liquid turns into a crystal. Further cooling results in moderate contraction as shown by CD. But if the liquid does not freeze at B, shrinkage continues smoothly through B on into the supercooled state to E. On further rapid removal of heat, the rate of contraction begins to change and from F to G it approximates that of the crystal. In the range E to F the atomic adjustments necessary for configurational contraction occur slowly enough to be measurable, and a glass transition temperature, T_f, is defined for a particular experiment by extending LF to cut the extended line BE.

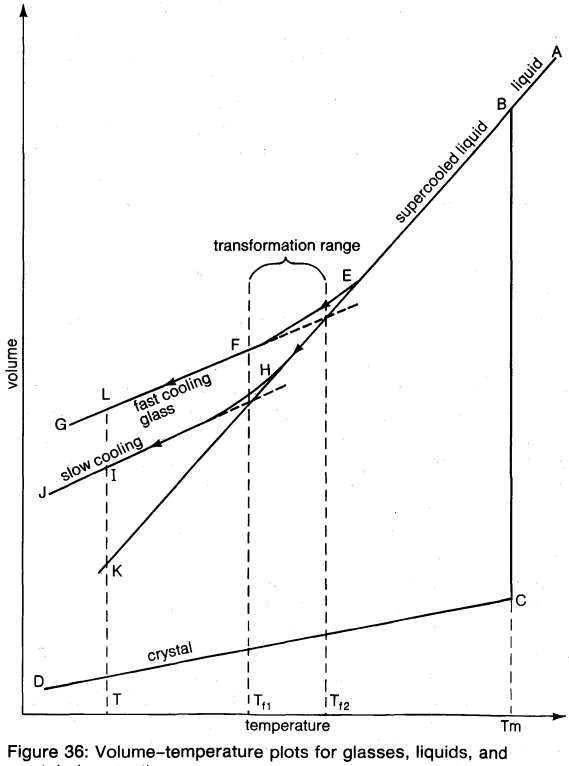


Figure 36: Volume–temperature plots for glasses, liquids, and crystals (see text).

It is more correct, however, to speak of a transformation range of temperature between the values T_{f2} and T_{f1}, which are defined by the attainable cooling rate limits. Below the transformation range the material is a glass and is in a thermodynamically unstable state. If it is cooled quickly enough to be following the line FG, and if the cooling is then stopped and the temperature held constant at T, say, the glass will slowly contract from L to K. The glass is then said to have been stabilized. The slower the rate of cooling, the further the line EK is followed, and BEHIJ is the curve obtained for a glass-forming liquid subject to a slower cooling rate than that experienced by the liquid represented by BEFG. Ordinarily, eventual departure from the equilibrium line will always occur because of the rapidly increasing times necessary for atomic rearrangement. Thus a particular glass composition has a "configurational temperature," or "fictive temperature" (T_{f1}, T_{f2}), which corresponds to the structural configuration frozen in at that temperature and persisting at lower temperatures.

Stabilized glass condition

Viscosity and temperature relationship. The variation of viscosity with temperatures indicates the ease of atomic rearrangements within the glass and is an essential property of glass that is utilized in industrial processes.

Molten glass may have a viscosity of about 10^3 poises (a poise is the unit of coefficient of viscosity and expresses the force required to maintain a unit difference in velocity between two parallel planes separated by one centimetre of fluid; one poise = one dyne-second per square centimetre) somewhat stiffer than honey on a cold day, and if cooled it gets much stiffer. At the temperatures at which molten glass can be shaped into useful articles its viscosity lies between 10^4 and 10^8 poises, well below the 10^{13} poises characteristic of glass in the transformation range. At this viscosity, the time required for the stress of distortion to be relaxed at constant strain is about one minute and the time scale for atomic rearrangements is similar. In the transformation range the physical properties are time-dependent, and below the transformation range the configuration is fixed, essentially, and the glass has the general properties of an elastic solid.

The effect of composition on glass properties. The wide range of the properties of glasses depends on their composition, and special effects result from the presence of various modifying agents in certain basic glass-forming materials (see below, *Structure of glass*).

Silica glass formers

One of the most important glass-formers is silica (SiO_2). Pure crystalline silica melts at 1,710° C (3,110° F). In pure form, silica glass exhibits such properties as low thermal expansion, high softening temperatures, and excellent chemical and electrical resistance. In pure form, it is relatively transparent over a wide range of wavelengths to visible and ultraviolet light, and to ultrasonic waves.

The high viscosity and melting temperature of silica glass are affected by the presence or absence of other materials. For example, if certain materials called fluxes are added, the most important being soda (Na_2O), both the viscosity and melting temperature can be reduced. If too much Na_2O is added, the resulting glass is readily attacked by water, but if stabilizing oxides, such as lime (CaO) and magnesia (MgO), are present in suitable amounts, the glass becomes more durable. Most commercial glass has a soda–lime–silica composition and is produced in vast

Table 1: Characteristics of Various Types of Glass

	fused silica	soda–lime–silica	borosilicate	alumino-silicate	lead
Approximate composition*	SiO_2 99.9% H_2O 0.1%	SiO_2 73% Al_2O_3 1% Na_2O 17% MgO 4% CaO 5%	SiO_2 81% Al_2O_3 2% B_2O_3 13% Na_2O 4%	SiO_2 62% Al_2O_3 17% B_2O_3 5% Na_2O 1% MgO 7% CaO 8%	SiO_2 56% Al_2O_3 2% Na_2O 4% K_2O 9% PbO 29%
Coefficient of thermal expansion (linear expansion per °C × 10^7)	5.5	93	33	42	89
Strain point† (°C) (viscosity about $10^{14.5}$ poise)	990	470	515	670	395
Annealing point‡ (°C) (viscosity about 10^{13} poise)	1,050	510	565	715	435
Softening point§‖ (°C) (viscosity about $10^{7.65}$ poise)	1,580	695	820	915	630
Young's modulus (lbs per sq in. × 10^6)	10.5	10	9.1	12.7	8.6
Refractive index (for sodium D line)	1.459	1.512	1.474	1.530	1.560
Dielectric constant (at 10^6 cycles per second and 20° C)	3.8	7.2	4.6	7.2	6.7
Density (g/cm³)	2.20	2.47	2.23	2.52	3.05

*These compositions are typical of the various glass types. †Temperature at which internal stresses are reduced significantly over a few hours. ‡Temperature at which internal stresses are reduced significantly over a few minutes. §Temperature at which glass will rapidly deform under its own weight. ‖ The strain point and annealing point roughly define the annealing range.

quantities for plate and sheet glass, for containers, and for lamp bulbs.

In soda–lime–silica glasses, if lime is replaced by lead oxide (PbO), and potash (K_2O) is used as a partial replacement for soda, lead–alkali–silicate glasses result that have lower softening points than lime glasses. The refractive indices, dispersive powers, and electrical resistance of these glasses are generally much greater than those of soda–lime–silica glasses.

Boric oxide (B_2O_3), itself a glass former, acts as a flux when present in silica and forms borosilicate glass, and the substitution of small percentages of alkali and alumina increases the chemical stability. It also exhibits low thermal expansion, high dielectric strength, and high softening temperature.

Aluminosilicate glasses find applications similar to those of borosilicates, but the former can stand higher operating temperatures; glasses with relatively high alumina contents and no boric oxide are exceptionally resistant to alkalies.

Nonsilica glass formers

The above glasses all have silica as the glass former. With other glass formers, glasses have special properties. For example, if boric oxide (the basis of Lindemann glass) is present, X-rays are transmitted and rare-earth glasses will exhibit low dispersion and a high refractive index. Phosphate glasses (used as optical glasses) based on phosphorus pentoxide (P_2O_5), are highly resistant to hydrofluoric acid, and act as efficient heat absorbers when iron oxide is added. Table 1 gives the compositions and physical properties of some typical commercial glasses of the types described.

STRUCTURE OF GLASS

Absence of long-range order. A glass is essentially a substance that has a frozen-in, liquid type of structure characteristic of that substance at a higher temperature than the actual temperature of its glassy state. Figure 37 shows, in two dimensions, the ways in which the atoms of an imaginary two-dimensional oxide (G_2O_3) would be arranged as a crystal (A) and as a glass (B). This imaginary oxide is the two-dimensional analogue of the three-dimensional tetrahedral silicon dioxide (SiO_2) in which each silicon atom is bonded to four oxygen atoms. In the crystal the structural unit is repeated regularly throughout the material and the crystal, therefore, shows long-range order. The structure of glass shows no such order. Although in both cases the number of bonds per atom is the same, in the amorphous structure of glass the bond angles are slightly distorted, and there is no long-range regularity.

X-ray evidence of structure. The broad X-ray-diffraction patterns for a glass are similar to those obtained for a liquid and they are in contrast to the crystal patterns that show sharp lines. The results are expressed in the form of radial distribution curves, two of which are shown in Figure 38. These graphs are plots of radial density; that is, they give the probability of finding a second atom at a given distance from the reference atom. For silica glass, a peak occurs at about 1.6 angstroms, which coincides with the average silicon–oxygen (Si–O) distance in crystalline silica and silicates. The area under this peak in the radial distribution curve shows that the number of oxygen atoms about each silicon atom is 4.3, which, within experimental error, is 4, and 4 is the coordination number of silicon by oxygen in crystalline silica and silicates. Other peaks and humps in the curve are thought to correspond to various characteristic interatomic distances in the silica. In the sodium–silica glass, for example, a new peak appears that probably represents the sodium-oxygen (Na–O) distance, giving an average of six oxygens near each sodium atom. Figure 38 also demonstrates an important property of glasses, namely, that they possess short-range order but no long-range order. The sharpness of the peaks is a measure of the amount of disorder present, and the peaks broaden to give a fairly smooth curve after a few interatomic distances confirming the lack of long-range order.

Functions of network modifiers and glass formers. Network modifiers, in the form of large metallic cations (pos-

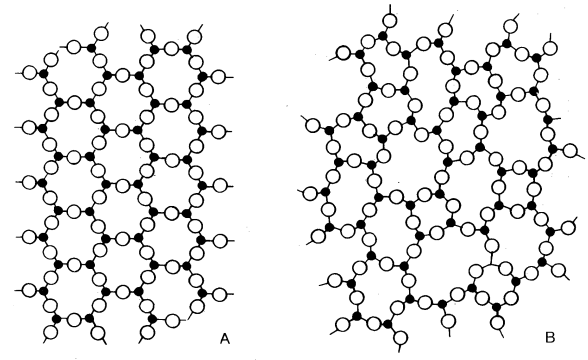


Figure 37: Structure of an imaginary oxide, G_2O_3, (A) as a crystal and (B) as a glass.

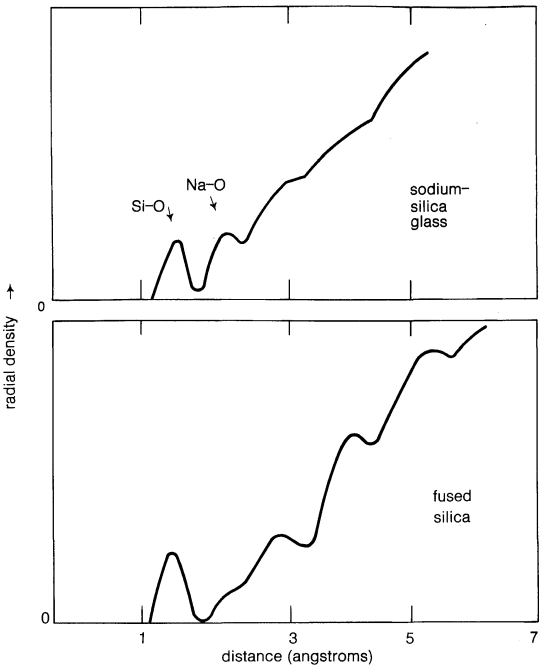


Figure 38: Radial distribution curves for a sodium-silica glass and for pure fused silica (see text).
Adapted from B.E. Warren and J. Biscoe in G.O. Jones, *Glass;* Chapman & Hall

itively charged atoms), weaken the structure. Although it occupies a space in the open structure, the modifier removes one of the strong directional oxygen–silicon bonds from the basic network and replaces it with a weaker nondirectional bond, thereby lowering the viscosity of the glass. Alkali and alkaline-earth ions act as network modifiers.

Examples of elements other than silicon whose oxides, sulfides, selenides, and tellurides form strong directional bonds and act as glass formers are boron, phosphorus, germanium, vanadium, and arsenic. Network modifiers, with their characteristically low bond strengths and property of lowering the viscosity, include sodium oxide and calcium oxide. There is also a class of substances whose oxides are known as intermediates, including titanium, aluminum, thorium, and beryllium. Alone, intermediates will not form glasses; however, they do so in combination with glass formers. Intermediates can enter the network if a suitable modifier oxide is available; if not, they act, themselves, as modifiers.

Spectroscopy of glasses. X-ray-diffraction studies of glasses give fairly reliable information about the number of oxygen atoms surrounding the small, highly charged, glass-forming cations, but far less about the positions of the modifying alkali ions. Spectroscopic techniques yield further useful information about the glass structure.

In spectroscopy, differences between allowed energies or quantum states of the system are measured, and these results are compared with theoretically calculated quantum states of a particular arrangement of atoms.

In the infrared region of the spectrum, the allowed energies, or quantum states, are associated with the vibrational motion of atoms but little theoretical progress toward understanding can be made because, in solids and liquids, the vibrations of all the atoms together must be considered. Spectral absorption peaks can, however, often be matched with observed peaks for crystals of known structure. Correspondence is taken to indicate the presence of particular groups of atoms, or "complexes."

In the ultraviolet region, the quantum states are those associated with the movement of an electron from a negative oxygen ion to a positive ion, or cation, but a large contribution is made by energy of vibrational modes (called phonons), which interact with the electronic energies to give intense broad absorption bands.

In the visible region, absorption bands arise from transition metal and rare-earth ions in solution in the glass. In this case the quantum states are only slightly modified by the nearest neighbours.

Deductions from ligand field theory

These nearest neighbours, which surround the central glass-forming ion, are known as ligands. Their modifications of the quantum states, and the symmetry of their arrangement around the central ion can be deduced from ligand field theory. This theory enables the relative energies of the absorption peaks to be derived and also allows the absorption intensities to be classified in groups according to the symmetry of arrangement of the nearest neighbours. It has been of great use in recent years in the study of coloured glasses.

Solution and colloidal colours in glass. Normally, in the visible region, only ions that give rise to colour need be considered when studying the transmission properties of glass because the characteristic absorptions due to vibration of the silicon–oxygen bond lie outside this region. The most common colouring ions are those of the transition metals cobalt (Co), chromium (Cr), copper (Cu), manganese (Mn), and iron (Fe). The colours are affected by three major factors, namely, the nature and concentration of the colouring ion, the composition of the base glass, and the oxidation state of the colouring ion.

Small quantities of the colouring ion, about one-tenth of a percent, are sufficient to colour glass. For this reason, the yellow-green colour caused by iron is a serious problem in the glass industry. This is alleviated by the choice of pure sands and by the addition of decolorizers. Ferrous iron absorbs energy in the near infrared, and its broad absorption band spreads into the visible spectrum causing a blue colour. The decolorizer contains an oxidizing agent, which converts iron cations with 2 charges (Fe^{2+}) to cations with 3 charges (Fe^{3+}), and also a little selenium and cobalt. Addition of the latter elements, together with the ferric iron, Fe^{3+}, give a residual absorption that is essentially neutral. The manganese dioxide, MnO_2, which was formerly used, oxidized the iron and gave a complementary purple colour, but was less stable.

Colour in glass can also be caused by colloidal precipitation of very small metallic crystals, such as those of copper or gold, to produce a "ruby" colour. The particles develop when a glass of suitable composition, previously melted under certain reducing conditions, is heat treated. The particle size is critical, between 10 and 1,000 angstroms, and selective absorption then gives the glass its red colour.

Nucleation and crystallization in glass-forming liquids. It was noted above that a supercooled glass-forming liquid is in a metastable state (a stability that is maintained only by a slight margin over an unstable state) and that if crystals or crystal nuclei are introduced and freezing can begin, the crystallization will spread throughout the liquid.

Two aspects of the crystallization process must be considered separately; namely, the rates of formation of nuclei and their growth. So many nuclei are present in normal freezing processes that the crystallization process depends essentially upon the growth rate, and as this is negative above the freezing point, and very large and positive just below it, the liquid freezes at, or very near, the expected freezing point. For liquids (typically the highly viscous glass formers) that can be greatly supercooled, however, the crystallization rate also depends critically on the rate of formation of nuclei. Nuclei can be introduced from outside (heterogenous nucleation) or can appear spontaneously within the supercooled liquid (homogeneous nucleation). Homogeneous nucleation occurs because of thermal fluctuations within the liquid; consideration of the contribution of the surface energy of the embryonic nuclei to their total free energy shows that, at any given temperature, there is a critical size below which the nuclei are unstable and redissolve, and above which they can grow.

Homogeneous nucleation

When nucleation and crystal growth occurs in glasses, it is known as devitrification. It is most marked in glasses in which the rates of nucleation and crystal growth nearly coincide when plotted as functions of temperature. Closeness of these curves results in nucleation and crystal growth over a wide temperature range. More stable glasses have a high viscosity at the freezing point, and crystal growth, which proceeds by diffusion process at the surface of the crystal, is slowed down. Heterogeneous nucleation,

however, can still occur, especially at glass surfaces, when impurity particles are present.

Studies of nucleation and crystallization in glasses have been made in recent years. One result of these efforts was the discovery of a glass containing a small amount of dissolved metal, such as gold or silver, which, after exposure to ultraviolet radiation and subsequent heat treatment, develops a colour caused by the growth of small metal particles in the glass. If this photosensitive glass is heat treated further, the particles act as nuclei, and opalescent regions of the glass devitrification products are formed wherever the glass was exposed to the ultraviolet radiation. Much less resistant to chemical attack than the parent glass, these regions are dissolved away while the parent glass remains. By the commercial use of this process, pieces and patterns of desired shapes can be obtained—for instance, fluid amplification elements.

Glass ceramics

A glass has been produced that can be changed almost completely into crystalline material—the glass ceramic. In application, an article of suitable glass composition fabricated by the usual method is subjected to appropriate heat treatment to convert it into a ceramic. Such materials are extremely inert chemically and can be designed to have either zero or a very small expansion coefficient.

SPECIFIC PROPERTIES OF GLASS

Mechanical properties. Glass rods with untouched surfaces have a tensile strength of 34,500 kilograms per square centimetre (500,000 pounds per square inch). By considering the energy changes involved in the surfaces formed by fracture the theoretical strength of glass is estimated to be about 69,000 kg/cm^2 (1,000,000 lb/in^2). Glass articles normally show strengths that are much lower, about 2,100 kg/cm^2 (30,000 lb/in^2); and this difference is due to submicroscopic surface flaws that, produced by the slightest contact with another body, act as stress magnifiers. Glass in ordinary use, therefore, is far less strong than it would be in a flawless state. Although the flaws have the same stress-reducing effect as dislocations in crystals, the exact identity of the flaws remains uncertain. Their presence, however, and number can be revealed by treating the glass with some modifying agent; for example, by immersing glass in molten lithium nitrate, the sodium in the surface is exchanged for lithium. The foreign layer that is produced is in tension when the glass cools and minute cracks develop at the flaws.

At temperatures below the transformation range, glass behaves as a solid and stresses will be present if the temperature varies within a glass object; a piece of hot glass cooled suddenly will contract more on the outside than on the inside; the further contraction of the outer layers is thus prevented and they are put into tension; the higher the thermal expansion coefficient of a glass is (the more it expands for each degree rise), the greater this tension will be. Some glasses expand very little. Thus, fused silica has excellent resistance to heat shock, while borosilicate glass can undergo repeated reheating below the transformation range without cracking. It makes excellent ovenware.

In the transformation range the glass behaves instantaneously as a solid, an immediate elastic strain being produced when an external force is applied, but this strain increases if the force is not removed. If the force is continually reduced to keep the strain constant, the internal stress gradually disappears. The rate of this change depends on the viscosity of the glass. When the coefficient of viscosity, η, equals 10^{13} poise, the stress is halved in about one minute. If a temperature gradient (a variation of temperature) exists in a piece of glass in the transformation range, the resulting stresses die away and the glass then has a temperature gradient but is stress free. On cooling to room temperature the hotter interior has to contract more than the outer layers and is, therefore, in tension while the outer layers are in compression. These stresses remain until the glass is reheated to the transformation or annealing range. The principle is applied in annealing, which is a process of minimizing the permanent stresses in glass by heating until they disappear and then cooling slowly so that only very small temperature gradients, and ultimately, therefore, only very small stresses, appear.

Annealing

Unless a glass article is properly annealed it might break spontaneously after cooling. Large volumes of stress-free uncracked glass are rare and difficult to achieve. Annealing for exacting uses, such as in some optical applications, moreover, involves more than the removal of thermal stresses; the atomic structure of the glass will vary within the piece unless all parts of it have had exactly the same thermal history through the transformation range.

Chemical properties. Common glasses are chemically inert but reactions with many substances do occur slowly. In contact with aqueous media the reaction begins with an exchange of hydronium ions, H_3O^+, for sodium ions, Na^+. (Hydronium ions exist in equilibrium with hydroxyl ions, OH^-, in water.) Immediately, a surface layer starts to build up in which the sodium ions in the glass have been replaced by hydronium ions. For the reaction to continue, sodium ions must diffuse through this surface layer. While the silica itself dissolves slowly, eventually the rate of diffusion of sodium ions and the rate of removal of silica may be such that the surface layer, although moving into the glass, will remain at a constant thickness.

With commercial glasses these conditions are hardly attainable. The complexity of the process makes it very difficult to design accelerated test procedures. Components such as alumina may be added to increase the chemical durability of the glass. The action of alumina is to provide sites that hold sodium ions on the surface, thus producing a more uniform layer of sodium ions and lowering their rate of diffusion. The occupancy of atomic sites on the glass surface determines the behaviour of glass when it is used as an electrode in determining hydrogen ion (H^+) concentrations. When measuring hydrogen ion concentrations, it is important that the surface sites on the electrode be occupied by protons. Normally, with glass electrodes, as the hydrogen ion concentration decreases, and the pH value rises (pH is a logarithmic expression of hydrogen ion concentration) to a sufficiently high pH value, the electrode emf (electromotive force or voltage) is no longer proportional to the hydrogen ion concentration; this produces an "alkali error" because the sites have become occupied by sodium ions instead of by protons. If a glass contains a high proportion of alumina, the change point in the measurement can occur at quite low pH values so that over a large range of pH the electrode can be used to determine sodium ion concentrations. There is a close correlation between the electrode behaviour and variation of durability with the pH of the attacking solution.

The effect of pH

Optical properties. Optical glass must be homogeneous. Inhomogeneities are of two general types, chemical and physical. Chemical inhomogeneities include gas bubbles and blisters, unmelted batch materials, and devitrification products, while physical inhomogeneity can result from the presence of strain in the glass caused by faulty annealing. Chemical and physical inhomogeneities may result in glass that differs in composition, and therefore in refractive index (the measure of the change in direction of light as it passes from one medium into another) from place to place in the glass.

Chemical and physical inhomogeneities

Batch materials for optical glass manufacture must be of exceptionally high purity in order to give a highly transparent glass free from colour. In order to ensure chemical homogeneity the molten glass is stirred to mix it thoroughly. Physical inhomogeneity is minimized by extremely slow annealing (see above *Mechanical properties*).

The optical characteristics of any glass can be completely described by giving the two values n and ν, n being the refractive index and ν the reciprocal of the dispersive power, a property related to the angular separation of two rays of light of different colours, passing through glass, and called the constringence. Thus any glass can be represented by a single point on an n-ν plot, as shown in Figure 39.

The tiny dark-shaded area (Figure 39) represents the entire range of optical glasses available in 1880. All earlier optical glasses were either soda–lime–silica crown glasses or potash–lead–silica flints, with a very limited range of optical properties and a dispersive power that increased with the refractive index. By controlled scientific experiments it was possible to add to these two glass types a range of optical glasses such as borosilicate crowns,

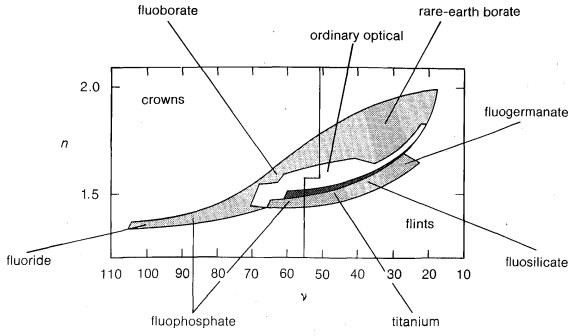


Figure 39: Refractive index, *n*, versus constringence, *v*, for the range of optical glasses (see text).
Glass in the Modern World by F.J. Terence Maloney, 1968, Doubleday & Company, Inc.

barium crowns, barium flints, and borate and phosphate glasses. From these, pairs of crown and flint glasses could be used in the manufacture of lenses that gave images free of colour fringes. The new glasses also had a greater diversity in both refractive index and dispersive power, and much more versatile optical systems became possible. The unshaded area of Figure 39 shows how far the range had been extended by 1934.

The remaining shaded area in Figure 39 represents glasses developed since 1934, owing largely to the introduction of rare-earth and fluoride types. New glasses of high dispersion combined with low refractive index, and of high refractive index combined with low dispersion, were especially sought. It was shown that optical glasses could be made by replacing silica with boric oxide, with additions of rare earths such as lanthana and thoria, and other oxides such as titania, and tantalum and tungsten oxides. These produced optical glasses with very high refractive indices but low dispersions.

The range of available dispersions and refractive indices was further extended by the use of fluosilicate flints, or "super flints"—glasses with dispersions greater than conventional flint glasses of the same refractive index. Figure 39 demonstrates the importance of glasses such as the fluoborates, phosphates and germanates, and even types of all-fluoride glasses. These glasses, with their very low dispersions and refractive indices, have an extended range of radiation transmission in both the infrared and ultraviolet portion of the electromagnetic spectrum.

Thus modern optical designers have available to them a very wide range of glasses with independently varying dispersions and refractive indices. This enables them to design versatile optical systems that actually require fewer glasses with slightly differing compositions than previously; digital computers aid in the work.

Electrical properties. All glasses containing alkali ions are electrolytic conductors and show great changes in resistivity with temperature. At room temperature, in addition to volume conductivity through the whole body of the glass, there is a conductivity restricted to the surface, caused by a layer of adsorbed moisture that decreases the surface resistivity. In a very humid atmosphere the surface conductivity of high-alkali glasses can be many times the volume conductivity. The effect is reduced by silicone surface treatment and disappears when the glass is tested in a vacuum.

In volume conductivity nearly all the current is carried by monovalent alkali cations. For this reason the resistivity is determined mainly by the amount of alkali present and, to a lesser degree, by the network structure and thermal history of the glass. At room temperatures the volume resistivity can be as low as 10^6 ohm centimetres for certain high-alkali glasses; for glasses in which trace alkali conductivity is blocked by the addition of divalent ions, it can be as high as 10^{30} ohm centimetres.

Semiconductors

A new class of semiconducting glasses has been discovered that shows a decrease in resistivity with increasing temperature, but no polarization. Conduction occurs by electrons or holes (fictitious electron vacancies for which the electrical charge is positive) or both, and such glasses can have a resistivity as low as 10^4 ohm centimetres at room temperature. Typical glasses of this type contain arsenic, selenium, tellurium, and germanium. Glasses that contain large amounts of transition metal ions (for example, vanadium oxide) also conduct by the passage of electrons and show the same low resistivities.

Amorphous semiconducting materials, in particular some chalcogenide glasses, have been used in electronic switches. An advantage claimed for amorphous switching devices, over conventional crystalline semiconductor switch types, is that they are basically less susceptible, or more insensitive to, the degrading caused by neutron or X-ray irradiation in "hot" environments or to the radiation of outer space. (R.W.D./Ed.)

Mechanical properties of solids

ELASTIC AND PLASTIC PROPERTIES

Many of the uses of solids depend on their ability to support stresses (applied load divided by the area of application); indeed, the ability of a solid to support normal loads (loads applied perpendicular to a surface) and shear loads (loads applied parallel to the surface) distinguishes a solid from a gas or liquid. The deformation (change of shape of the solid) that results from the application of a stress may be divided into three general types: elastic, plastic, and anelastic. These may be best defined with reference to the dependence on the rate at which the stress is applied and the permanence of the deformation when the stress is removed.

Types of deformation

Elastic deformation is an instantaneous shape change in response to an applied stress and is completely and immediately recovered when the stress is removed; *i.e.*, there is no permanent "set." In contrast to this, plastic deformation results in a permanent shape change that depends on the rate at which a stress is applied and that is not recovered on removal of the stress. Anelastic deformation, occupying a middle ground between these two extremes, is a recoverable deformation that depends on the rate of application of the stress and on the time during which it remains applied. These three types of deformation have been commonly experienced in the stretching of a steel spring or of a rubber band (elastic deformation), the bending of a paper clip (plastic deformation), and the decreasing amplitude of vibration of a plucked violin string (caused by anelastic deformation). Unfortunately for the investigators of these phenomena, it is not easy to separate these three classical types of behaviour, and the overall result of the application of a stress is a combination of all these types.

THERMAL PROPERTIES

The excitation of lattice vibrations and electrons on increasing the temperature of a solid has a major effect on many properties of solids. Attempts at understanding these thermal effects have played a central role in the development of solid-state physics. The early quantitative experiments on the specific heat of solids in the 18th century were interpreted as the introduction of heat as a "caloric fluid" that had properties such as weightlessness and transparency and that seeped from hot to cold. It was not until 19th-century work on the behaviour of gases that heat was interpreted as the motion of molecules and that the kinetic molecular theory stating this in specific terms became accepted. The application of principles that had been developed for gases to an understanding of the thermal properties of solids was not possible in the context of classical physics, and the attempt provided some motivation for the development of quantum mechanics early in the 1900s.

Early theory of heat

The two major thermal phenomena in solids are the specific heat (or heat capacity: the quantity of heat that will raise the temperature of a unit mass 1°) and the conduction of heat. Based on the understanding of gases, it was realized that each atom stored energy in its vibrational modes (frequency and amplitude) as kinetic energy (energy of motion) of the atom and as potential energy of the distorted bonds that join the atom to its nearest neighbours. The amount of energy per mode in the temperature range where the solid behaved classically—*i.e.*,

where the atomic motion can be described without the use of quantum mechanics—was one-half of the product of Boltzmann's constant, k (3.3×10^{-24} calorie per degree Kelvin) and the absolute temperature, T (degrees Kelvin): $\frac{1}{2}\ kT$. Since the vibration of each atom in a solid can be described as motion in three independent directions (*i.e.,* three kinetic and three potential energy modes), the classical energy of a mole of atoms in a solid is three times the product of Boltzmann's constant, k, the temperature, T, and Avogadro's number, N, 6×10^{23} (the number of atoms in the atomic weight in grams of an element): $3\ NkT$. The heat capacity, the rate of change of energy with temperature, is $3Nk$ (or about six calories per mole). In addition to energy stored as lattice vibrations (phonons), excitation of electrons to higher energy states may also occur. In metals this electronic specific heat increases linearly with temperatures and is small compared to the lattice vibration term.

Electron and phonon transport

Conduction of heat along a temperature gradient in a solid can involve the transport of energy by electrons, by phonons, or by both. Transport of heat by electrons accounts for the majority of the thermal conductivity in metals at room temperature and above. The thermal currents due to electron and phonon transport, however, are additive and at low temperatures the heat conductivity is due solely to phonons. Motion of phonons and electrons through the lattice is limited by scattering from crystalline defects such as impurities, grain boundaries, external surfaces, dislocations, etc. and from other phonons and electrons. These scattering processes determine the thermal conductivity of the solid. The presence of silver impurity ions in an ionic crystal such as potassium chloride, for example, causes a drop in conductivity. In the scattering process involved here, a silver ion momentarily traps the phonon, which leads to a decrease in the thermal conductivity.

Electrical and magnetic characteristics

Variation in metal conductivity

Materials display a wide range of electrical behaviour, typified by the electrical conductity of such elements as aluminum, silicon, and phosphorus, which are neighbours in the periodic table of elements. Wires made of each of these materials about the size and length of wires in a lamp cord and connected across the terminals of a 12-volt auto battery would conduct greatly different currents. Through the aluminum wire a current of about 100 amperes would flow, through the silicon wire a current many times smaller would flow, about 10^{-6} ampere, and through the phosphorus wire a current of only about 10^{-15} ampere would flow.

This wide range of electrical behaviour among aluminum, a useful metal, silicon, a widely used semiconductor, and phosphorus, a good insulator, typifies the enormous variety of electrical applications of materials now possible. Why should the adjacent elements, aluminum with 13 electrons, valence 3, silicon with 14 electrons, valence 4, and phosphorus with 15 electrons, valence 5, behave so differently? The answer requires understanding of the basic characteristics of electrons in solids.

The prevailing conception of the electronic structure of matter was developed rather recently. The discovery of the electron and its properties in the early part of the 20th century and the development of quantum mechanics during the mid-1920s led to formulation of a detailed theory of the electronic structure of solids in the 1930s. The theory is rather good, overall, even though certain fine details are not yet clear. This deficiency has not prevented the technical application of knowledge of the electronic structure of solids; indeed, the rapid development of solid-state devices has its basis in the close interplay among the developing theory, basic experiment, and practical applications.

CONDUCTORS

Wide variability of electrical behaviour is not limited to the three examples just described; rather, this variability is common as is illustrated for a number of materials in Figure 40. The group of materials at the bottom, the metals,

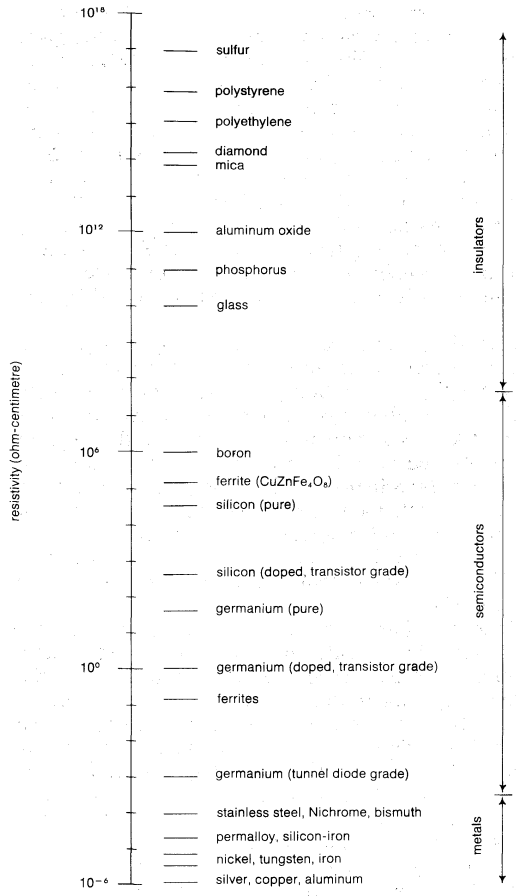


Figure 40: The electrical resistivity of various materials at room temperature. The unit, ohm-centimetre, is the resistivity across a one-centimetre cube of the material.

are good conductors. The individual atoms in the metals are assumed to have lost electrons (the valence electrons) to the entire solid. These free electrons move as a group in an electric field and hence conduct current. Indeed, early in the 20th century a quantitative theory of conductivity was developed to explain electrical conductivity in terms of the number of free electrons, their charge, and their degree of mobility. Later refinements were added by application of quantum mechanics, but the basic principles still apply. (C.A.W./H.K.B.)

Free-electron model. In the simple (that is, nontransition) metals there is a clear distinction between the conduction electrons and the tightly bound core electrons closer to the nucleus. The number of conduction electrons per atom, called the metallic valence, is equal in these metals to the number of the column in the periodic table; for example, I for sodium, II for magnesium, III for aluminum, and so on (see Figure 41).

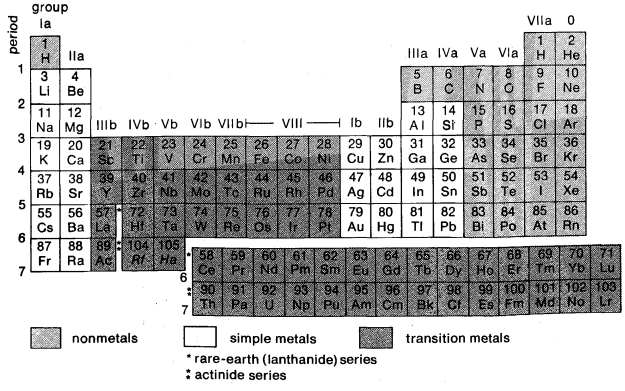


Figure 41: Periodic Table of the elements showing simple metals, transition metals, and nonmetals.

Quantum theory, which describes the emission or absorption of energy by atoms or molecules as taking place by steps, each step being the emission or absorption of an amount of energy called a quantum, must be used to describe the states of these conduction electrons in the metal. Free electrons may travel through the entire metal, but because they are waves, they may move only with a specific set of velocities. The electron wavelength must be such that a whole number of wavelengths (or half wavelengths for some methods of treating the crystal boundaries) equals the length of the crystal. It is more convenient to specify a particular electronic state by its wave number, which is 2π (2×3.1416) divided by the wavelength.

Wave number

Although the electron can move in any direction in the crystal, which is visualized in the shape of a cube, it is evident that the motion of the electron is similarly restricted in the other two dimensions. Thus, the wave number represents a value in the direction of the electron motion, restricted in each of the three dimensions to one of the discrete (specific) velocities. The possible electronic states in a crystal, then, are represented by a lattice of points in wave-number space, each state corresponding to a wave number k, which terminates on one of these points. The larger the crystal, the more closely spaced are the possible states.

Pauli principle

The Pauli principle states that not more than two electrons (spinning in opposite directions) can have the same wave number. Since the smaller the wave number, the lower the energy, the most stable arrangement of electrons is obtained by letting two electrons have each of the wave numbers in a region in wave-number space bounded by a sphere, called the Fermi sphere, large enough to encompass all the conduction electrons (see Figure 42). The states outside the Fermi sphere have higher energy than those inside and are not occupied. In simple metals there are so many conduction electrons that the states with wave numbers near the Fermi sphere are very energetic, the electrons having velocities corresponding to 1 percent of the speed of light.

Conduction of electric current. In the ground state of a metal, the most stable state, there are as many electrons going one way as another, and no net current flows. If an electric field is applied, however, each electron is accelerated, and correspondingly its wave number shifts in the direction of the force. This results in a displacement of the Fermi sphere and in an electronic current.

If the electrons were truly free, they would accelerate indefinitely in the presence of an electric field, and the current would continue to grow. Impurities, defects, and thermal vibrations of the lattice, however, scatter the electrons and keep the current at a constant value. An electron of mass m, on the average, will move for a time symbolized by the Greek letter tau, τ, the scattering time, before it is deflected. During that time it will be propelled by a force equal to its own charge (e) multiplied by the electric field symbolized by E (*i.e.*, as eE) and will increase its velocity by an amount $eE\tau/m$. The electron contributes to the current its charge times the added velocity, and the current density, if N represents the number of electrons in unit volume, is $Ne^2E\tau/m$. Thus, the current density is proportional to the electrical field (Ohm's law); the proportionality constant, symbolized by the Greek letter sigma, σ (with $\sigma = Ne^2\tau/m$), is the conductivity. It is large in metals because of the high density of conduction electrons, particularly large if the number of defects is small and the temperature is low, conditions that lead to a long scattering time, τ.

Conduction of heat. The theory of thermal conductivity is quite similar. The large number of essentially free electrons, which transport thermal energy from hotter to cooler sections of the material, accounts for the high thermal conductivity of metals. Difference in temperature in two parts of the metal produces the flow of thermal energy, and the proportionality constant is the thermal conductivity. The ratio of the thermal to the electrical conductivity is equal to a constant (the Wiedemann–Franz ratio) times the temperature. There are also thermoelectric effects, involving both electric fields and temperature differences, which are understood on the same basis, and are brought to use in the thermocouple, which measures temperatures electrically.

Although the Pauli principle has little effect on conduction properties, it becomes quite important in the heat capacity. In the classical physics preceding quantum theory, electrons were expected, because of their large numbers, to make a major contribution to the heat capacity of a metal. In the quantum mechanical description, however, electrons lying well within the Fermi sphere cannot take on this thermal energy, because all the electron states in this energy range are already occupied. The contribution of the electrons to the heat capacity is reduced to the point that the ratio is of the order of 1:400 in most metals at room temperature. There is a similar suppression of the response of a metal to a magnetic field. Each electron behaves as a tiny magnet by virtue of its spin, and in the presence of a magnetic field, it tends to align with that field, leading to a magnetization within the metal. Because only those electrons near the Fermi sphere may reverse their spins, however, the resulting contribution, known as Pauli magnetic susceptibility, is much smaller than would be expected according to classical theory.

In the above description the electrons are free to move within the metal but not to leave the metal. The forces that hold the electron in the metal are the same electrostatic forces that hold the metal itself together. They were first calculated for the alkali metals; then, comparison with the corresponding binding of electrons to the isolated atom led to determination of the cohesive energy for the metal. Cohesion in metals derives from closer proximity of the free negative electrons to the positive metallic ions in the metal than exists for electrons bound to isolated atoms.

Pseudopotentials. In the free-electron representation, interaction between free electrons and the lattice of ions was neglected except in the general sense that applies to cohesion. While this interaction is known to be strong, the success of the foregoing assumptions seems to indicate that the effect of the interaction is insignificant. This paradox is resolved through the concept of a pseudopotential. Conduction electrons, having high energy, move past the ions with very little deflection, as a speeding car

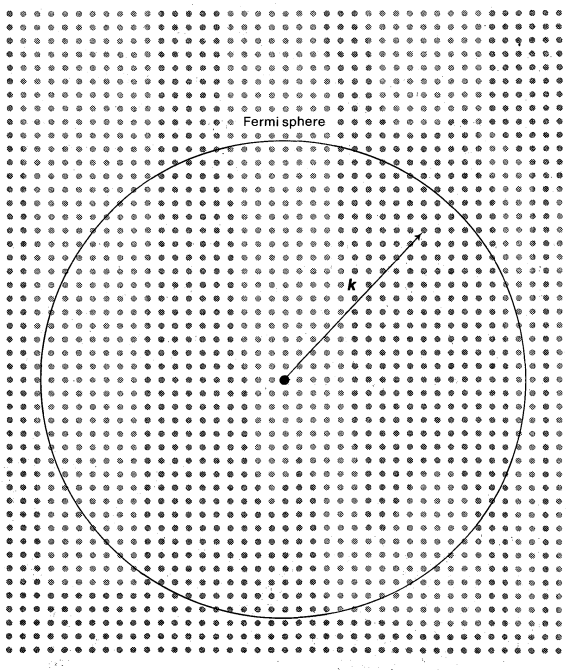


Figure 42: The Fermi surface, which separates conduction electrons from nonconducting space. Each dot represents an electron state and, if it is occupied by an electron, it represents the electron's velocity and a wave number k. The wave number may be represented by an arrow from the centre to the dot. Because no electron has an energy high enough—i.e., a wave number high enough—to place it outside the surface of the sphere, no arrow can be drawn to any point outside the surface (see text).

rides relatively smoothly over a bumpy road and a slower car rises and falls with every bump and depression. This effect may be incorporated in the quantum theory of the electron motion by mathematically considering only conduction electron states and deliberately excluding the core electrons bound to the nuclei.

Diffraction by pseudopotential

Although the interaction between conduction electrons and the lattice is then represented by a weak pseudopotential, there are inevitable effects of that interaction. In the same manner that an X-ray moving through a crystal may be diffracted, the electron wave may be diffracted by the pseudopotential. Either diffraction can occur, however, only if the wave in question satisfies the so-called Bragg condition; *i.e.,* that the wave number terminate on a Bragg reflection plane in wave-number space. These planes contain wave numbers with magnitudes of the order of 2π (2×3.1416) divided by atomic spacing. For electromagnetic waves, this condition is satisfied by X-rays.

Bragg planes and Fermi spheres. Because of the high density of electrons in metals, the wavelengths of free electrons in the Fermi sphere are also comparable to spacing between atoms, so that it frequently happens that Bragg reflection planes intersect the Fermi sphere in wave-number space. Only a small number of states, however, lie exactly at the plane and can be diffracted. The others traverse the crystal without diffraction and behave as if completely free.

The Bragg condition presumes perfect arrangement of atoms in the periodic lattice. When there are defects or thermal distortions, even electrons not at Bragg planes may be diffracted. Such diffraction caused by defects is called electron scattering and is the process that limits the conductivity. The strength of this scattering may be directly calculated from a knowledge of the pseudopotential associated with each ion and of the arrangement of ions in the defective crystal. Electrical resistivity, which is the inverse of the conductivity, tends to be proportional to the number of defects. Usually the change in resistivity of a metal is about one micro-ohm-centimetre per atomic percent of defects (one ohm is the resistance of a circuit in which a potential difference of one volt generates a current of one ampere).

When a metal crystal is melted, the distinction between diffraction and scattering of electrons becomes unclear. It is nevertheless possible, again in terms of pseudopotentials, to calculate the rate of transitions and therefore the conductivity of the liquid metal.

Although only a tiny fraction of the electrons in a perfect metallic crystal are diffracted, the effects of the diffraction may be seen by the application of a magnetic field. An electron moving in a magnetic field reacts to a force perpendicular to its motion (this is exactly the force that drives the armature in an electric motor). An electron with its wave number at the Fermi surface will change its direction of motion by moving along the Fermi sphere. In the absence of diffraction it moves around the Fermi sphere, returning periodically to its initial wave number with a frequency called the cyclotron frequency. The motion on the Fermi sphere would be along a circle made by the intersection of the Fermi sphere and a plane perpendicular to the magnetic field.

Such motion of electron states across the sphere brings many to the Bragg plane, where they can be diffracted. In doing so, they jump to the Fermi sphere at the opposite Bragg plane and continue on their way. The wave number returns to its initial value after traversing a shorter set of arcs than would be the case if there were no diffraction planes; therefore, its cyclotron frequency is correspondingly higher.

It is not difficult to visualize the real orbit of the diffracted electron within the crystal in a magnetic field. The diffraction causes its velocity to change abruptly, but it cannot jump instantaneously from one place to another in the crystal. Thus, its orbit is made up of a series of arcs, always combined into a large orbit on an atomic scale, extending over many thousands of atom distances.

Reduced-zone representation. Studies of the electronic properties of metals in which such orbit shapes are of direct interest can be more easily done by means of a different representation of the electron states, the reduced-zone representation.

The region so defined, called the Brillouin (or reduced) zone, is bounded by Bragg reflection planes and can be constructed for three-dimensional lattices. The Fermi surface is constructed in the reduced zone by translating each segment of the Fermi sphere so that every pair of states between which diffraction can occur appears at the same wave number. Such translations of the arcs in which diffracted electrons move lead to a segment of the Fermi surface in the reduced zone that is precisely the shape of the electron orbit in real space.

Such a three-dimensional construction of the Fermi surface of aluminum is shown in Figure 43. All possible orbital paths in a magnetic field may be obtained by constructing intersections of planes perpendicular to the magnetic field with these surfaces. Thus, they provide a tremendous amount of information at a glance.

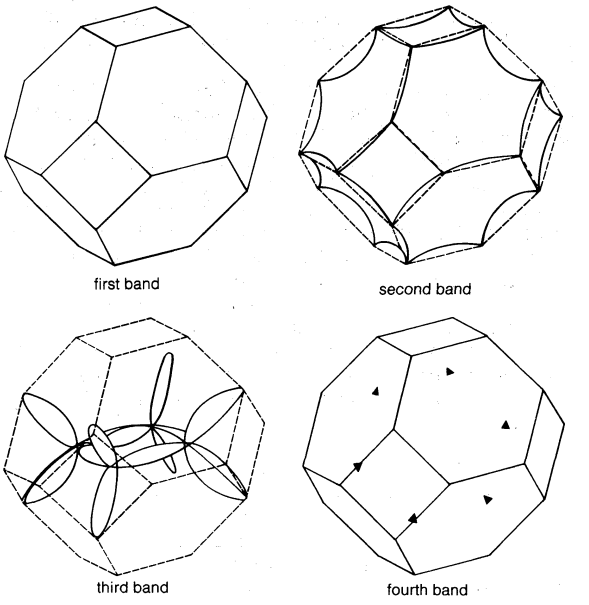


Figure 43: The Fermi surface of aluminum.

A wide range of properties is affected by the existence of diffraction and the corresponding changes in the Fermi surface. One of the simplest is the Hall effect, which arises when a magnetic field is applied perpendicularly to the direction of a current flow: the electrons are deflected and accumulate at the surface, causing an electric field, the Hall field, just large enough to prevent further accumulation. Thus, the Hall field is a transverse field proportional to both the current and to the applied magnetic field. The proportionality proves to depend only on the density of electrons and their charge if there is no diffraction.

The Hall effect

The presence of diffraction strongly modifies the Hall field, and thus the Hall effect may be used as an experimental tool to study the Fermi surface. The magnetic field also reduces the current along the applied electric field, an effect called magnetoresistance, which also gives information about the Fermi surface.

Certain other effects give more direct information about the orbits in a magnetic field. One, called cyclotron resonance, is a measurement of the cyclotron frequencies of various orbits in the metal. Another, the magnetoacoustic effect, uses sound waves to measure the diameter of electron orbits and therefore of the Fermi surface. The most precise measurements of the Fermi surface have been made with the De Haas–Van Alphen effect, in which fluctuations in magnetic susceptibility are measured by increasing or decreasing a magnetic field. These fluctuations depend only upon the cross-sectional areas of the Fermi surface and give precise measurements of the Fermi surface areas.

Studies of the Fermi surfaces using these techniques have confirmed the general features of this diffraction theory of the Fermi surface. Discrepancies found in calculations of

areas and dimensions are believed to be due to the fact that the pseudopotential is not arbitrarily weak. Because of the pseudopotential influence, an electron begins to diffract before it quite reaches the Bragg reflection plane; the true orbits do not have sharp cusps of the type shown; rather the cusps are rounded off and the Fermi surfaces are correspondingly rounded off by the effects of the finite pseudopotential. The pseudopotentials themselves and their resulting effects may be calculated and again tested by experiment.

Band structure and properties of metal groups. In the monovalent metals, those in Groups Ia and Ib of the periodic table (see Figure 41), the free-electron Fermi sphere lies entirely in the reduced zone and no diffraction is expected. The electrons in the Ia metals are expected and found to be like free electrons. The pseudopotential, however, is so strong in the noble metals, copper, silver, and gold, that the Fermi surface bows out toward the zone faces.

The polyvalent metals (those of Groups IIa, IIIa, IVa, and IIb) have complicated Fermi surfaces, such as the aluminum surface shown in Figure 43. All have been thoroughly studied, and the reduced-zone representation has proven a good starting point in all cases.

Electronic structures of the transition metals are more complicated. In addition to the conduction electrons there are electrons, called *d* electrons as in the atom, which have energies near the Fermi energy but which retain much of their nucleus-bound character rather than becoming free. It is possible to describe the conduction electrons and *d* electrons together in terms of very complicated Fermi surfaces, but for many purposes it is preferable to separate the *d* electrons conceptually. This separation is particularly appropriate in the rare earths where the states in question are *f* states rather than *d* states.

Atomic *d* states

In the noble metals, *d* electrons lie below the Fermi energy and therefore have no Fermi surface. Conduction-electron states, however, and *d* states in the noble metals interact to produce a large pseudopotential and the distortion of the Fermi surface mentioned above. The presence of the *d* states also leads to a contraction of the lattice in the noble metals in comparison to the other monovalent metals and the corresponding high density of electrons makes the conductivity high.

Transition metals have varying numbers of *d* electrons. Many properties depend principally upon the electron-to-atom ratio, the continuation of metallic valence through the transition series indicated at the top of the periodic table. The properties of transition metal alloys depend upon the average electron-to-atom ratio for the alloy system. In all cases the *d* states give a large concentration of states near the Fermi energy and therefore an enhancement of the properties such as the specific heat and the Pauli paramagnetism.

The *d* states also give rise to the possibility of ferromagnetism in transition metals. Imagine a pair of such states, with opposite electron spins, which lie just below the Fermi energy. If there are electrons in both of these states, there will be an electrostatic repulsion between them. In some cases this repulsion causes only one of the states to be occupied, but it could be occupied with either spin. A local moment (something that behaves as a tiny bar magnet) is said to have been formed. Such a system will have a large response to an applied magnetic field. The field will tend to align the magnetic moments of these electrons, and in this case the alignment can take place without any increase in electronic energy.

When local moments are formed in transition metals, there will ordinarily be an interaction between the moments on different atoms (usually called an exchange interaction) just as there are interactions between bar magnets; thus, in iron, the exchange interaction favours the parallel alignment of the moments on neighbouring atoms so that the lowest energy will be obtained with all moments aligned, which is the ferromagnetic state, the state of ordinary magnetic iron. In some materials the exchange interaction favours an antiparallel alignment, and the lowest state is called anti-ferromagnetic. In either case, the magnetic state is well ordered, but a phase transition will occur if the temperature is raised sufficiently. At this temperature, called the Curie temperature for ferromagnets, the magnetic system becomes disordered, just as the crystalline system becomes disordered at the melting temperature, although there are fundamental differences in the two transitions.

Exchange interaction

Lattice vibrations. The theory of the lattice vibrations requires an understanding of the interactions between the ions in a metal. There are two important factors involved in these interactions: the direct electrostatic interaction, and an indirect interaction, which arises as conduction electrons are successively deflected by two ions. The latter can be calculated in terms of the pseudopotential and, when added to the direct interaction, gives an effective interaction between ions.

Calculations of the vibration characteristics in terms of these interactions have not proved very reliable; but, with minor adjustments, they can be made to agree with experimental results for simple metals. The application to transition metals has not been made.

Lattice vibrations (phonons) affect many of the properties of a metal. The effect of thermal vibrations on conductivity was mentioned earlier. The interaction of a vibration of a given intensity with an electron, called the electron-phonon interaction, may be calculated from the pseudopotential. Probability of a vibration scattering an electron is proportional to the intensity of the vibration; that intensity is, in turn, proportional to the absolute temperature at high temperatures, where the resistivity is found to be proportional to temperature. At low temperatures, the resistivity drops, although, because of scattering by impurities and defects, a residual resistance remains. Another effect of the vibrations is that the speed with which electrons at the Fermi surface move is decreased.

The lattice vibrations can also cause an indirect interaction between pairs of electrons comparable to the interaction electrons cause between pairs of ions. An electron moving through the lattice may jostle the ions, producing a distortion that lowers the energy of an electron subsequently passing through the region. This type of interaction may cause the electronic system to become unstable and form a new phase.

This new phase of the electron system is called superconductivity (see below *Superconductivity*). In many metals (lead, for example) resistivity drops slowly as the temperature drops until at a critical temperature, the resistivity suddenly becomes zero. This phenomenon does not mean that the electrons are no longer scattered, only that scattering no longer reduces the current. The nature of this superconducting state is understood in terms of the indirect interaction between electrons. Electrons moving in opposite directions form so-called Cooper pairs. In the superconducting state one electron scatters inevitably when the other does, and the current carried by the pair is unchanged. The superconducting state represents a kind of order in the electronic system that is analogous to the magnetic order in ferromagnetism.

Surface phenomena. Extensive theoretical study has been devoted to the metallic surface as well as to the behaviour of electrons and ions within the metal, which concern the bulk properties of metals. An understanding of the metal surface requires an understanding of the forces that hold electrons within the metal. An electron striking the surface from within moves slightly beyond the last plane of atoms before returning. There is a layer of negative charge, caused by these electrons, beyond the metallic surface and a compensating positive charge from the ions slightly inside. This double layer of charge provides the electric field that prevents the electrons from escaping. The calculations of the force of this field lead to the finding that an electron with the Fermi energy lacks, by a few electron volts, the energy required to escape from the metal. The barrier is known as the work function of the metal. The work function is very sensitive to the nature of the surface and to any adsorbed (adhering to the surface) atoms, and this theory tells little about surfaces of real metals. It does, however, describe the physics of the work function.

Thermionic emission. Electrons can escape from a metal

if the temperature is high enough; then a small number of electrons will have sufficient energy to overcome the work function and escape. This is thermionic emission, the mechanism by which electrons are introduced into the vacuum by a hot cathode in a vacuum tube. Electrons may also escape by the photoelectric effect, which results from the quantum-mechanical nature of light. The light arrives in quanta of energy, called photons, which vary according to the wavelength of the light. In light of short wavelengths, ordinarily in the visible, the photon energies may be greater than the few electron volts required to free the electron from the metal surface.

Tunnelling. There has been extensive research on the removal of electrons through quantum-mechanical tunnelling. A thin layer of oxide is formed on the surface of the metal and a second metal deposited upon the oxide. The oxide is an insulator, and the electrons respond to it much as they would to a narrow evacuated space; an electron at the Fermi energy does not have the energy needed to traverse it. There is, however, a small probability that, because of the wave nature of the electron, it will nevertheless be transmitted; this is the tunnel effect. Such tunnelling is of particular interest when one or both of the metals are superconductors, because it provides a way of studying the electronic states near the Fermi energy in the superconductor (this is called Giaever tunnelling). If the oxide is sufficiently thin, it is possible for supercurrents to flow without impedance through the oxide, a phenomenon known as the Josephson effect.

Tunnel effect

Skin effect and photoelectric emission. Because of the high conductivity of metals, electromagnetic waves are reflected from a metal surface (this is particularly familiar in the case of light waves that give a fresh metal surface its shiny appearance). The reflection occurs by the generation of electric currents in the surface, which act as antennas reradiating the light outward. Thus, when light is reflected, there are always currents flowing in the surface of the metal, but they are restricted to quite near the surface. This is the skin effect, and the depth of penetration is called the skin depth. It is the field within the skin depth that can give conduction electrons the energy necessary for photoelectric emission. It is the current within the skin depth that, because of metallic resistance, may heat the metal with energy extracted from the reflecting light, resulting in imperfect reflection. This aspect of optical absorption was understood in the earliest days of the free-electron theory of metals and is called Drude absorption, after the German physicist Paul Drude. (W.A.H./Ed.)

SEMICONDUCTORS AND INSULATORS

In contrast to metals through which electric current can flow easily, semiconductors and insulators (Figure 40) offer more resistance to electric charges moving through them. Moreover, whereas heating reduces the electrical conductivity of good conductors and cooling enhances it, semiconductors and insulators are transformed into better conductors when heated and poorer ones when cooled.

General properties of semiconductors and insulators. Characteristic of the electrical resistivity of a poor conductor is its great sensitivity to temperature, pressure, and frequency of electromagnetic radiation. It is also sensitive to its own purity and crystal perfection. For example, the resistivity of the insulator cadmium sulfide (CdS) decreases by 13 orders of magnitude (to one part in 10^{13} of the original value, because one order of magnitude is a factor of 10) upon the addition of less than 0.01 percent of a particular element, referred to as a type of impurity atom; the corresponding change in such typical semiconductors as silicon and germanium is six to seven orders of magnitude. On the other hand, the addition of impurities to typical metals produces a very much smaller change in resistivity, less than one order of magnitude, and the change is of the opposite sign; that is, the addition of impurities increases the resistivity.

Of particular interest is the dependence of resistivity on temperature. In fact, this property is used as a criterion for distinguishing semiconductors and insulators from metals. The resistivity of semiconductors and insulators characteristically increases by many orders of magnitude as the temperature is lowered from room temperature to 1 K (approximately −272° C [−458° F]). On the other hand, the corresponding change in the resistivity of metals is a decrease of several orders of magnitude. That is, the resistivity of metals decreases with decreasing temperature, while that of semiconductors and insulators increases with decreasing temperature.

The classification of materials according to the magnitude and temperature dependence of their electrical resistivity can be quite different from a classification based on visual inspection. Such semiconductors as silicon and germanium, for example, exhibit a metallic lustre, while a metal such as rhenium trioxide (ReO_3) is a reddish, almost transparent solid. Metallic lustre is associated with a high reflectivity at visible light frequencies and is usually, though not necessarily, found in materials with high direct-current conductivities. The electrical conductivity of all solids is dependent on the frequency of the electromagnetic radiation: for metals, conductivity tends to decrease with increasing frequency; for semiconductors and insulators, on the other hand, conductivity tends to increase with increasing frequency. Thus, at ultraviolet frequencies (higher than those of visible light, the optical region) the conductivity of all solid materials is about the same order of magnitude.

Visual properties

Although the conductivity at optical and ultraviolet frequencies is similar for the various classes of materials, the effect of light exposure on low-frequency conductivity (*i.e.*, alternating currents in the material of frequencies lower than those of the radiation) is different. For example, the direct-current conductivity of cadmium sulfide can change by eight orders of magnitude (100,000,000 times) upon exposure to visible light. The increase in electrical conductivity exhibited by certain solids when exposed to light is called photoconductivity. This effect is common in insulators and semiconductors, but it is not observed in metals and semimetals; in semiconductors, such as silicon and germanium, photoconduction is produced by infrared as well as by visible light excitation.

Photoconductivity

The reason solids are classified according to their direct-current resistivities is partly because of a high correlation between this property and other physical properties. There is also a practical reason for doing so: most of the industrial applications of semiconductors and insulators make use of the low-frequency electrical properties. These applications characteristically exploit the great sensitivity of the resistivity of low-conductivity materials to small variations of external parameters. For example, the large photoconductivity of cadmium sulfide makes this an ideal material for an electrical switch triggered by light. Indeed, rectification (one-directional conduction) and other properties of semiconductors are exploited in the modern electronics industry and computer technology. In its infancy the electronics industry employed semiconductors as rectifiers in crystal sets for radio reception. These crystals were soon replaced by the then more reliable and effective vacuum tube, which controlled current by a relatively weak flow of free electrons through a vacuum, and which dominated the electronics industry for four decades until the invention of the transistor in 1947. (A transistor is a device that utilizes a semiconductor to control or to amplify, or both, small electrical currents. In 1956 three Bell Telephone Laboratories scientists, William Shockley, John Bardeen, and Walter Brattain, were awarded the Nobel Prize for Physics for their investigations of semiconductors and for their discovery of the transistor effect.) Because of this invention and the technological development that it inspired, semiconductor components have replaced vacuum tubes for most purposes. Their compactness, simplicity, durability, and reliability are all responsible for the general industial acceptance of solid-state devices for a wide variety of applications: from the transistor radio in the home to the hardware of giant computers; from solar cells for converting light into electricity to the electronic circuitry of rockets and satellites in the space program.

Mechanisms of conduction. In order to understand the nature of electronic conduction in semiconductors and insulators, it is necessary to study in some detail the conduction mechanisms appropriate to these materials.

Electrical conduction is associated with the flow of charge between two electrodes attached to two points on these materials, from a positively charged anode to a negatively charged cathode or the reverse, and depends upon the concentration (N) of charged particles, or carriers, in the material. Whereas electrons are the dominant current carriers in metals, semimetals, and semiconductors, for some insulators both electrons and the heavier ions (charged atoms) contribute to the conduction of current.

Mobility of charged particles. To transport effectively electric charges from one electrode to another, the particles must resist collision and scattering along the path of the charge. Thus, charged particles in a solid are not accelerated indefinitely by the application of an external electric field because collisions force these particles to achieve some average velocity, called the drift velocity (v_D); each scattering event tends to randomize the velocity achieved in the acceleration process, so that after each such event the charged particle is effectively returned to zero average velocity, and the acceleration process caused by the electric field must begin again. The mobility, symbolized by the Greek letter mu (μ), which measures the effectiveness of the charge transport by a given charged particle in a conductor, semiconductor, or insulator, is thus defined as the drift velocity achieved through application of unit electric field (E), being equal to the drift velocity divided by the electric field; *i.e.*, $\mu = v_D/E$. That is, as the external field is increased, the drift velocity increases proportionately. Because the drift velocity depends on the magnitude of the applied electric field, the mobility of a particle of mass m and electric charge e can be more directly related to the mean time between scattering events, symbolized by the Greek letter tau, τ, through an equation in which mobility equals the product of the charge and the mean time between scattering events divided by the mass, or $\mu = e\tau/m$. In this way mobility is related directly to the interval of time between collisions; that is, to the effectiveness of a charged particle in avoiding collisions.

Drift velocity

Electrical conductivity (σ) equals the product of the mobility of each carrier (μ) and the concentration (number per unit volume) of carriers (N) with charge e; it may be defined as $\sigma = Ne\mu$. This definition of electrical conductivity is particularly useful for semiconductors and insulators because of the independence of the two factors, the carrier density factor (N), which is highly sensitive to such external parameters as temperature, optical excitation, and others, and the mobility factor (μ), which is sensitive to scattering mechanisms within the material.

Mobility in solids. Whereas conduction by ions is generally ineffective because of low mobility, conduction by electrons in metals, semimetals, semiconductors, and insulators is efficient. In fact, because of quantum mechanical principles, an electron travelling through a perfect crystal lattice is not scattered. The present understanding of electronic conduction in solids goes back to the classical Drude theory (1900), so called for Paul Drude, the German physicist who presented it, and according to which it was supposed that a dense gas of free electrons, interacting strongly with each other, existed in a metal. This theory has since been modified to explain why only a fraction of the available electrons participate in conduction and why the electron scattering times are longer than predicted by classical theory.

Electron diffraction

The wavelike nature of electrons gives rise to diffraction (resulting when obstacles interrupt a sequence of waves) rather than to scattering (resulting when a stream of moving particles meets an obstacle), as the wave traverses a crystal lattice (an orderly arrangement of atoms which act as obstacles). Thus, in certain directions the waves interfere constructively to yield diffraction peaks of intensity, whereas in other directions destructive interference results. From this point of view, although no scattering occurs, electronic conduction in a periodic (repeating) array of atoms is modified by the presence of the lattice. For most crystalline semiconductors and insulators, the effect of the periodic lattice on the conduction process can be described by an equation in which the electron mobility (μ) is equal to the product of the charge and the mean time between scattering events divided by a factor called the effective mass (m^*): $\mu = e\tau/m^*$ (see below). In semiconductors the effective mass tends to be considerably lower than the free electron mass (m)—*i.e.*, the mass of an electron at rest—whereas in insulators the effective mass tends to be higher than the free mass.

Electron scattering

The most important room temperature scattering mechanism in solids is usually the scattering of electrons by the thermal (heat) vibrations of the atoms of the lattice. Because of this mechanism, electrical resistivity varies with temperature, a relationship that is generally valid for materials in which the carrier density is independent of temperature. As the temperature is lowered, the thermal lattice vibrations become less important, and scattering is dominated by crystalline imperfections, such as lattice vacancies or interstitial atoms, lattice dislocations, impurity atoms, and crystal boundaries. Thus, electron scattering is increased by the addition of impurity atoms or by destroying the crystalline order of the solid; in fact, for amorphous (noncrystalline) materials, scattering by these imperfection mechanisms is important even at room temperature.

Because of the increasing importance of lattice vibrations with increasing temperature, the mobility of electrons in solids tends to decrease with increasing temperature. Moreover, because the carrier concentration (density) in metals and semimetals is essentially temperature independent, the increased scattering with increasing temperature leads to a characteristic decrease in conductivity with increasing temperature. For semiconductors and insulators the number of carriers is extremely temperature dependent; in fact, the carrier density increases exponentially with increasing temperature. Such a rapid increase in the carrier density completely overshadows the very much slower decrease in the mobility; thus the conductivity of semiconductors and insulators characteristically increases with increasing temperature.

Activation energy

The strong temperature dependence of the carrier density in semiconductors and insulators is intimately connected with an activation energy, or barrier, which restricts carriers to bound, nonconducting states. This barrier can be overcome by thermal excitation of the particles, which has an exponential temperature dependence through the Boltzmann factor that relates the temperature to the energy of the particle. Thus, the carrier density (N) is equal to the product of the density of the available bound carriers (N_o) and the exponential of minus the ratio of the activation energy (E_g) to twice the Boltzmann constant (k) times the absolute (Kelvin) temperature (T), or $N = N_o \exp(-E_g/2kT)$. The temperature dependence is extremely high, especially in the temperature range below the characteristic temperature T, approximately equal to E_g/k. In insulators this region is generally equal to or greater than 10^5 K, and the room temperature conductivity is very small. For pure semiconductors, the characteristic temperature is equal to or greater than $T_c \geqq 10^4$ K, and even here the room temperature conductivity is low. Because of the strong temperature dependence of the conductivity in semiconductors, a constant temperature must be maintained for stable operation of semiconductor devices. In such devices it is important to minimize any power flow and power loss that could cause heating effects, because these would greatly increase the number of carriers, ultimately resulting in avalanche breakdown, a phenomenon that is similar to breakdown processes occurring in the air during a thunderstorm to produce lightning.

Conduction in semiconductors and insulators. Conductivity in very pure semiconductors is characterized by extremely strong temperature dependence. As impurities are added, however, the temperature dependence of the carrier density becomes weaker; upon heavy doping with impurity atoms, the carrier density becomes essentially temperature independent, thus exhibiting semimetallic behaviour.

To gain a better understanding of the effect of impurity atoms on electrical conductivity, it is instructive to take a more microscopic view of a solid and to inquire into the nature of chemical bonding. In this way the interactions of single ions with their neighbours can be examined, and the microscopic interactions can be related to the macroscopic properties of the material. Contrasting with this chemical view of matter is the physical approach, whereby

the solid is considered as an immobile structure composed of units arranged in a precise geometric pattern that can be repeated in any direction to extend the structure infinitely. Such a structure is periodic, and this periodicity is exploited to describe the electronic behaviour.

Chemical approach. Perhaps the most powerful application of the chemical approach has been directed toward the question of which atoms and crystal structures favour the formation of insulators, semiconductors, and metals. The approach is well illustrated by the important semiconductors silicon and germanium, which are chemically related to the element carbon and the atoms of which have the same chemical bonding as do the carbon atoms in diamond crystal structure.

Chemical bonding

Each carbon, silicon, or germanium atom has four valence electrons. (By gaining, losing, or sharing valence electrons, atoms can bind with one another.) Because there are about eight valence states of about the same energy the valence shells of carbon, silicon, and germanium are only half filled. To occupy the four remaining empty states and fill up the shell, each atom shares one electron with each of four neighbouring atoms. Thus, silicon and germanium (like carbon) crystallize in the diamond structure, in which each atom is bound to its four nearest neighbours in a tetrahedral arrangement.

Bonding characterized by the sharing of an electron between neighbouring atoms is called covalent bonding. The covalent bonds in diamond are so strong that it is extremely difficult to remove an electron from any bond. It is the strength of covalent bonding between carbon atoms that accounts for the extreme hardness of diamonds. In germanium and silicon, however, the covalent bonds are weaker because the valence electrons are in shells that are farther from the nucleus than the valence shell of the carbon atom. In such cases it is relatively easy to promote an electron from a covalent bond into a conducting state, and much of semiconductor technology is concerned with controlling the number of broken bonds that can contribute electrons to the conduction process. Increasing the temperature provides thermal energy for breaking bonds, thereby freeing electrons for conduction purposes.

Addition of impurities

The periodic table of the elements is an invaluable aid in comparing their properties. Elements that this system places in the same vertical column, or group, are similar because they have similar electron structures. The elements carbon, silicon, germanium, and tin are in Group IV of the periodic table. Impurities from neighbouring Groups V or III can be introduced substitutionally into the diamond structure of these elements. While the presence of the impurities has little effect on the crystal structure, they profoundly affect electrical conductivity. For example, antimony (Sb), an element in Group V, has five electrons in the valence shell, one electron more than that needed to satisfy the covalent bonding requirement. When antimony is used as an impurity atom in silicon, only a small amount of energy is needed to liberate the extra electron, which is weakly bound to the positively charged antimony ion (Sb^+). Impurities that donate electrons to the conduction process are called donors; here, electrical conduction is by negatively charged carriers and the materials are called *n*-type.

On the other hand, indium (In), an element in Group III, has only three electrons in its valence shell, which means a one-electron deficiency relative to the covalent bonding requirement. Because the covalent bond is so strong when indium is used as an impurity atom in silicon, the electron deficiency in the indium atom can be satisfied by taking an electron from a covalent bond elsewhere in the material. The covalent bond that now lacks an electron is called a hole and acts like a positive charge; it is free to hop from covalent bond to covalent bond in the material. Impurities that receive electrons are called acceptors; thus, an effective positive charge, a *p*-type, has been introduced into the conduction process.

Because the impurity atoms in semiconductors and insulators tend to be relatively well isolated from each other, the chemical approach is particularly fruitful. For example, the chemical mass-action law (the rate of a chemical reaction depends on the concentrations of the reacting substances) can be applied to determine the degree of ionization of the impurity atoms and hence the carrier density in the semiconductor.

Compound semiconductors

Silicon, germanium, diamond, and gray tin form the Group IV semiconductors, so called because their atomic structure puts them into the fourth group, or column, of the periodic table. (Other forms of carbon and tin, such as graphite and white tin, do not act as semiconductors.) A class of closely related materials is the III-V compound semiconductors, in which one atom of an element from the third group of the periodic table is surrounded by four neighbouring atoms of an element from the fifth group of the periodic table and conversely, so that the number of atoms from each group is equal. By transferring an electron from a fifth-group atom to a third-group atom, each lattice site becomes occupied by an ion surrounded by four oppositely charged ions in a tetrahedral arrangement, with the electronic bonding similar to that of the diamond structure. Actually, however, the crystal structure for the III-V compound semiconductors is the zinc-blende structure, which is like the diamond structure except that nearest-neighbour atoms are of a different species. An example of a III-V semiconductor is gallium arsenide, in which the bonding is mainly covalent, though some ionic bonding, and therefore Coulomb attraction (the electrostatic attraction between oppositely charged particles), exists between the oppositely charged ions. Because covalent bonding dominates these compounds, they have semiconducting properties similar to the Group IV materials.

Hopping process

The chemical structure approach is useful not only for understanding compound semiconductors but also for explaining molecular solids, which include all solids of organic molecules as well as such materials as solid hydrogen, which has a hexagonal arrangement of hydrogen (H_2) molecules. In molecular solids the dominant conduction mechanism is the hopping of a single electron from one molecular complex to another. Because this process typically involves an activation energy, however, it is not an efficient conduction mechanism. Consequently, molecular solids tend to be poor conductors of electricity.

Ionic crystal conduction

The hopping process is also an important conduction mechanism in ionic crystals. Here, however, the process involves either a lattice vacancy or an interstitial ion, rather than simply an electron. Just as in the case of electron hopping in molecular crystals, the hopping ions in ionic crystals must overcome a potential activation energy barrier by virtue of their thermal motion. The hopping barriers are lower for small ions than they are for larger ions; hence, a solid with small ions (*e.g.*, H^+ ions in ice) will tend to have relatively higher ionic conductivity. The relative sizes of the ions are important in determining the barrier height. In photosensitive materials, such as silver bromide (AgBr), the barrier of the smaller positive silver ions (Ag^+) is much less than that of the larger negative ions (Br^-), a difference that is an essential ingredient in the photographic process.

It is found that ionic conductivity is enhanced by the addition (doping) of doubly charged positive ions to alkali halides—*e.g.*, adding calcium chloride to potasium chloride. The simplest explanation of this is that the doubly charged calcium ions of calcium chloride are more mobile than the singly charged potassium ions, but this is not the explanation. Rather, since the crystal tends to remain electrically neutral, there will tend to be a positive ion vacancy assciated with each calcium ion (Figure 44). Conductivity will then occur through the vacancy mechanism.

A practical aspect of both ionic and electronic conductivity of insulators comes from so-called space-charge effects. For example, if it is assumed that conduction is by positive vacancy migration and that the crystal is in an electric field but that the electrodes do not touch the crystal; then, after some time, there will exist a layer of positive charge in the crystal near the negative electrode and a layer of negative charge near the positive electrode. These space charges will set up an electric field inside the crystal in the direction opposite to that of the applied field, which will decrease the ionic currents. Similar problems are associated with electronic conductivity in insulating crystals and can occur even when the electrodes touch

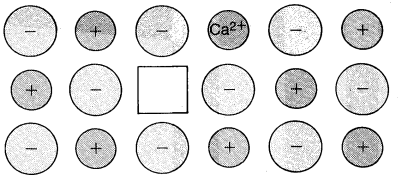


Figure 44: Production of a vacancy to enhance ionic conductivity by adding calcium chloride to potassium chloride. For each calcium ion (Ca^{2+}) there tends to be a positive ion vacancy, indicated as a square, which permits conductivity through the vacancy mechanism.
From C. Kittel, *Introduction to Solid State Physics* (1966); by permission of John Wiley & Sons, Inc.

the crystal. In some cases, in fact, the building of space charge is an important factor in lowering the conductivity of the crystal.

Amorphous solid semiconductors

The chemical approach has also proved to be very effective in dealing with amorphous solids that do not exhibit any long-range periodic structure. Such solids are noted for their durability, their insensitivity to variations of temperature and pressure, and their insensitivity to the introduction of impurities, vacancies, and interstitial atoms. As noted above, the most common example of an amorphous material is glass, which consists mostly of a random array of silicon dioxide (SiO_2) molecules as well as impurities. The crystalline form of pure silicon dioxide is quartz, a well-ordered structure. When the ordering is destroyed, a glass results, and some of the properties of the material are changed. For example, fused quartz, the amorphous version of quartz, has a much lower conductivity than crystalline quartz; the ultraviolet properties of fused quartz and crystalline quartz, however, are almost identical.

Physical approach. While the chemical approach is advantageous in studying the more complicated materials, such as molecular and amorphous solids, the physical approach is probably more advantageous in dealing with simple, highly ordered crystalline solids, thus leading to an understanding of semiconductors and insulators. A crystalline solid is conceived of as an infinite, perfectly periodic array of atoms. Initially, this point of view may seem more complicated than that of the chemical approach, in which only a few interacting atoms are considered; a great deal of progress can be made, however, by exploiting the symmetry of the periodic structure, especially the translational symmetry (the movement of a configuration to new coordinates without changing its symmetry) of the perfectly periodic array.

Energy bands and gaps

As a direct consequence of translational symmetry, the discrete energy levels of the free atom spread out to form two kinds of energy bands in the solid, valence bands and conduction bands. The valence bands are those occupied by electrons and the conduction bands are normally unoccupied. If the conduction bands of a semiconductor become partly filled by thermal excitation, the solid becomes conducting. For materials in which the electrons are tightly bound to the atoms, the width of the energy bands is sufficiently small for each band of energy levels to be separated from neighbouring bands by an energy gap. Whereas there are no available valence states in the energy gaps, the number of available states in each energy band would be completely filled if each atom would contribute exactly two electrons.

The result for the number of available states is of great consequence because of the Pauli exclusion principle (*i.e.,* only one electron may occupy a particular energy state), and the result does, in fact, imply that every solid with an even number of electrons is an insulator, and every solid with an odd number is a metal.

The situation in actual solids is not quite so simple for several reasons. Band degeneracies (*i.e.,* several levels with the same energy) occur because of rotational point symmetries (the configuration remains unchanged upon performing a rotation) in the crystal structures. Such band degeneracies characteristically are present in the bands associated with the valence electrons of semiconductors that have the diamond and zinc-blende crystal structure. A second complication, prevalent in semimetals and metals, is the overlapping of neighbouring energy bands. Overlapping arises when the energy that binds electrons to their ion cores becomes weaker, causing electrons to move farther away, thus increasing the band widths and, consequently, decreasing the band gaps from a finite value to zero and, finally, to the overlapping condition. A third complication is associated with the complex arrangement of atoms within the solid. For example, the concept of band occupation given above indicates that solid hydrogen should be a metal (a hydrogen atom has only one electron, an odd number), whereas the solid is actually an insulator consisting of a hexagonal lattice of hydrogen (H_2) molecules. There is, however, a high-pressure phase of solid hydrogen that is metallic, because when the distance between interacting molecules is reduced, the crystal structure changes, with a resulting change in the electrical conductivity.

It is also possible to form, from the same atom, solids with different electrical properties depending on their crystal structure, a striking example being tin. White tin crystallizes in a tetragonal arrangement and is a high-conductivity metal, while gray tin is a covalently bonded semiconductor crystallizing in the diamond structure. In the case of tin, certain partly filled energy levels (p levels) of the free atom become conduction bands for metallic white tin, whereas the greater symmetry of the diamond structure of gray tin results in a splitting of the p bands into filled bonding and empty antibonding states, thus giving rise to semiconducting behaviour.

Theoretical approximations

The physical approach to the electronic properties of solids uses the translation symmetry of the periodic structure to establish the existence of valence and conduction energy bands. The detailed treatment of energy bands, however, is an exceedingly difficult matter and can be carried out only in an approximate way. Much of the theoretical work results in equations that employ only a small number of factors determined by experiment. Among these factors are the energy gap, which separates the occupied valence states from the empty conduction states, and the effective mass, which describes the dynamic motion of electrons in a solid. By introducing the effective mass concept, it is possible to treat the electrons in a solid as free electrons that possess a mass different from the free-electron mass. The effective mass approximation, whereby the effect of the periodic lattice is taken into account by introducing an effective mass, is extremely useful in dealing with many of the physical phenomena in semiconductors and insulators.

The lattice

The lattice, however, cannot be completely ignored. The atoms of the lattice are in a state of constant vibration, and the characteristic, or normal, mode frequencies of these lattice vibrations have been studied extensively for semiconductors and insulators. The number of normal modes of vibration is completely determined by the crystal structure and the number of atoms per unit cell. For example, in the diamond structure there are three acoustic (sound) modes and three optical (light) modes. The characteristic frequencies of lattice vibrations are determined by a variety of experimental techniques.

Lattice vibrations have an important influence on electrical conductivity because they provide the dominant scattering mechanism for solids at room temperature. In the scattering process, the energy of the electrons is either increased or decreased by the characteristic energies of the lattice. The conduction electrons contribute to both the electrical conductivity and the thermal conductivity of solids. Metals are characteristically good conductors of both heat and electricity, whereas electrical insulators are also typically good thermal insulators.

Statistical properties

The physical approach to solids uses the translational symmetry associated with the periodic lattice to establish the characteristic vibrational frequencies for the lattice. The physical view also exploits the statistical properties of a large ensemble of particles to determine such fundamental properties as the carrier densities of solids. Like electrons, ensembles of particles obey the Pauli exclusion principle, and the statistical behaviour of such ensembles is described by the Fermi–Dirac distribution function, $f(E)$, which expresses the probability that a given state of

energy E will be occupied at temperature. This relationship can be expressed by the following equation, in which the Fermi energy, or chemical potential, is symbolized by the Greek letter zeta, ζ, and k is Boltzmann's constant:

$$f(E) = [1 + \exp\{(E - \zeta)/kT\}]^{-1}. \quad (7)$$

The significance of the Fermi energy (ζ) is that, at $T = 0$ K, all the states up to energy ζ are completely occupied, while those above energy ζ are completely empty. At higher temperatures a fraction of the states above the Fermi energy will be occupied and, consequently, some states below ζ will be empty in accordance with the distribution function $f(E)$. The distribution function can be applied directly to metals to find, at a given temperature, the number of excited electrons above the Fermi energy; such excited electrons are directly responsible for the electronic contribution to the specific heat of metals and for the electronic Pauli paramagnetism of metals.

In applying the Fermi–Dirac distribution to semiconductors, it is necessary to take into account the presence of an energy gap above the valence bands that is forbidden to electron states. For pure semiconductors, at temperature equal to 0 K, the states below the energy gap are completely occupied, those above are completely empty, and the Fermi energy lies within the band gap. If the energy gap is small compared with the product of Boltzmann's constant and temperature, kT, then the Fermi–Dirac distribution function implies that the thermal energy will be strongly effective in exciting electrons into conduction states located in the conduction band. For every electron that is excited into a conduction state, an unfilled valence state is created in the valence bands; unfilled states called holes and, for conduction purposes, behave very much like positive charges with masses comparable to electron masses.

For typical semiconductors, such as germanium and silicon at room temperature, the energy gap is very much greater than thermal energies, so that only a few thermally excited carriers are present. In this limit, the temperature dependence of the carrier concentration for both electrons and holes is exponential and of the form $N = N_0 \exp(-E_g/2kT)$, as explained above under *Activation energy*. Both the excited electrons and the holes that are left behind contribute to the conduction process, and for a semiconductor of high purity they are the predominant current carriers at room temperature.

Nonconducting states

The physical approach to solids is also useful in treating certain nonconducting excited states in semiconductors that are associated with the attraction of an electron to the hole that it has left behind. Such states, which are commonly called excitons, either lie below the internal ionization energy of the solid or lie within the energy gap, just below the condcuting states. An exciton can be transferred from one lattice site to another by a hopping mechanism. Excitons are generally more important in semiconductors and insulators with large band gaps, such as ionic crystals, than in those with small band gaps, although weak exciton effects have been observed in the latter. Because the exciton does not obey Fermi–Dirac statistics whereas the electron does, it is not possible to obtain a filled band of excitons. The Coulomb attraction of the exciton is in many ways suggestive of the attraction of the electron to the hydrogen ion in the hydrogen atom or of the electron to the positron (a negatively charged electron) in positronium (a short-lived system containing an electron and positron that revolve around a common centre). Theoretical treatments of excitons based on models similar to the hydrogen atom or to positronium have been very successful in explaining many of the characteristic properties of excitons.

Nonconducting states within the energy gap are also created by the presence of impurities. Because the free charges associated with impurities tend to screen the attraction between the electron and the hole associated with it, however, the exciton levels are generally observed in materials of higher purity than are desirable for the study of impurity levels.

Extrinsic and intrinsic semiconductors. From the physical point of view, shallow impurity levels in a solid (impurities that have a potential not too different from the host atoms) are generally treated by using models based on a hydrogen-like atom with an electron of effective mass. Impurity levels, unlike states within the conduction and valence bands, tend to be sharp, or discrete; in the simplest cases, the discrete donor (excess electron) levels lie just below the conduction band minimum, while the acceptor (deficiency of electrons) levels lie just above the valence band. Because these levels are easily ionized to create electrons or holes in conducting states, impurities provide a ready source of charged carriers for electrical conduction.

When conduction is dominated by impurity mechanisms, the semiconductor is called extrinsic, in contrast with intrinsic semiconductors, in which conductivity is mainly associated with the thermal excitation of carriers. An extrinsic semiconductor is generally classified as *n*- or *p*-type, depending on whether the impurity has an excess of negative charge (*n*-type) or a deficiency of negative charge (*p*-type). Under some circumstances, an intrinsic semiconductor can also be classified as *n*-type or *p*-type, because the mobility of the electrons and holes may be quite different although their carrier densities are closely equal. In such cases the conduction process itself is dominated by the carriers with the highest mobility.

Measurement of conductivity. A direct measurement of the carrier concentration in a semiconductor is obtained from a phenomenon, called the Hall effect after its discoverer, which makes use of the deflection of the electric current by a transverse direct-current magnetic field. The Hall measurement provides both the sign and concentration of the carriers. A separate measurement of the conductivity allows one to calculate the mobility of the carriers. Because these factors are so important for the characterization of semiconductors, conductivity and Hall-effect measurements are routinely carried out at room temperature and at liquid nitrogen temperature (77 K [−196° C or −321° F]) when semiconducting crystals are produced. The temperature dependence of carrier density helps to determine whether the semiconductor is intrinsic or extrinsic.

The conductivity or resistivity of a solid is itself dependent on the magnitude and direction of the magnetic field relative to the current flow. This magnetoresistance effect is associated with the curling up of the electron trajectories in the magnetic field, thereby decreasing the mean free path (the average distance that a particle travels between successive collisions) of the carriers. Such experiments yield valuable information on the directional dependence (anisotropy) of the effective masses of the carriers.

Magnetoresistance effect

The most direct information, however, on the effective mass for electrons and holes in semiconductors is provided by a cyclotron resonance experiment in which the charged particles can be considered to execute closed orbits in a plane perpendicular to the magnetic field. The experiments establish that the electrons in germanium have an effective mass that is approximately 10 times smaller than the free-electron mass; even smaller effective masses have been found in other small-gap semiconductors.

Measurement of energy gaps. The other dominant parameter (a constant, in equations, to which any desired value can be given) that characterizes semiconductors and insulators is the band gap separating the valence states from the conduction states. The most direct measurement of this energy gap involves the excitation of an electron in a valence state to a state in the conduction band through the absorption of a photon. The process, called the internal photoelectric effect (the action by which a substance emits free electrons as a result of absorbing light energy), is closely related to the photoelectric effect at a metallic surface. The energy gap is determined as the threshold for the onset of the photoelectric process and typically occurs at infrared or optical frequencies. Optical spectral techniques also provide a powerful tool for the study of discrete exciton and impurity states. The optical excitation of carriers is also exploited in a variety of photoconductive detectors, which are characterized by fast response, high sensitivity, and availability over a wide range of infrared and optical frequencies.

Internal photoelectric effect

Principles of semiconductor applications. *Photographic process.* Perhaps the most common application of the photoelectric effect is in the photographic process. Although this process has been studied extensively for many years, the detailed mechanism is not yet completely understood. Photographic film contains small crystallites of silver salts, such as silver bromide or silver iodide, which are darkened when exposed to light and subsequently developed. The initial exposure to light sets the image in the photographic process; it is associated with the following mechanism.

The incident light frees an electron, which becomes trapped at some impurity site (*e.g.,* thallium). The silver ions, which have a relatively high mobility, diffuse through the lattice and migrate to those sites where the electrons have been trapped. At one of these sites the silver ion becomes neutralized, forming a neutral silver atom and thereby restoring the impurity site to its original state. In this way each impurity site can collect several neutral silver atoms, forming a collection of silver particles called the latent image. When the film is developed, the small silver particles in the latent image grow, and finally the silver salts that are in the ionic form are washed out, retaining the colloidal silver particles. In colour photography a separate image is formed for each of the three primary colours, red, green, and blue.

Luminescence. Also occurring in semiconductors and insulators is a process inverse to the photoelectric effect, one in which the recombination of an electron with a hole or with an impurity centre results in the emission of a photon. The effect, which is called luminescence, is essential in many devices using semiconductors and insulators; for example, a laser is a device that emits light by virtue of releasing radiation stored by creating electron and hole pairs.

Luminescence is also basic to the operation of scintillation counters, which are so important in detecting and measuring high-energy particles. In these devices a high-energy gamma ray (photons with frequencies much higher than visible light) is made to produce a larger number of photons in the visible range (photons with much lower frequencies). The mechanism for such photon production is as follows: the high-energy gamma ray generates very energetic electrons and holes that rapidly become degraded by collisions with other electrons; thus, when a highly energetic gamma ray is absorbed by an insulating material, such as sodium iodide (NaI), it effectively produces a large number of excited electrons, which, in turn, can recombine with an impurity centre or a hole by a luminescent process to produce many photons of visible light, which are then detected optically.

Hot-electron effects. Another property of semiconductors and insulators that has great practical significance is the nonlinear response (*i.e.,* nonproportionality) of the current flow to the applied electric field. For small electric fields the current flow is directly proportional to the applied electric field, and the electrical conductivity (the proportionality constant) is independent of the magnitude of the applied field. For sufficiently large electric fields, however, the conductivity itself can become dependent on the field, either increasing with increasing electric field (giving rise to avalanche effects), or decreasing, or even becoming negative. Nonlinear effects in semiconductors and insulators are sometimes called hot-electron effects because they involve a nonlinear dependence of the Fermi–Dirac distribution function on the applied electric field.

Avalanche breakdown

Avalanche breakdown, or sparking, is a common hot-electron effect in insulators. In the presence of a sufficiently large electric field, a conduction electron can pick up enough energy in a mean free path to induce ionization in the solid. With this mechanism for the creation of additional carriers, the current flow can increase catastrophically and the concomitant local heating can result in actual physical damage to the material. Similar avalanche processes occur in gases and are responsible for the operation of such devices as Geiger counters.

Practical devices are also based on hot-electron effects that result in a decrease in current flow with increasing applied electric field. Such a nonlinear effect (the Gunn effect) was discovered in gallium arsenide (GaAs) by a British-United States physicist, J.B. Gunn, in 1963 and is the basis for the Gunn oscillator, which supplies large quantities of power at microwave frequencies. In such nonlinear semiconductors, small electric fields excite carriers into high-mobility conduction states; but above a certain threshold field, excitation into low-mobility states becomes energetically possible. With increasing electric fields, the low-mobility states are occupied with increasing probability, and a decrease in conductivity results. This negative conductivity is basic to the operation of the Gunn oscillator.

Junction effects. Solid-state devices can make use of either bulk effects or junction effects. (Some typical examples of bulk-effect devices are Gunn oscillators, photoconductors, and ruby lasers.) Junctions exist between different semiconductors, or between semiconductors and metals, and are particularly important in devices based on a single junction.

When two dissimilar materials (in semiconductor electronics the most common are silicon and gallium arsenide) are placed in contact, there will be a net flow of electrons from one material to the other until the Fermi levels or chemical potentials of the two materials are equal. Thus, in equilibrium, it is possible to take an electron across the junction between the two materials without the expenditure of any energy. In equilibrium there is no net current across the junction, and the hole current is equal and opposite to the electron current.

The *p–n* junction

Perhaps the most common junction is the *p–n* junction, which is formed by contact between *n*-type and *p*-type samples of the same material. In the immediate neighbourhood of the junction there is a depletion layer, a region in which the carrier concentration is low, reflecting the fact that there are no allowed conduction states at the Fermi energy in this region of the crystal (see Figure 45B). In the depletion layer the electrons must tunnel (pass through at very high speeds) to overcome the potential barrier associated with the contact potential. For practical applications of *p–n* junctions it is important that the junctions be kept narrow and that their physical characteristics be carefully controlled.

The properties of the junction can be altered by superimposing a voltage across it or by modifying the junction characteristics (for example, the width). The effect of an external bias on a *p–n* junction is illustrated in Figure 45.

From S.M. Sze, *Physics of Semiconductor Devices* (1969); John Wiley & Sons, Inc.

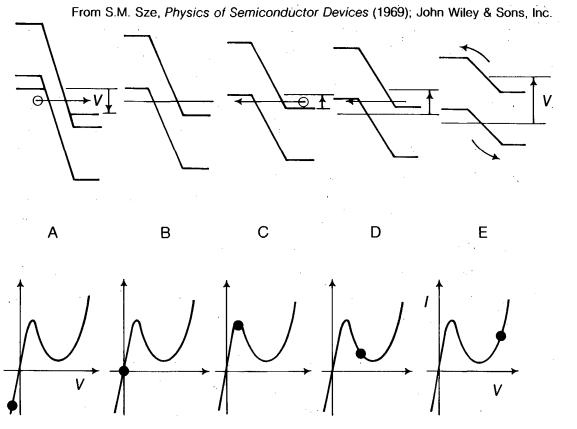


Figure 45: Simplified energy-band diagrams of tunnel diode at: (A) Reverse bias. (B) Thermal equilibrium. (C) Forward bias such that peak current is obtained. (D) Forward bias such that the valley current is approached. (E) Forward bias with thermal current flowing (see text).

Application of a reverse bias (Figure 45A) produces a large reverse current, which is illustrated by the dot directly below the *I-V* (ampere-voltage) curve. The reverse bias increases the potential of the *p*-type region relative to the *n*-type region, and conversely the forward bias increases the potential of the *n*-type region relative to the *p*-type region. Application of a forward bias (Figures 45C, 45D, 45E) eventually leads to a dramatic decrease in current or a negative conductivity as the number of available states vanish (Figure 45D). Other mechanisms, for example photon- or phonon-assisted tunnelling, provide some

current for biassing voltages. Furthermore, by application of an extremely large forward bias, large currents once again become possible (Figure 45E).

Semiconductor and metal junctions

A semiconductor in contact with a metal also forms a barrier with carriers removed, on the semiconductor side of the barrier. These junctions are important as rectifying contacts and operate in the following manner: because the semiconductor effectively lowers the work function (energy needed to eject an electron) of the metal, electrons can be pulled out of the metal and into conducting states of the semiconductor with the application of reasonably small electrical fields. The reverse bias, however, does not produce currents, because the filled levels of the semiconductor are below the Fermi level of the metal.

In addition to rectifying contacts, it is also possible to produce internal photoemission with a semiconductor–metal junction. The light can be incident on the junction, either through the semiconductor (it must be of low frequency so that bulk absorption does not occur) or directly onto the very thin metal coating that covers the semiconductor. In either arrangement, internal photoemission occurs at a lower threshold than for a metal-vacuum interface. The threshold can also be changed by the application of an external potential across the junction.

The internal photoemission process is basic to the operation of a solar cell, which consists of a *p–n* junction covering the entire external surface of a semiconductor. Light is absorbed in the *p–n* junction, creating a potential difference that in turn gives rise to a tunnelling current. Thus, the light is converted directly into electrical energy. The experimental conversion efficiency for the process is about 10 percent. Because solid-state electronics is widely used in space statellites, solar cells provide adequate power for the necessary circuitry.

A *p–n* junction not only can convert light into electrical energy but can also accomplish the inverse process; that is, convert electrical energy directly into light. The emitted radiation, which is produced by the recombination of holes and electrons that have been injected into the junction region by application of a forward bias, has a frequency characteristic of the energy gap. By polishing crystal faces on either side of a *p–n* junction, a resonant optical cavity, or Fabry–Perot interferometer, is formed from which the recombination radiation can be emitted coherently—that is, the emitted waves are in phase (in step) with each other—thereby producing laser action. Semiconducting diode lasers are of very small size and typically operate at an infrared frequency between 0.04 and two electron-volts, depending on the material. The tuning of a particular laser frequency is accomplished either by temperature variation or by application of an external magnetic field or mechanical stress.

Pressure effects. Application of pressure results in a lattice deformation, which changes the lattice constant and hence the energy gap of the semiconductor; this mechanism is essentially that behind the pressure tuning of lasers. The same kind of mechanical deformation can also greatly influence the electric conductivity of a solid, and the resulting current flow gives rise to piezoelectric effects, the production of electric charges when certain crystals are deformed by mechanical stress, or, conversely, the deformation of a crystal caused by the application of a voltage. This conversion of mechanical motion into an electric current is exploited in such practical devices as phonograph pickups and sonar detectors. Piezoelectric materials that are used for such applications typically have a strong coupling between the electronic band structure and the characteristic modes for the lattice vibrations. An important class of piezoelectric materials are the ferroelectrics (ionic crystals in which the molecules are aligned with one another), such as barium titanite.

Thermoelectric effects. Semiconductors also have proved to be among the most useful materials for the production of devices based on thermoelectric power conversion. The Seebeck effect, discovered in 1821, represents a means of converting heat energy directly into electrical energy, whereas the Peltier effect (1834) represents a means of using electrical energy for refrigeration. Practical devices based on these two effects make use of a junction between two dissimilar conductors. For thermoelectric generation (Seebeck effect), a potential difference is set up across a junction when it is heated. In thermoelectric refrigeration (Peltier effect), cooling of a junction results when a current is passed through it. A current passing down a wire or through a junction represents the transport of both charge (electric current) and kinetic energy (heat current). Lattice-thermal conduction provides a mechanism for curtailing the temperature difference established by the flow of kinetic energy. The presence of *p*- and *n*-type materials on either side of a *p–n* junction provides high-mobility carriers of opposite sign, which act to enhance thermoelectric generation. The *p–n* junctions in the semiconductors lead telluride (PbTe) and bismuth telluride (Bi_2Te_3) have been found to be among the best materials for thermoelectric refrigeration. Although these devices have not replaced mechanical refrigerators for economic reasons, they can be effectively applied in situations that exploit their small size and weight, such as outer-space travel. (Ed.)

MAGNETISM IN SOLIDS

Few subjects in science are more difficult to understand than magnetism. A magnetic field is invisible, it can be detected only by its effect on moving charges or on other magnetic fields, and the state of magnetization of a particular piece of material is not easy to measure. Yet magnetism is important technologically for such devices as motors, relay armatures, electrical transformers, and speakers for communications systems, and many natural phenomena are a result of magnetism, such as sunspots, the Van Allen radiation belts, and the Northern Lights (see EARTH: *The Earth's main magnetic field*).

Observing the magnetic field

One of the obstacles to understanding magnetic solids is the difficulty of making measurements on them. Since lines of magnetic force cannot be seen and in fact are simply a convenient way of thinking about magnetic interaction, indirect methods of observing magnetic fields must be used; for example, the torsional, or twisting, effect on a permanent magnet. A magnetic field is produced by a coil of wire when a current flows through the coil. If the coil has 50 turns per metre and carries a current of one ampere, the magnetic field at the centre of the coil is said to be 50 ampere-turns per metre. This is about the average size of the Earth's magnetic field and is rather small. If an iron rod is inserted inside the coil, though, a much higher field results, perhaps 10,000 times as much. Such a field, 500,000 ampere-turns per metre, is large enough to be extremely useful. A field of this size is large enough to drive an electric motor, to activate an electric-power generator, or to power a loudspeaker.

Paramagnetism and diamagnetism. Not all materials inserted into the coil would increase the magnetic field by such a large amount. The paramagnetic solids would increase the field only slightly. Such solids include aluminum, sodium, manganese, chromium, magnesium, potassium, titanium, and many others. Another large group would actually decrease the strength of the field but again only slightly; they are called diamagnetic. They include copper, zinc, germanium, silver, gold, and bismuth. Neither the paramagnetic nor the diamagnetic materials are technically important, but it is instructive to examine the reason for their behaviour.

Paramagnetism occurs in matter because small magnetic dipoles on the individual atom are slightly aligned by a magnetic field. (A magnetic dipole of an atom can be considered to be a tiny permanent magnet.) The effect is small; the increase in magnetic field produced by paramagnetic materials is commonly only about one part in 100,000, at the greatest.

Diamagnetism is caused by a change in the orbital motion of electrons that occurs upon application of a magnetic field. The change causes formation of induced dipoles that are aligned opposite to the applied field; hence, they reduce the overall strength of the field. All materials possess some diamagnetic characteristics, even those that are paramagnetic or ferromagnetic; in these last cases, the positive effects overwhelm the negative diamagnetic features. Even at its strongest, the diamagnetism of most materials is very weak; for copper, silver, and gold

the influence is only a few millionths of the external field.

Ferromagnetism. The strong magnetic effect that is called ferromagnetism is by far the most interesting. It exists in a number of metals and alloys, but by far the most important are the elements iron, cobalt, nickel, and their alloys. The first magnetic materials known to man were not these pure metals and alloys, however, but lodestone, a type of iron oxide. This material, which occurs in nature in great quantity, was known for its magnetic qualities at least as early as 600 BC. Modern alloys of this same kind, called ferrites, are exceedingly important.

Electrons and magnetism

Just as the conduction of electric current is a property of the electrons (*i.e.,* their charge), so is magnetism a fundamental property of the electrons. Each electron behaves like a small permanent magnet of a certain size, a fundamental unit given the name Bohr magneton. This tiny dipole is an inherent characteristic of the electron, and no magnetic field is required to produce it. Whereas the electrical conductivity of solids is caused by the motion of the free electrons, the electrons important to ferromagnetism in solids are the ion-core electrons. In fact, not all of them contribute because they partly cancel one another. The ferromagnetic materials (*e.g.,* iron, cobalt, and nickel) are those for which the cancellation is not complete and the effects of some of the inherent dipoles add. Furthermore, dipoles interact so as to align themselves parallel to each other in certain directions, even in the absence of a magnetic field. Because different materials have different numbers of uncompensated dipoles, the strength of magnetic materials differs. For example, cobalt and iron have roughly the same strength, but nickel only about one-third as much.

Notions about the origins and details of magnetism developed gradually, culminating about 1910 in a theory by the French physicist Pierre-Ernest Weiss. His theory of magnetization and demagnetization utilizes the concept of domains, small regions in which dipoles on adjacent ions spontaneously align themselves parallel. Direct observation of domains required another two decades of work, but they were observed in the 1930s by Francis Bitter, an American physicist. By now, a great deal is known about the structure of domains and the details of the magnetization process.

Ferromagnetism disappears above a certain temperature, called the Curie temperature. The self-alignment capability of the dipoles gradually disappears as the temperature goes up, because the increasing energy of the thermal vibrations misaligns the dipoles. Finally, the dipoles lose their self-aligning capability completely, and the solid is no longer ferromagnetic. The Curie temperature of nickel is 358° C (676° F), of iron, 770° C (1,418° F), and of cobalt, 1,131° C (2,068° F). Alloys of these elements have intermediate Curie temperature in general; obviously, a ferromagnetic material must have a Curie temperature higher than any temperature at which it is actually used.

Soft magnetic materials. Soft ferromagnetic materials become demagnetized easily when the external field is removed. Hard magnetic materials become magnetized and demagnetized with great difficulty. The most widely used soft magnetic material is transformer steel, an alloy of iron with a few percent of silicon. Another group consists of iron–nickel alloys, one of 50–50, the other around 20–80, both being called Permalloys. A third alloy, with certain characteristics superior to either of these groups, has a composition of around 78 percent nickel, 18 percent iron, and 4 percent molybdenum; appropriately, it is called Supermalloy.

The ferrites

Magnetic materials made from metals become useless at high frequencies, because energy losses caused by eddy currents are much too high to be tolerated. The problem is solved through use of magnetic insulators, principally oxides of iron. One important group of these oxides, known as ferrites, is based on the iron oxide compound Fe_3O_4. This is the naturally occurring magnetic material that constitutes lodestone.

An interesting and promising material is a special group of magnetic iron oxides that exhibit the garnet structure. A typical example is yttrium iron oxide, $Y_3Fe_5O_{12}$. Tiny cylindrical domains can be formed across thin strips of this material, the direction of magnetization in a domain being opposite to that of the material surrounding it. These little cylinders behave much like tiny magnetic "bubbles"; they are stable in size and can be moved about very rapidly by a small magnetic field. The bubbles are so small—approximately 0.001 millimetre (0.00004 inch) in diameter—that their density in a sheet can be extremely high, perhaps 154,000 per square centimetre (1,000,000 per square inch). They can also be made to step from site to site hundreds of thousands of times per second with expenditure of very small energy. They are now being employed in high-density memory units for certain computers, as for example microcomputer systems designed for use in machine shops.

Hard magnetic materials. Striking as are the advances that have been made in soft magnetic materials, advances in the hard magnetic materials for permanent magnets are just as startling. Three periods characterize the development of permanent magnets in the last century. The first, from about 1880 to 1930, saw a development of special steels; the magneto of the early automobiles and the old telephone receiver got their size and shape from the quality of these steels. In the second period, from 1930 to 1960, the Alnicos (*i.e.,* alloys of iron with copper, titanium, aluminum, nickel, and cobalt) were supreme, yielding smaller, stronger magnets, though so brittle and hard that they could not be fashioned by machine. Finally, after 1960, the ceramic magnets and alloys of the rare earths have been developed. The successes achieved in these three periods depend on a particular characteristic of magnetic materials.

The word permanent applied to magnets means "not easily demagnetized." This characteristic is measured by the size of field necessary to reduce the magnetization of a permanent magnet to zero, called the coercive field. Good permanent magnets must have coercive fields of 50,000 ampere-turns per metre or more. Such quality was first achieved by use of physically hard tool steels (a typical alloy had 71 percent iron, 17 percent molybdenum, 12 percent cobalt), which produced large magnetic fields because they contained substantial amounts of iron and cobalt. They also contained many finely dispersed carbides that impede motion of the domain walls. Accordingly, they were not easily demagnetized; domain walls separating domains of opposite polarity could not easily move through the material.

The permanent-magnet tool steels were replaced by materials whose high coercive force comes from the actual shape of magnetized regions in the alloys. The alloy Alnico is typical. It consists of extremely fine needlelike particles of magnetic material imbedded in a nonmagnetic binder. When solidified and heat-treated in a magnetic field, the long particles tend to line up with the field, yielding a strong magnet. If the particles are fine enough, domain walls cannot form in them, and the individual particles cannot be readily demagnetized.

Small-particle theory in the 1950s led to attempts to produce permanent magnets by dispersing fine suspensions of iron in plastic or a metallic binder. Several products were made and sold, but manufacturing problems are difficult, and these magnets are not of great commercial importance.

Anisotropy

The third phase of permanent-magnet development has used a property known as crystalline anisotropy. Simply put, certain directions in a crystal are magnetically preferred over others. For iron, the cube edge is preferred; for nickel, the cube diagonal; and, for cobalt, the axis normal to the hexagonal plane. In switching the direction of magnetization from one preferred direction to another, parts of the crystal will be at one time or another in a higher energy state—hence, a large coercive force is required to change the direction of magnetization.

Attempts have been made for many years to find alloys that would utilize this strong directional preference. Until recently, they were largely unsuccessful however. Work on ferrites, though, led to discovery of a hexagonal oxide of iron, $BaFe_{12}O_{19}$, that has high coercive force. Like the cubic ferrites based on the Fe_3O_4 structure, however, it does not produce the highest magnetic fields, since much

of its volume is taken up by nonmagnetic ions. Even so, many ceramic magnets are made from this material.

Efforts have been undertaken to develop a metallic alloy with hexagonal structure, large anisotropy, and large density of a high-moment ion. Thus, it would be a strong magnet to give high field and a hard magnet to resist demagnetization. Such a group of materials has been found, a hexagonal compound of cobalt with some of the rare-earth metals. A compound of cobalt with samarium, Co_5Sm, makes a permanent magnet more than 10 times as strong as the best Alnico.

Interaction of radiation with matter

The electromagnetic spectrum covers an immense range of frequencies. That part most commonly used, visible light, is but a tiny part of the known range. Even so, this discussion is concentrated on the interaction of the visible spectrum with matter (see also the articles RADIATION and ELECTROMAGNETIC RADIATION).

Electromagnetic waves are pulsations of electric and magnetic fields. In a vacuum they travel with a velocity of 300,000 kilometres per second (186,000 miles per second). Their frequency, ν, and wavelength, λ, are related to their velocity, c, by a simple equation in which the velocity equals the product of wavelength and frequency: $\nu\lambda = c$. A stationary observer who could detect the electric field of the waves as it passed by would find that its strength, E, would have some maximum amplitude, A, and would vary with time in a very regular fashion. Furthermore, the electric field is maximum at right angles to the direction of motion of the wave.

The energy of electromagnetic waves is quantized; that is, the energy comes in discrete, well-defined units called photons, the size of which is given by the product, $h\nu$, of a universal number called Planck's constant, h, and the frequency, ν, of the wave. Strong radiation fields contain more photons than weak ones, but all photons of a given frequency have the same energy.

Essential features of electromagnetic waves

Thus, electromagnetic waves have four essential features: (1) they travel with a certain speed, 300,000 kilometres per second in vacuum but less in solids, liquids, or gases; (2) each wave has a particular frequency that is determined by the manner in which it is excited; (3) the wave consists of electric and magnetic fields at right angles to the direction of propagation, and these fields vary sinusoidally (*i.e.*, their values when charted follow a sine curve) with time as observed by a stationary observer; (4) the energy of the wave is packaged into individual units, called photons, whose energy has magnitude $h\nu$.

For all the value of materials as building materials, for all their value in electrical and magnetic uses, nothing about materials is more important to mankind than this relationship to electromagnetic radiation. The infrared in the Sun's rays provides warmth; more detailed information is provided by sight than by any other sense; food and clothing come ultimately from chemical interactions of plants with radiation. The sections that follow deal with the way solids behave when they are illuminated with radiation, the changes in the internal structure induced by the energy of incoming waves, and the way radiation is generated in the materials.

BEHAVIOUR OF SOLIDS ILLUMINATED WITH RADIATION

Most materials do not emit light of their own; mostly, they simply reflect light, transmit it, or absorb it. How matter behaves when illuminated depends on what the electrons do. If the light photon has a frequency such that its energy matches the energy needed to excite an electron to a higher state, then the photon may be absorbed. The electron, which may be an ion-core electron or one of the free electrons in a solid, may return to its initial state from this higher energy excited state by re-emission of the photon of the same frequency ν. Detection of this photon permits viewing of an object by reflected (that is, by re-emitted) light. The process of absorption and re-emission of photons is very selective. If the energy of the incoming photon does not closely match the required excitation energy, no excitation occurs. The sharpness of absorption lines for gas atoms can be noted by observation of the dark lines observed in spectra of nebulae and stars, including the Sun. Radiation falling on solids and liquids must meet this same requirement of energy matching, called resonance, but solids have such an abundance of frequencies of excitation possible that excitations may exist in broad ranges of frequency.

Transparency of ice

Ice is transparent to visible light because no excitations exist whose energy matches photons in the visible part of the spectrum. Excitation levels do exist in the ultraviolet, so ice is opaque to ultraviolet. Radiation in this part of the spectrum has wavelengths shorter than 4,000 angstrom units; *i.e.*, frequencies greater than 10^{15} hertz (or cycles per second). Photons of this frequency have energies of about 1.6×10^{-19} calorie; *i.e.*, about four electron volts. This is the energy required to free a valence electron in the ice crystal. Ice (liquid water as well) also absorbs photons with frequencies in the infrared; were it not so, all lakes and rivers might be frozen permanently.

The warming of ice and water by infrared radiation shows another feature of excitation of electrons by photons. Energy of excitation in solids and liquids need not be re-emitted as the same type of photons—the energy may be transferred to thermal vibrations (that is, into heat), or it may produce other physical or chemical changes in matter.

Transparency of insulators

Insulators such as glass, quartz, diamond, and rocksalt crystals are transparent to visible light for the same reason as ice. No excitation resonances exist in the visible spectrum, because the valence electrons are so tightly bound that photons with energy in the ultraviolet are required to free them. Addition of impurities with resonances in the visible range, however, reduces the transparency and may give the crystal a colour. Crystalline aluminum oxide, corundum, is transparent to visible light; addition of a little chromium gives it a red colour to produce the gemstone ruby; addition of titanium and iron produces the blue sapphire.

Metals are opaque to visible light; they also reflect light with little loss. This reflection comes entirely from the free electrons. These electrons are able to move so freely that no electric field may be propagated in the solid, and light penetrates only a few hundred angstroms. The motion of the electrons, in fact, causes the light to be re-emitted; indeed, if the surface of the metal is smooth, the re-emitted radiation from all parts of the surface is in phase, and the surface acts like a mirror.

Metals are excellent reflectors of radiation not only in the visible part of the spectrum but also for radiation of longer wavelength, far into the infrared ($\lambda = 1$ to 500 micrometres). Gradually, the depth of penetration of the wave increases, and transformation of the incident energy to heat increases. Power lines conducting electric current alternating at 60 cycles per second to heat a lamp filament are supplying radiation of wavelength 5,000,000 metres to the metal. All of this energy is absorbed in the filament; none is reflected (neglecting inductive effects of the tiny coil). Waves with even lower frequency have been detected in nature by geophysicists using sensitive magnetometers. These waves (of wavelength up to 10,000,000,000 metres) are thought to penetrate deeply into the Earth—far down into the molten metallic core.

Optical properties of semiconductors lie between those of metals and those of insulators. They are commonly transparent at radio frequencies and down into the infrared. In the visible region they are opaque like metals; incident photons are absorbed by the excitation of valence electrons into higher energy states. For germanium the crossover wavelength is about 15,000 angstroms, for silicon about 12,000.

The most useful radiations in the other direction (*i.e.*, to shorter wavelengths) occur in the X-ray region (from one to 1,000 angstroms). These waves have energy sufficiently large to excite the ion-core electrons in all materials, whether metals, semiconductors, or insulators. Thus, discrete line absorption occurs at well-defined frequencies. Furthermore, at wavelengths in the range of one to five angstroms, diffraction effects occur, permitting crystal-structure determination to be made (see above *Crystalline solids*).

GENERATION OF ELECTROMAGNETIC RADIATION

The immediate re-emission of incident radiation associated with reflection or transmission of incident photons is not really generation of radiation. Solids (and liquids and gases), however, can be self-luminous; they may generate photons from energy that is supplied to the solid in some other form. The photons emitted may have a sharp, well-defined frequency, or they may cover a rather broad range of frequencies.

Conversion of energy into photons

Energy can be converted into photons in a variety of ways. The most important is the conversion of heat into light—thermoluminescence. Another way, electron bombardment, is important in luminescence of both solids and gases. Fission and fusion reactions produce high-energy photons as well as ions and elementary particles of high kinetic energy. Some solids have the capability of converting photons of incoming radiation into photons of another frequency, photoluminescence. Others produce light by mechanical means, by rupture of materials, a phenomenon called triboluminescence; still others convert chemical energy into light, chemiluminescence; finally, some solids convert electrical energy into photons directly, electroluminescence.

The variety of ways in which these phenomena can occur is vast, almost beyond counting if one considers that they can occur in combination. Therefore, the discussion below is limited to a few examples that are important because of practical value or because of insight they give into basic features of the emission processes.

Monochromatic radiation. Photons of a single frequency—*i.e.,* monochromatic radiation—are produced when either atoms or molecules in some excited state revert to a lower state. These photons may be of high energy, as for example X-rays or gamma rays, if the states are of widely different energy; that is, different by more than 1,000 electron volts. Photons are in the visible spectrum if the energy difference is from 1.5 to three electron volts. The radiation is in the infrared or even radio wavelengths if the transitions involve energy differences of fractions of an electron volt.

Excitation of electrons to the higher states can be provided by several means. An electric discharge—a spark—can eject electrons from low levels into higher levels in the atom. A continuous electric arc, as in a neon tube or a sodium-vapour lamp, may produce excited states. Incident photons or high-energy electrons may provide energy for excitation.

Monochromatic light from isolated atoms (as in a gas) is of immense value. By its use the electronic structure of atoms has been deduced and many practical light sources have been developed: neon, sodium-vapour and mercury-vapour lamps, fluorescent lamps, and X-ray tubes. An additional important feature, laser action, is also being exploited for various purposes.

Laser emission

Laser emission results from a special type of interaction of light with atoms. Not only can radiation be absorbed if the photon energy matches an excitation energy of an atom, but an additional photon (of the same energy) can be emitted if the atom is already in the excited state. This emission process, called stimulated emission, results in amplification of light, in contrast to diminution caused by absorption. Amplification will predominate if more atoms are in the excited state than in the lower state and, in fact, sufficiently more to compensate for loss of photons that escape without participating in the process. Many materials are suitable for holding the excited atoms: gases (carbon dioxide is an example), liquids (such as organic molecules in alcohol), or solids (ruby or glass containing atoms of the rare-earth metal neodymium).

Band of radiation. Every photon has a certain frequency, but photons within a band of frequencies will be emitted if the allowed electron transitions form a continuous range. Much of the radiation suitable for practical application is of this type; three examples are described in the following paragraphs.

1. Excitation of free electrons in metals by varying electric fields can be used to excite radio waves. By means of the electric field generated by an oscillator, the free electrons in an antenna are accelerated and decelerated. These electrons oscillating in phase in the antenna radiate photons whose frequency is that of the oscillator. The electrons themselves do not possess an inherent resonant frequency; rather, the frequency is set by the oscillator and the antenna. Enormous amounts of power, hundreds of thousands of watts (corresponding to myriads of individual photons), can be radiated in the radio and television frequencies.

2. Radiation from solids by thermally induced motion of atoms is by far the most important scheme of energy conversion. The allowed frequencies of vibration of atoms and groups of atoms in solids form a continuous band; hence, the radiation itself covers a wide range of frequencies. The greatest amount of light comes in the infrared and the visible parts of the electromagnetic spectrum; thus, hot objects can be used as sources of both heat and visible light.

Metals, semiconductors, and insulators, both solid and liquid, produce such radiation. The efficiency of production of light depends partly on the temperature of the material; the maximum energy in the spectrum comes at shorter wavelengths as the temperature of the solid is increased. Thus, a heat lamp may operate at 800° C (1,500° F), an incandescent lamp at 2,500° C (4,500° F). The energy radiated in the visible part of the spectrum from a lamp filament—*e.g.,* a 100-watt lamp—is only a few percent of the electrical energy supplied to the lamp. Even so, a 100-watt lamp filament radiates about 10^{19} photons per second in the visible range.

3. The abrupt deceleration of high-energy electrons when they strike solid matter also produces radiation in a broad band. This effect is extensively employed to produce X-rays. With operating voltages of from 10,000 to 50,000 volts, X-rays of wavelengths close to one angstrom are generated when the incident electrons are slowed by collision with lattice ions of the target. Of course, the incident electrons may also cause generation of line spectra characteristic of the target ions, most typically tungsten or molybdenum. The overall emission spectrum therefore consists of sharp line spectra superimposed on broad-band (so-called white) radiation.

Production of X-rays

Photoluminescence. Materials may emit radiation of one wavelength when they are irradiated with photons of another wavelength. This event, known as photoluminescence, is most widely used to produce visible light, but it also occurs in the X-ray region and in other regions of the electromagnetic spectrum. Radiation emitted immediately upon absorption of the incident light is called fluorescent radiation. Such radiation in the visible spectrum is frequently produced when ultraviolet light, often the characteristic lines of mercury, is absorbed either by an inorganic compound (in many instances a metal oxide, a phosphide, or a sulfide) or by an organic material such as motor oil. The absorption of the incident photons is accompanied by excitation of valence electrons into some excited state. Recovery to the original state does not occur by a single-step process involving re-emission of a photon of the original wavelength, but rather the excited state decays in two or more steps to the original state. As a consequence, the re-emitted photons are of less energy (that is, of longer wavelength) than the original photons.

Practical use of fluorescence in lighting uses the mercury lines excited by electrons streaming through mercury vapour. These ultraviolet lines are absorbed by a mixture of compounds (called phosphors) coated on the inside of the glass envelope. These phosphors, blended to give the desired colour, produce visible light.

Visible light can also be produced by these same phosphors if the incident "radiation" is electrons. This process—cathodoluminescence—is widely used in cathode-ray tubes. Television display tubes and their predecessor—tubes for cathode-ray oscilloscopes—use this mode of light excitation.

Many other types of interactions of radiation with matter exist, some of which have practical application while others have only scientific interest; for details of these, see MEASUREMENT AND OBSERVATION.

(C.A.W./H.K.B.)

LIQUID STATE

Clear distinctions between the liquids, gaseous, and solid phases hold only for those substances whose molecules are composed of a small number of atoms. When the number exceeds about 20, the liquid may often be cooled below the true melting point to form a glass (see *Amorphous solids*), which has many of the mechanical properties of a solid but lacks crystalline order. If the number of atoms in the molecule exceeds about 100–200, the classification into solids, liquids, and gases ceases to be useful. At low temperatures such substances are usually glasses or amorphous solids, and their rigidity falls with increasing temperature—*i.e.*, they do not have fixed melting points; some may, however, form true liquids. With these large molecules, the gaseous state is not attainable because they decompose chemically before the temperature is high enough for the liquid to evaporate. Synthetic and natural high polymers (*e.g.*, nylon and rubber) behave in this way.

If the molecules are large, rigid, and either roughly planar or linear, as in cholesteryl acetate or *p*-azoxyanisole, the solid may melt to an anisotropic liquid (not uniform in all directions) in which the molecules are free to move about but have great difficulty in rotating. Such a state is called liquid crystal, and the anisotropy produces changes of the refractive index (a measure of the change in direction of light when it passes from one medium into another) with the direction of the incident light and hence leads to unusual optical effects; however, no inorganic compounds and fewer than one in 200 of the known organic compounds form liquid crystals. The theory of normal liquids is, therefore, predominantly the theory of the behaviour of substances consisting of simple molecules.

Thus, a liquid lacks both the strong spatial order of a solid, though it has the high density of solids, and the absence of order of a gas that results from the low density of gases; *i.e.*, gas molecules are relatively free of each other's influence. The combination of high density and of partial order in liquids has led to difficulties in developing quantitatively acceptable theories of liquids. Understanding of the liquid state, as of all states of matter, came with the kinetic molecular theory, which stated that matter consisted of particles in constant motion and that this motion was the manifestation of thermal energy. The greater the thermal energy of the particle, the faster it moved.

In very general terms, the particles may be molecules, atoms, ions, or electrons. In a gas, the particles are far enough from one another and are moving fast enough to escape each other's influence, which may be of various kinds, such as attraction or repulsion due to electrical charges, and specific forces of attraction that involve the electrons orbiting around atomic nuclei. The motion of particles is in a straight line, and the collisions that result occur with no loss of energy although an exchange of energies may result between colliding particles. When a gas is cooled, its particles move more slowly, and those slow enough to linger in each other's vicinity will coalesce, because a force of attraction will overcome their lowered kinetic energy and, by definition, thermal energy. Each particle, when it joins the liquid state with others, gives up a measure of heat called the latent heat of liquefaction, but each continues to move at the same speed within the liquid as long as the temperature remains at the condensation point. The distances the particles can travel in a liquid without colliding are of the order of molecular diameters. As the liquid is cooled, the particles move more slowly still, until at the freezing temperature the attractive energy produces so high a density that the liquid freezes into the solid state. They continue to vibrate, however, at the same speed, as long as the temperature remains at the freezing point, and the freezing process releases their latent heat of fusion. Heating a solid provides the particles with the heat of fusion necessary to escape one another's influence enough to move about in the liquid state, and further heating provides the liquid particles with their heat of evaporation, which enables them to escape one another altogether and form the vapour of the liquid, or the gaseous state. This starkly simplified view of the states of matter ignores many complicating factors, most important being the fact that no two particles need be moving at the same speed in a gas or liquid or solid and the related fact that even in a solid some particles may have acquired all the energy they need to be gas particles, while even in a gas some particles may be practically motionless for a brief time. It is the average kinetic energy of the particles that has to be considered, together with the fact that the motion is random. At the interface between liquid and gas and between liquid and solid, an exchange of particles is always taking place, slow gas molecules condensing at the liquid surface and fast liquid molecules escaping into the gas. An equilibrium state is reached in any closed system, so that the number of exchanges in either direction is the same. Because the kinetic energy of particles in the liquid state can be defined only in statistical terms (*i.e.*, every possible value can be found), discussion of the liquid state, as of the gaseous state, involves formulations in terms of probability functions.

Transition to the liquid state

The nature and structure of liquids at equilibrium

BEHAVIOUR OF LIQUIDS AT EQUILIBRIUM

The number of phases present in a system at equilibrium is, as explained earlier, governed by the phase rule derived by J. W. Gibbs. All the properties of a homogeneous phase of a single substance (*e.g.*, a pure liquid) are determined if, in the absence of external forces, the two variables of pressure and temperature are specified. Such a system is thus said to have two degrees of freedom. If there is more than one phase (*e.g.*, ice and water) or more than one substance present (*e.g.*, water and alcohol) or both, the number of degrees of freedom (F) is related to the number of phases (P) and the number of substances or components (C), by Gibbs phase rule, which states that the sum of the number of phases and degrees of freedom equals the sum of the number of components and the number two. This is written $P + F = C + 2$. Hence, for a single substance (when $C = 1$), the number of coexisting phases cannot exceed three, since F cannot be negative, though it can be zero. Solid, liquid, and gas can exist together in equilibrium only for one fixed pair of values of pressure and temperature, and at this point $F = 0$, since, being fixed, there is no

Degree of freedom

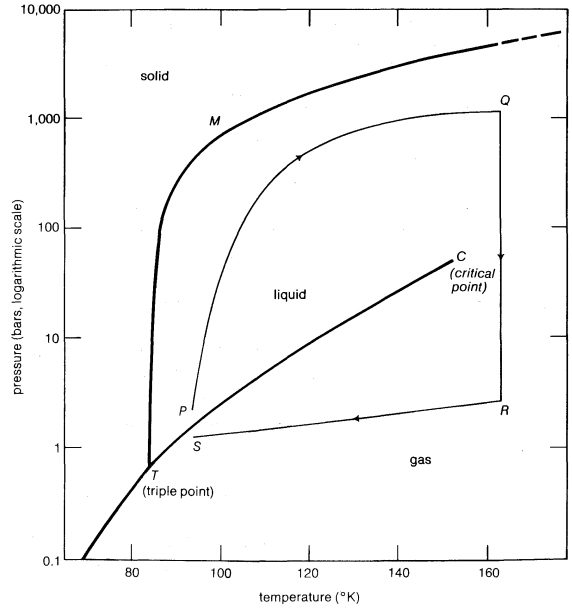


Figure 46: *Phase diagram of argon.*
Heavy lines show the boundaries between liquid and gas (*TC*) and between liquid and solid (*TM*). Line *PQRS* describes path from typically liquid state (*P*) to typically gaseous state (*S*) (see text).

freedom to choose the values of these variables. With two phases—solid and liquid or liquid and gas—the equation becomes $2 + F = 1 + 2$; hence, $F = 1$; *i.e.*, the system has one degree of freedom: either the pressure or the temperature can be chosen at will (within certain ranges), and the other is fixed by the nature of the substance.

These relations are summarized in Figure 46, the phase diagram of argon, in which one phase is shown as an area, two as a line, and three as the meeting of the lines at the triple point, *T*. The diagram shows that the liquid phase is merely a part of a single fluid state. There are abrupt changes in physical properties (*e.g.*, density, heat capacity, viscosity, dielectric constant) on crossing the lines that separate the liquid from the solid on one side and from the gas on the other. The second line, the vapour-pressure curve, *TC* (obtained by measuring the pressure at every temperature), does not continue to arbitrarily high temperatures but ends at the temperature and pressure that describe the critical point, *C*. At the critical point, phases in equilibrium become identical and form one phase. A liquid at the pressure and temperature indicated at point *P* can be heated without changing its density along the line *PQ;* at *Q*, the pressure is lowered, while the temperature is kept constant, and the volume of the liquid is expanded along *QR;* at *R*, the gas is cooled along *RS*, to *S*, where it is changed from a liquid to a gas without undergoing that abrupt change of properties associated with boiling or evaporation. The path *PQRS* demonstrates the essential continuity of state between liquid and gas, which differ in degree but which comprise together the one fluid state. Strictly speaking, the term liquid should be applied only to the more dense of the two phases on the line *TC*, but it is generally extended to any dense fluid state at low temperatures; that is, to the area lying within the angle *CTM*.

The line *TM* is the melting curve and is initially much steeper than the vapour-pressure curve; hence, as the pressure is changed, the temperature does not change much, and the melting temperature is little affected by pressure. No substance has been found to have a critical point on this line, and there are theoretical reasons for supposing that it continues indefinitely to high temperatures and pressures, until the substance is so compressed that the molecules break up into atoms, ions, and electrons. At pressures above 10^6 bars (one bar is roughly equal to the pressure of the atmosphere at sea level), it is believed that most substances pass into a metallic state.

There are a few substances, such as water and bismuth, for which the density of the solid is less than that of the liquid at the triple point. For these two substances, the volume change for solid to liquid is negative; hence, the melting temperatures fall with increasing pressure. With water, this negative volume change (*i.e.*, shrinking) persists to 2.1 kilobars and −22° C (−8° F), at which point the normal form of ice changes to a denser form, and, thereafter, the change in volume is positive.

It is possible to cool a gas at constant pressure to a temperature lower than that of the vapour-pressure line without producing immediate condensation, since the liquid phase forms readily only in the presence of suitable nuclei (*e.g.*, dust particles or ions) about which the drops can grow. Unless the gas is scrupulously cleaned, such nuclei remain; a supercooled vapour state is unstable, and liquid will ultimately condense. It is similarly possible to superheat a liquid to a temperature where, though still a liquid, the gas is the stable phase. Again, this occurs most readily with clean liquids heated in smooth vessels, because gas bubbles will collect around foreign particles or sharp points. When the superheated liquid changes to gas, it does so with almost explosive violence. A liquid also may be supercooled to below its freezing temperature.

Triple-point temperature

Triple-point temperatures range from 14 K (0 K equals −273.15° C [−459.67° F]) for hydrogen to temperatures too high for accurate measurement. Triple-point pressures are generally low, that of carbon dioxide at 5.2 bars being one of the highest. Most are around 10^{-3} bar, and those of some hydrocarbons are as low as 10^{-7} bar. The normal melting point of a substance is defined as the melting temperature at a pressure of one atmosphere (exactly equivalent to 1.01325 bars); it differs little from the triple point because of the steepness of melting lines (*TM* in Figure 46). Critical temperatures (maximum temperature at which a gas can be liquefied by pressure) range from helium, at 5.2 K, to temperatures too high to measure. Critical pressures (at which liquid and vapour phases are identical) are generally about 40–100 bars. The normal boiling point is the temperature at which the vapour pressure reaches one atmosphere. The normal liquid range is defined as the temperature interval between the normal melting point and the normal boiling point, but such a restriction is artificial, the true liquid range being from triple point to critical point. Substances whose triple-point pressures are above atmospheric (*e.g.*, carbon dioxide) have no normal liquid range but sublime at atmospheric pressure.

Many of the properties of a liquid near its triple point are closer to those of the solid than to those of the gas. It has a high density (typically 0.5–1.5 grams per cubic centimetre), a high refractive index (a measure of the change in direction of light rays passing from one medium into another; for liquids the index varies from 1.3 to 1.8), a high heat capacity at constant pressure (two to four joules per gram per K, one joule being equal to 0.239 calorie), and a low compressibility ($0.5–1 \times 10^{-4}$ per bar). The compressibility falls to values characteristic of a solid (0.1×10^{-4} per bar, or less) as the pressure increases. A simple and widely used equation describes the change of specific volume with pressure. If $V(p)$ is the volume at pressure p, $V(0)$ is volume at zero pressure, and A and B are positive parameters (constants whose values may be arbitrarily assigned), then the difference in volume resulting from a change in pressure equals the product of A, the pressure, and the volume at zero pressure, divided by the sum of B and the pressure. This is written:

$$V(0) - V(p) = \frac{ApV(0)}{(B+p)}. \qquad (8)$$

The pressure B is close to the pressure at which the compressibility has fallen to half its initial value and is generally about 500 bars for liquids near their triple points. It falls rapidly with increasing temperature.

As a liquid is heated along its vapour-pressure curve, *TC*, its density falls and its compressibility rises. Conversely, the density of the saturated vapour in equilibrium with the liquid rises; *i.e.*, the number of gas molecules in a fixed space above the liquid increases. Liquid and gas states approach each other with increasing rapidity as the temperature approaches *C*, until at this point they become identical and have a density about one-third of that of the liquid at *T*. The change of saturated-gas density (ρ_g) and liquid density (ρ_l) can be expressed by a simple equation when the temperature is close to the critical. If ρ_c is the density at the critical temperature T_c, then the difference between densities equals the difference between temperatures raised to a factor called beta, β. This is written:

$$(\rho_l - \rho_c) = (\rho_c - \rho_g) = (T_c - T)^{\beta}, \qquad (9)$$

and β is about 0.34. The compressibility and the heat capacity of the gas at constant pressure (C_p) become infinite as T approaches T_c from above along the path of constant density. The infinite compressibility implies that the pressure no longer restrains local fluctuations of density. The fluctuations grow to such an extent that their size is comparable with the wavelength of light, which is therefore strongly scattered. Hence, at the critical point, a normally transparent liquid is almost opaque and usually dark brown in colour. The classical description of the critical point and the results of modern measurement do not agree in detail, but recent considerations of thermodynamic stability show that there are certain regularities in behaviour that are common to all substances.

Shape of drops

Between a liquid and its vapour there is a dividing surface that has a measurable tension; work must be done to increase the area of the surface at constant temperature. Hence, in the absence of gravity or during free fall, the equilibrium shape of a volume of liquid is one that has a minimum area; *i.e.*, a sphere. In the Earth's field this shape is found only for small drops, for which the gravitational forces, since they are proportional to the volume, are negligible compared with surface forces, which are pro-

portional to the area. The surface tension falls with rising temperature and vanishes at the critical point. There is a similar dividing surface between two immiscible liquids, but this usually has lower tension. It is believed that there is a tension also between a liquid and a solid, though it is not directly measurable because of the rigidity of the solid; it may be inferred, however, under certain assumptions, from the angle of contact between the liquid and the solid; *i.e.,* the angle at which the liquid's surface meets the solid. If this angle is zero, the liquid surface is parallel to the solid surface and is said to wet the solid completely.

MOLECULAR STRUCTURE OF LIQUIDS

A liquid lacks regular arrangement of its molecules; *i.e.,* it lacks the regular crystalline order of a solid. The molecular structure of a liquid, therefore, can be described and measured only on the basis of probability. Probability is computed by a mathematical operation that yields distribution functions; a distribution function measures the probable distribution of some property of molecules through the liquid. Distribution functions that describe liquid structure are formulated as follows.

Distribution functions for liquids

If the volume of a minute quantity of liquid that may be called an element of the liquid, symbolized by $d\tau$, is small compared with the volume of a molecule and is in a fixed position symbolized by τ, sometimes it will contain the centre of a molecule, and sometimes it will not. The probability that it will do so, in a liquid at equilibrium, is the product of the number of molecules in the total volume, symbolized by N, and of the ratio of $d\tau$ to the total volume, V. Written out, this equals $Nd\tau/V$. The ratio N/V is called the number density and is written as n. Thus, probability is the product $nd\tau$. If two elements of volume are considered, $d\tau_1$, or $d\tau_2$, and if they are separated by a distance symbolized as τ_{12}, which is large compared with the size of a molecule, the probability that each such element of volume contains the centre of a molecule is given by the product of the probabilities of independent occupancy; that is, by $nd\tau_1 \times nd\tau_2$, or $n^2 d\tau_1 d\tau_2$. If τ_{12}, the distance between τ_1 and τ_2, is comparable with the size of a molecule, the probabilities are not independent, and the joint probability can be written as the product of the probabilities of independent occupancy (of a particular state) and a function of τ_{12} denoted by $g^{(2)}(\tau_{12})$. This function is called the pair distribution function and is the simplest measure of the structure of a liquid.

If the pair distribution function equals one—that is, if $g^{(2)} = 1$—this value corresponds to no order or structure, since the positions of the molecules are mutually independent. Such a limit of no structure at all is reached in a gas in which the molecules are far apart; *i.e.,* in effect, a hypothetically perfect gas. In a more dense gas, where molecules do affect one another, it is found that a useful approximation to $g^{(2)}$ is the value called the Boltzmann factor of the energy of interaction of a pair of molecules, symbolized as $u(r)$. The energy can be found from the properties of a dilute gas, from calculations using quantum mechanics, and in some cases from molecular-beam-scattering experiments. For the noble gas argon, which is highly unreactive, $u(r)$ is large and positive (*i.e.,* the molecules repel each other) if r is less than 3.3 angstroms, and $u(r)$ is negative (*i.e.,* the molecules attract one another) if the distance is greater than 3.3 angstroms; the value reduces to zero at large separations of the molecules.

More detailed mathematical calculations in terms of molecular forces are beyond the scope of this discussion.

In a liquid, the pair distribution function $g^{(2)}$ is a more complex function of r and depends on both the number density, n, and the absolute temperature, T. When r is less than the collision diameter (3.3 angstroms for argon), $g^{(2)}$ is essentially zero, since two molecules cannot "overlap." At slightly larger separations—when $g^{(2)}$ is greater than one—there is a more than average chance of finding a pair of molecules. This enhancement arises in part from the Boltzmann factor of $u(r)$ but, more importantly, from the geometric necessity of there being many pairs at such separations in a liquid whose density is little less than that of a solid. Minor maxima and minima at larger separations are the remnants in the liquid of the much stronger extrema found in the solid (for which the pair distribution function $g^{(2)}$ along a crystal axis consists of a set of many narrow peaks separated by regions in which it is zero). Thus, a liquid lacks the long-range order of a solid but shows appreciable order at short separations.

Importance of distribution function

A knowledge of the pair distribution function $g^{(2)}$ is the minimum information needed for the discussion of the thermodynamic properties of a liquid, since molecules interact predominantly in pairs, through the potential energy $u(r)$, and since $g^{(2)}$ describes the statistics of the distribution of such pairs. The macroscopic properties of internal or cohesive energy of a liquid, u, and of the departure of the pressure from its value in a perfect gas can be calculated from averages over $g^{(2)}$. These results show that a knowledge of $g^{(2)}$ is the link between the microscopic energies and forces and the macroscopic properties.

The theory of liquids at equilibrium reduces to a large degree to the measurement and calculation of the pair distribution function $g^{(2)}$. Finer details of the structure can be codified formally in higher distribution functions that specify the probabilities of there being molecules at three specified positions, but little is known about these functions.

PAIR DISTRIBUTION FUNCTION

There are two methods of measuring the pair distribution function $g(r)$ (from which the superscript may be omitted): first, by X-ray or neutron diffraction from simple fluids and, second, by computer simulation of the molecular structure and motions in a liquid. In the first, the liquid is exposed to a specific, single type (monochromatic) radiation and the observed results are then subjected to a mathematical treatment known as a Fourier transform.

The second method of obtaining the pair distribution function $g(r)$ supposes that the energy of interaction, $u(r)$, for the liquid under study is known. A computer model of a liquid is set up, in which between 100 and 1,000 molecules are contained within a cube. There are now two methods of proceeding, by Monte Carlo calculation or by what is called molecular dynamics; only the latter need be discussed here. Each molecule is assigned a random position and velocity, and Newton's equations of motion are solved to calculate the path of each molecule in the changing field of all the others. A molecule that leaves the cell is deemed to be replaced by a new one entering through the corresponding point on the opposite wall with equal velocity. After a few collisions per molecule, the distribution of velocities conforms with equations worked out by the British physicist James Clerk Maxwell, and after a longer time the mean positions are those appropriate to the density and mean kinetic energy (*i.e.,* temperature) of the liquid. Functions such as the pair distribution function $g(r)$ can now be evaluated by taking suitable averages as the system evolves in time. Since 1958 such computer experiments have added more to the knowledge of the molecular structure of simple liquids than all the theoretical work of the previous century.

Use of a computer model

The theory of liquids at equilibrium cannot be discussed without extensive mathematical analysis. It is, to a large degree, the theory of the calculation of the total correlation function, $h(r)$, for which $h(r)$ is $g(r) - 1$, in terms of the pair potential $u(r)$, which, for the purposes of this theory, is assumed to be known. Equations can be derived that can be interpreted by saying that the total function h is composed of a direct term, c, and of chains of direct terms through all other molecules in the liquid. The direct function c can be determined from X-ray-diffraction experiments and also from other calculations, including computer results based on the assumption that a liquid behaves to a first approximation as a hard-sphere fluid.

The structure of a simple liquid (*e.g.,* argon) is determined largely by the problem of packing at high density molecules with strong repulsive forces between them. This approach is not a complete theory of liquids due to its basis in computer experiments. In any event, molecules as polar as water, for example, cannot be included quantitatively in any general scheme—such liquids must be discussed individually. No simple theory can account for the fact that water has a maximum density at 4° C (39° F) and expands on cooling below that temperature.

Mechanical and thermal properties of liquids

TRANSPORT PROPERTIES

A liquid in which there exists a gradient of the concentration, temperature, or velocity is not at equilibrium and responds to these variations, which may be considered "forces," by an appropriate movement of a quantity of matter, energy, or momentum through a unit area, called a flux. Unless the gradients are extremely large (as in a shock wave), the flux is proportional to the force, and this relation is embodied in several laws of diffusion and conduction, for which the mathematical formulations are beyond the scope of this article.

Velocity is a vector property, and so the laws governing the transport of momentum and, hence, of viscous flow are more complicated than those of diffusion and conduction. The simplest case is the fluid that obeys Newton's laws, for which stress is proportional to rate of strain.

Diffusion

Diffusion coefficients are about 10^{-9} m^2/s near the triple point and rise with increasing temperature and decreasing density. The thermal conductivity of a liquid is about 10 times that of the gas. It generally rises with increasing temperature and appears to become infinitely large at the critical point in the same way as the heat capacity at constant volume. The thermal conductivities of liquid metals are unusually high, since most of the energy is carried by the free electrons. Hence, the thermal conductivity is closely related to the electrical conductivity.

The shear viscosity varies with the molecular complexity of the liquid; thus, in argon and liquid air it is a fraction of that in water, and in heavy oils it rises to hundreds of times that. The bulk viscosity was not detected in a monatomic liquid until 1965 and is discussed below in connection with the speed of sound.

The theory of transport properties (that is, of the relation between certain properties such as diffusion and viscosity and the mass, structure, and interaction potential of the molecules) is in a less developed state than that of the equilibrium properties. In a gas, the mechanism of transport is by convection on a molecular scale; that is, matter, energy, and momentum are carried by the actual movement of molecules. In a liquid, this is true also for diffusion, but the the major contribution to viscosity and thermal conductivity is through collisions. When a pair of molecules collide there is virtually instantaneous transfer of energy and momentum over a distance equal to the molecular diameter. In a gas, this distance is negligible compared with the mean free path of molecular travel, and so such transfer is unimportant; in a liquid with the molecules touching, it is dominant. The molecules spend much of their time moving in "cells" to which they are confined by their neighbours, but this restriction on their movement does not hinder transfer by collisions. Hence, diffusion is slow in a liquid, but the coefficients of viscosity and thermal conductivity are large compared with those of a gas. This simple division into transfer by convection and transfer by collision has been put into more quantitative form by making simplifying assumptions about the movement of molecules within their cells and about the activation energy needed to escape to a neighbouring cell. The resulting theory, however, although pictorially attractive, is not a satisfactory route to a theory of transport processes.

A proper theory of transport properties must proceed through the distribution functions. Since the liquid is not at equilibrium, these will be functions not only of the separations of the molecules but also of their velocities; moreover, they will change with time.

Some of the most useful results have come not from a direct attack on the central theoretical problem but from the use of theory to interrelate different properties of the liquid. One approach, started by Albert Einstein and continued by others, introduces a correlation between a function at time zero and its value at a later time, t.

Transport properties of dense fluids

A theory of the transport properties of dense fluids was developed by assuming (1) that the molecules were hard spheres, so that only binary collisions occur, and (2) that the velocities of two colliding molecules are uncorrelated before collision. Velocity correlation is neglected in the description of the fluid. This approach is satisfactory at moderate densities but less so as the true liquid density is reached. The cells in which molecules are effectively confined at high densities mean that two neighbouring molecules may well collide more than once before they diffuse apart. Hence, their velocities are not uncorrelated at the second and subsequent collisions. Attempts to remove this defect lead to mathematical difficulties that do not lend themselves to ready summary; these are now an active field of research.

PROPAGATION OF SOUND IN LIQUIDS

A sound wave is a series of longitudinal compressions and expansions that travels through a liquid at a speed of about one kilometre per second, or about three times the speed of sound in air. The compressions and expansions are adiabatic (changes that take place without transfer of heat) and reversible if the frequency is not too high. Conduction of energy from the hot (compressed) to the cold (expanded) regions of the liquid introduces irreversible effects, which are dissipative, and so such conduction leads to the absorption of the sound. A longitudinal compression (in the direction of the wave) is a combination of a uniform compression and a shearing stress (a force that causes one plane of a substance to glide past an adjacent plane). Hence, both bulk and shear viscosity also govern the propagation of sound in a liquid.

These dissipative effects are small in simple liquids, which may be virtually "transparent" to sound at frequencies up to 100 megahertz (10^8 cycles per second). At these high frequencies, that elusive coefficient, the bulk viscosity, can be observed. More complex liquids, such as hydrocarbons, esters, and even so simple a liquid as carbon disulfide, show strong absorption of sound below one megahertz. The absorption is accompanied by dispersion—that is, by the change of speed with frequency—and that can be described formally in terms of a frequency-dependent bulk viscosity. The cause is either a failure of the vibrational modes of the molecule to acquire and lose energy, if the time of heating and cooling is less than 10^{-6} of a second, or a structural change of the liquid on compression, which needs a similar time to occur. The first is certainly the cause in carbon disulfide, and the latter is exemplified by the association and dissociation of acetic acid into paired molecules and hence the strong absorption of sound of high frequency in this liquid. Hydrocarbons may combine both effects.

ELECTRICAL AND MAGNETIC PROPERTIES

If a liquid is placed in a static electric field, the field exerts a force on any free carriers of electric charge in the liquid, and the liquid, therefore, conducts electricity. Such carriers are of two kinds, mobile electrons and ions. The former are present in abundance in liquid metals, which have conductivities that are generally about one-third of the conductivity of the solid. The fall in conductivity on melting arises from the greater disorder of the positive ions in the liquid and hence their greater ability to scatter electrons. The contribution of the ions is small, less than 5 percent in most liquid metals, but is the sole cause of conductivity in molten salts (consisting of negative and positive ions) and in their aqueous solutions. Such conductivities vary widely but are much lower than those of liquid metals.

Non-ionic liquids

Non-ionic liquids (composed of molecules that do not dissociate into ions) have negligible conductivities, but they are polarized by an electric field; that is, the liquid develops positive and negative poles and also a dipole moment (pole strength times distance between poles) that is oriented against the field, from which the liquid therefore acquires energy. This polarization is of three kinds; in the first, called electron polarization, the electrons in each atom are displaced from their usual positions, thus giving each molecule a small dipole moment. The molecules act independently, and their reaction to the static electric field is the same as to the fluctuating electrical component of the field of electromagnetic radiation. The contribution of electron polarization to the dielectric constant (a measure of the electric stress) of the liquid is numerically equal

Atomic polarization

to the square root of its refractive index. A second effect of an electric field is to produce atomic polarization, in which there is a relative change in the mean positions of the atomic nuclei within one molecule. This generally small effect is observed at radio frequencies but not at optical, and so it is missing from the refractive index. The third effect is that of orientation polarization. If the molecules have permanent dipole moments, they are partially aligned by the field and contribute heavily to the polarization. Thus, the dielectric constant of a nonpolar liquid, such as a hydrocarbon, is about 2, that of a weakly polar liquid, such as chloroform or ethyl ether, about 5, while those of highly polar liquids, such as ethanol and water, lie between 25 and 80.

The partial alignment occurs in a static electrical field or in one whose frequency is less than the reciprocal of the time it takes the polar molecule to rotate. Hence, orientation polarization does not contribute to the refractive index. The degree of alignment is less than might be expected from independent dipoles because of the polar intermolecular forces. The full effects of the independent dipoles can be observed only in the gas phase or, in a reasonable measure, in a dilute solution of a polar substance in a nonpolar solvent.

Most liquids are diamagnetic. The exceptions are those whose molecules have unpaired electrons (*e.g.*, nitric oxide) and the liquid metals. Even those metals that are ferromagnetic (magnetic) in the solid state are only paramagnetic in the liquid state.

ABSORPTION AND SCATTERING OF RADIATION

Knowledge of the structure of liquids is derived principally from the study of their interaction with radiation. One example, the determination of $g(r)$ at equilibrium from X-ray or neutron diffraction, has been touched upon above under *Pair distribution function.* In the more general case of the scattering of radiation, the cause of the scattering is different for different kinds of radiation; thus, light is scattered by the outer electrons of an atom, X-rays by all the electrons, and neutrons by the nuclei. If the atoms in liquids move as entities, however, then all types of scattering can be described by the atomic or molecular distribution functions. (Jo.S.R.)

Solutions and solubilities

The ability of liquids to dissolve solids, other liquids, or gases has long been recognized as one of the fundamental phenomena of nature encountered in daily life. The practical importance of solutions and the need to understand their properties have challenged numerous writers since the Ionian philosophers and Aristotle. Though many physicists and chemists have devoted themselves to a study of solutions, this is today still an incompletely understood subject under active investigation.

A solution is a mixture of two or more chemically distinct substances, a mixture that is said to be homogeneous on the molecular scale—the composition at any one point in the mixture is the same as that at any other point. This is in contrast to a suspension (or slurry), in which small discontinuous particles are surrounded by a continuous fluid. Although the word solution is commonly applied to the liquid state of matter, solutions of solids and gases are also possible; brass, for example, is a solution of copper and zinc, and air is a solution primarily of oxygen and nitrogen with a few other gases present in relatively small amounts.

The ability of one substance to dissolve another depends always on the chemical nature of the substances, frequently on the temperature, and occasionally on the pressure. Water, for example, readily dissolves methyl alcohol but does not dissolve mercury; it barely dissolves benzene at room temperature but does so increasingly as the temperature rises. While the solubility in water of the gases present in air is extremely small at atmospheric pressure, it becomes appreciable at high pressures where, in many cases, the solubility of a gas is (approximately) proportional to its pressure. Thus, a diver must breathe air (four-fifths nitrogen) at a pressure corresponding to the pressure around him, and, as he goes deeper, more air dissolves in his blood. When he ascends, the solubility of the gases decreases so that they leave his blood, forming bubbles in the blood vessels. This condition (known as the bends) is extremely painful and may cause death; it can be alleviated by breathing, instead of air, a mixture of helium and oxygen because the solubility of helium in blood is much lower than that of nitrogen.

Effects of temperature and pressure

The solubility of one fluid in another may be complete or partial; thus, at room temperature, water and methyl alcohol mix in all proportions, but 100 grams of water dissolve only 0.07 gram of benzene. Though it is generally supposed that all gases are completely miscible—*i.e.*, mutually soluble in all proportions—this is true only at normal pressures (pressure is usually expressed in atmospheres, one atmosphere being the pressure exerted by a column of mercury 760 millimetres high, or about 14.7 pounds per square inch). At high pressures, pairs of chemically unlike gases may exhibit only limited miscibility; for example, at 20° C (68° F) helium and xenon are completely miscible at pressures below 200 atmospheres but become increasingly immiscible as the pressure rises.

The ability of a liquid to dissolve selectively forms the basis of common separation operations in chemical and related industries. A mixture of two gases, carbon dioxide and nitrogen, can be separated by bringing it into contact with ethanolamine, a liquid solvent that readily dissolves carbon dioxide but hardly dissolves nitrogen. In this process, called absorption, the dissolved carbon dioxide is later recovered, and the solvent made usable again by heating the carbon dioxide-rich solvent, since the solubility of a gas in a liquid usually (but not always) decreases with rising temperature. A similar absorption operation can remove a pollutant such as sulfur dioxide from smokestack gases in a plant using sulfur-containing coal or petroleum as fuel.

The process wherein a dissolved substance is transferred from one liquid to another is called extraction. As an example, phenolic pollutants (organic compounds of the types known as phenol, cresol, and resorcinol) are frequently found in industrial aqueous waste streams, and, since these phenolics are damaging to marine life, it is important to remove them before dumping the waste stream into a lake or river. One way to do so is to bring the waste stream into contact with a water-insoluble solvent (*e.g.*, an organic liquid such as a high-boiling hydrocarbon) that has a strong affinity for the phenolic pollutant. The solubility of the phenolic in the solvent divided by that in water is called the distribution coefficient, and it is clear that for an efficient extraction process it is desirable to have as large a distribution coefficient as possible.

CLASSES OF SOLUTIONS

Electrolytes and nonelectrolytes. Broadly speaking, liquid mixtures can be classified as either solutions of electrolytes or solutions of nonelectrolytes. Electrolytes are substances that can dissociate into electrically charged particles called ions, while nonelectrolytes consist of molecules that bear no net electric charge. Thus when ordinary salt (sodium chloride, formula NaCl) is dissolved in water, it forms an electrolytic solution, dissociating into positive sodium ions (Na^+) and negative chloride ions (Cl^-), whereas sugar dissolved in water maintains its molecular integrity and does not dissociate. Because of its omnipresence, water is the most common solvent for electrolytes; the ocean is a solution of electrolytes. Electrolyte solutions, however, are also formed by other solvents (such as ammonia and sulfur dioxide) that have a large dielectric constant (a measure of the ability of a fluid to decrease the forces of attraction and repulsion between charged particles). The energy required to separate an ion pair (*i.e.*, one ion of positive charge and one ion of negative charge) varies inversely with the dielectric constant, and, therefore, appreciable dissociation into separate ions occurs only in solvents with large dielectric constants.

Ion pairs

Most electrolytes (for example, salts) are nonvolatile, which means that they have essentially no tendency to enter the vapour phase. There are, however, some notable exceptions, such as hydrogen chloride (HCl), which

is readily soluble in water, where it forms hydrogen ions (H^+) and chloride ions (Cl^-). At normal temperature and pressure, pure hydrogen chloride is a gas, and, in the absence of water or some other ionizing solvent, hydrogen chloride exists in molecular, rather than ionic, form.

Solutions of electrolytes readily conduct electricity, whereas nonelectrolyte solutions do not. A dilute solution of hydrogen chloride in water is a good electrical conductor, but a dilute solution of hydrogen chloride in a hydrocarbon is a good insulator. Because of the large difference in dielectric constants, hydrogen chloride is ionized in water but not in hydrocarbons.

Weak electrolytes. While classification under electrolyte-solution or nonelectrolyte-solution headings is often useful, some solutions have properties near the boundary between these two broad classes. Although such substances as ordinary salt and hydrogen chloride are strong electrolytes—*i.e.,* they dissociate completely in an ionizing solvent—there are many substances, called weak electrolytes, that dissociate to only a small extent in ionizing solvents. For example, in aqueous solution, acetic acid can dissociate into a positive hydrogen ion (H^+) and a negative acetate ion (CH_3COO^-), but it does so to a very limited extent; in an aqueous solution containing 50 grams acetic acid and 1,000 grams water, less than 1 percent of the acetic acid molecules are dissociated into ions. Therefore, a solution of acetic acid in water exhibits some properties associated with electrolyte solutions (*e.g.,* it is a fair conductor of electricity), but in general terms it is more properly classified as a nonelectrolyte solution. By similar reasoning, an aqueous solution of carbon dioxide is also considered a nonelectrolyte solution even though carbon dioxide and water have a slight tendency to form carbonic acid, which, in turn, dissociates to a small extent to hydrogen ions and bicarbonate ions (HCO_3^-).

Endothermic and exothermic solutions. When two substances mix to form a solution, heat is either evolved (an exothermic process) or absorbed (an endothermic process); only in the special case of an ideal solution do substances mix without any heat effect. Most simple molecules mix with a small endothermic heat of solution, while exothermic heats of solution are observed when the components interact strongly with one another. An extreme example of an exothermic heat of mixing is provided by adding an aqueous solution of sodium hydroxide, a powerful base, to an aqueous solution of hydrogen chloride, a powerful acid; the hydroxide ions (OH^-) of the base combine with the hydrogen ions (H^+) of the acid to form water, a highly exothermic reaction that yields 75,300 calories per 100 grams of water formed. In nonelectrolyte solutions, heat effects are usually endothermic and much smaller, often about 100 calories, when roughly equal parts are mixed to form 100 grams of mixture.

Volume changes caused by the formation of a solution

Formation of a solution usually is accompanied by a small change in volume. If equal parts of benzene and stannic chloride are mixed, the temperature drops; if the mixture is then heated slightly to bring its temperature back to that of the unmixed liquids, the volume increases by about 2 percent. On the other hand, mixing roughly equal parts of acetone and chloroform produces a small decrease in volume, about 0.2 percent. It frequently happens that mixtures with endothermic heats of mixing expand—*i.e.,* show small increases in volume—while mixtures with exothermic heats of mixing tend to contract.

A large decrease in volume occurs when a gas is dissolved in a liquid. For example, at 0° C (32° F) and atmospheric pressure, the volume of 28 grams of nitrogen is 22,400 cubic centimetres. When these 28 grams of nitrogen are dissolved in an excess of water, the volume of the water increases only 40 cubic centimetres; the decrease in volume accompanying the dissolution of 28 grams of nitrogen in water is therefore 22,360 cubic centimetres. In this case, it is said that the nitrogen gas has been condensed into a liquid, the word condense meaning "to make dense"—*i.e.,* to decrease the volume.

PROPERTIES OF SOLUTIONS

Composition ratios. The composition of a liquid solution means the composition of that solution in the bulk—that is, of that part that is not near the surface. The interface (the surface separating them) between the liquid solution and some other phase (for example, a gas such as air) has a composition that differs, sometimes very much, from that of the bulk. The environment at an interface is significantly different from that in the bulk of the liquid, and in a solution the molecules of a particular component may prefer one environment over the other. If the molecules of one component in the solution prefer to be at the interface as opposed to the bulk, it is said that this component is positively adsorbed at the interface. In aqueous solutions of organic liquids, the organic component is usually positively adsorbed at the solution–air interface; as a result, it is often possible to separate a mixture of an organic solute from water by a process called froth separation. Air is bubbled vigorously into the solution, and a froth is formed. The composition of the froth differs from that of the bulk because the organic solute concentrates at the interfacial region. The froth is mechanically removed and collapsed, and, if further separation is desired, a new froth is generated. The tendency of some dissolved molecules to congregate at the surface has been utilized in water conservation. A certain type of alcohol, when added to water, concentrates at the surface to form a barrier to evaporating water molecules. In warm climates, therefore, water loss by evaporation from lakes can be significantly reduced by introducing a solute that adsorbs positively at the lake–air interface.

The composition of a solution can be expressed in a variety of ways, the simplest of which is to use weight fraction, or weight percent; for example, the salt content of seawater is about 3.5 weight percent; *i.e.,* of 100 grams of seawater, 3.5 grams is salt. For a fundamental understanding of solution properties, however, it is often useful to express composition in terms of molecular units such as molecular concentration, molality, or mole fraction. To understand these terms, it is necessary to define atomic and molecular weights. The atomic weight of elements is a relative figure, with one atom of the carbon-12 isotope being assigned the atomic weight of 12; the atomic weight of hydrogen is then approximately 1, of oxygen approximately 16, and the molecular weight of water (H_2O) 18. The atomic and molecular theory of matter asserts that the atomic weight of any element in grams must contain the same number of atoms as the atomic weight in grams (the gram-atomic weight) of any other element. Thus, two grams of molecular hydrogen (H_2)—its gram-molecular weight—contain the same number of molecules as 18 grams of water or as 32 grams of oxygen molecules (O_2). Further, a specified volume of any gas (at low pressure) contains the same number of molecules as the same volume of any other gas, at the same temperature and pressure. At standard temperature and pressure (0° C and 1 atmosphere) the volume of one gram-molecular weight of any gas has been determined experimentally to be approximately 22.4 litres. The number of molecules in this volume of gas, or in the gram-molecular weight of any compound, is called Avogadro's number.

Role of atomic weight

Molarity. By molecular concentration is meant the number of molecules of a particular component per unit volume. Since the number of molecules in a litre or even a cubic centimetre is enormous, it has become common practice to use what are called molar, rather than molecular, quantities. A mole is the gram-molecular weight of a substance and, therefore, also Avogadro's number of molecules (6.02×10^{23}). Thus, the number of moles in a sample is the weight of the sample divided by the molecular weight of the substance; and it is also the number of molecules in the sample divided by Avogadro's number. Instead of using molecular concentration, it is more convenient to use molar concentration; instead of saying, for example, that the concentration is 12.04×10^{23} molecules per litre, it is simpler to say that it is two moles per litre. Concentration in moles per litre (*i.e.,* molarity) is usually designated by the letter M.

Molality. In electrolyte solutions it is common to distinguish between the solvent (usually water) and the dissolved substance, or solute, which dissociates into ions. For these solutions it is useful to express composition in

terms of molality, designated as m, a unit proportional to the number of undissociated solute molecules (or, alternatively, to the number of ions) per 1,000 grams of solvent. The number of molecules or ions in 1,000 grams of solvent usually is very large, so molality is defined as the number of moles per 1,000 grams of solvent.

Formality. Many compounds do not exist in the form of molecules, either in the pure substances or in their solutions. The particles that make up sodium chloride (NaCl), for example, are sodium ions (Na^+) and chloride ions (Cl^-), and although equal numbers of these two ions are present in any sample of sodium chloride, no Na^+ ion is associated with a particular Cl^- ion to form a neutral molecule having the composition implied by the formula. Therefore, even though the compositions of such compounds are well defined, it would be erroneous to express concentrations of their solutions in terms of molecular weights. A useful concept in cases of this kind is that of the formula weight, defined as the sum of the weights of the atoms in the formula of the compound; thus, the formula weight of sodium chloride is the sum of the atomic weights of sodium and chlorine, 23 plus 35.5, or 58.5, and a solution containing 58.5 grams of sodium chloride per litre is said to have a concentration of one formal, or 1 *f.*

Formula weight

Mole fraction and mole percentage. It often is useful to express the composition of nonelectrolyte solutions in terms of mole fraction or mole percentage. In a binary mixture—*i.e.,* a mixture of two components, 1 and 2—there are two mole fractions, x_1 and x_2, which must obey the relation $x_1 + x_2 = 1$. The mole fraction x_1 means the fraction of molecules of species 1 in the solution, and x_2 means the fraction of molecules of species 2 in the solution. (Mole percentage is the mole fraction multiplied by 100.)

Volume fraction. The composition of a nonelectrolyte solution containing very large molecules, known as polymers, is most conveniently expressed by the volume fraction (Φ)—*i.e.,* the volume of polymer used to prepare the solution divided by the sum of that volume of polymer and the volume of the solvent.

Equilibrium properties. A quantitative description of liquid-solution properties when the system is in equilibrium is provided by relating the vapour pressure of the solution to its composition. The vapour pressure of a liquid, pure or mixed, is the pressure exerted by those molecules that escape from the liquid to form a separate vapour phase above the liquid. If a quantity of liquid is placed in an evacuated, closed container the volume of which is slightly larger than that of the liquid, most of the container is filled with the liquid, but, immediately above the liquid surface, a vapour phase forms consisting of molecules that have passed through the liquid surface from liquid to gas; the pressure exerted by that vapour phase is called the vapour (or saturation) pressure. For a pure liquid, this pressure depends only on the temperature, the best known example being the normal boiling point, which is that temperature at which the vapour pressure is equal to the pressure of the atmosphere. Figure 47 shows vapour pressures for a few common liquids. The vapour pressure is one atmosphere at 100° C (212° F) for water, at 78.5° C (173.3° F) for ethyl alcohol, and at 125.7° C (258.3° F) for octane. In a liquid solution the component with the higher vapour pressure is called the light component, and that with the lower vapour pressure is called the heavy component.

Vapour pressure

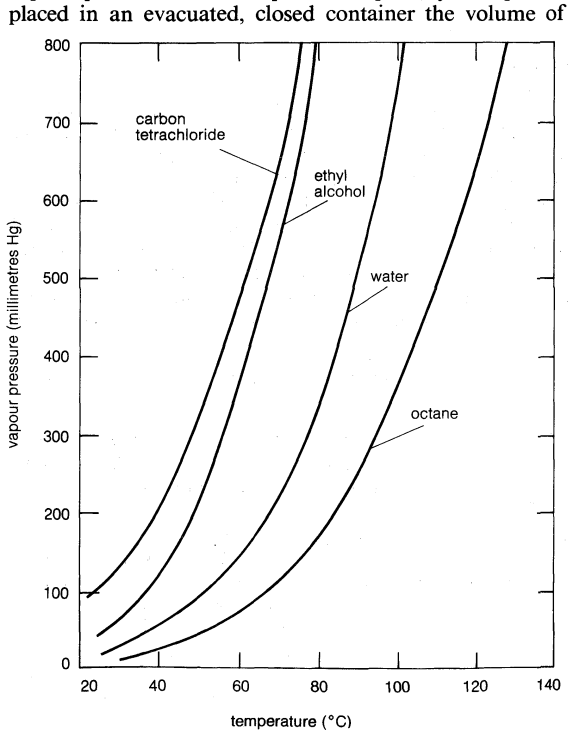


Figure 47: Vapour pressures of several pure liquids.

In a liquid mixture, the vapour pressure depends not only on the temperature but also on the composition, and the key problem in understanding the properties of solutions lies in determining this composition dependence. The simplest approximation is to assume that, at constant temperature, the vapour pressure of a solution is a linear function of its composition (*i.e.,* as one increases, so does the other in such proportion that, when the values are plotted, the resulting graph is a straight line). A mixture following this approximation is called an ideal solution.

Fugacity. In a pure liquid, the vapour generated by its ecscaping molecules necessarily has the same composition as that of the liquid. In a mixture, however, the composition of the vapour is not the same as that of the liquid; the vapour is richer in that component whose molecules have greater tendency to escape from the liquid phase. This tendency is measured by fugacity, a term derived from the Latin *fugere* ("to escape, to fly away"), which is also the root of the word fugitive. The fugacity of a component in a mixture is (essentially) the pressure that the component exerts in the vapour phase when the vapour is in equilibrium with the liquid mixture. (A state of equilibrium is attained when all the properties remain constant in time and there is no net transfer of energy or matter between the vapour and the liquid.) If the vapour phase can be considered to be an ideal gas (*i.e.,* the molecules in the gas phase are assumed to act independently and without any influence on each other), then the fugacity of a component, i, is equal to its partial pressure, which is defined as the product of the total vapour pressure, P, and the vapour-phase mole fraction, y_i. Assuming ideal gas behaviour for the vapour phase, the fugacity (y_iP) equals the product of the liquid-phase mole fraction, x_i, the vapour pressure of pure liquid at the same temperature as that of the mixture, P_i and the activity coefficient, γ_i. The real concentration of a substance may not be an accurate measure of its effectiveness because of physical and chemical interactions, in which case an effective concentration must be used, called the activity. The activity is given by the product of the mole fraction x_i and the activity coefficient γ_i. The equation is:

The activity coefficient

$$y_iP = \gamma_i x_i P_i^\circ. \qquad (10)$$

Raoult's law. In a real solution, the activity coefficient, γ_i, depends on both temperature and composition, but, in an ideal solution, γ_i equals 1 for all components in the mixture. For an ideal binary mixture then, the above equation becomes, for components 1 and 2, $y_1P = x_1P_1^\circ$, and $y_2P = x_2P_2^\circ$, respectively. Upon adding these equations—remembering that $x_1 + x_2 = 1$, and $y_1 + y_2 = 1$—total pressure, P, is shown to be expressed by the equation $P = x_1P_1^\circ + x_2P_2^\circ = x_1[P_1^\circ - P_2^\circ] + P_2^\circ$, which is a linear function of x_1.

Assuming $\gamma_1 = \gamma_2 = 1$, equations for y_1P and y_2P express what is commonly known as Raoult's law, which states that at constant temperature the partial pressure of a component in a liquid mixture is proportional to its mole fraction in that mixture (*i.e.,* each component exerts a pressure that depends directly on the number of its molecules present). It is unfortunate that the word law is associated with this relation, because only very few mixtures behave according to the equations for ideal binary

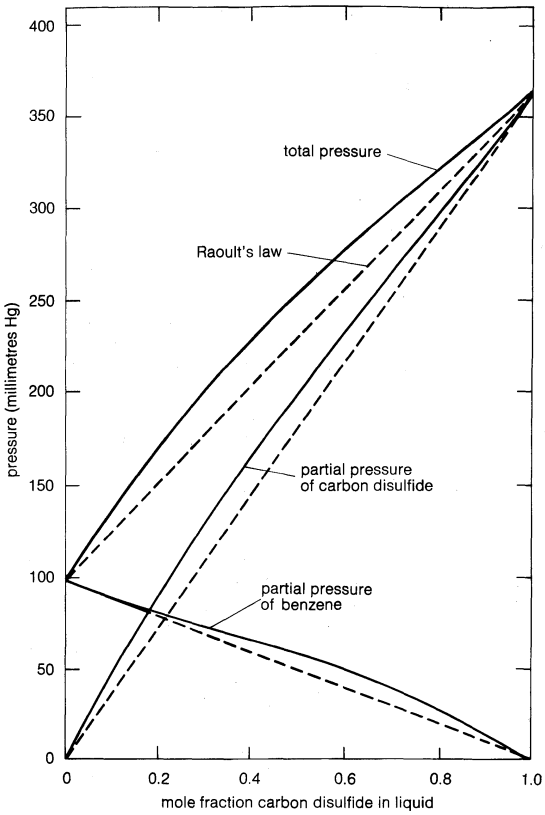


Figure 48: Total pressure and partial pressures for the system benzene–carbon disulfide at 25° C, or 77° F (see text).

mixtures. In most cases the activity coefficient, γ_i, is not equal to unity. When γ_i is greater than 1, there are positive deviations from Raoult's law; when γ_i is less than 1, there are negative deviations from Raoult's law.

An example of a binary system that exhibits positive deviations from Raoult's law is represented in Figure 48, the partial pressures and the total pressure being related to the liquid-phase composition; if Raoult's law were valid, all the lines would be straight, as indicated. As a practical result of these relationships, it is often possible by a series of repeated vaporizations and condensations to separate a liquid mixture into its components, a sequence of steps called fractional distillation.

Azeotropic mixtures

When the vapour in equilibrium with a liquid mixture has a composition identical to that of the liquid, the mixture is called an azeotrope. It is not possible to separate an azeotropic mixture by fractional distillation because no change in composition is achieved by a series of vaporizations and condensations. Azeotropic mixtures are common. At the azeotropic composition, the total pressure (at constant temperature) is always either a maximum or a minimum, and the boiling temperature (at constant pressure) is always either a minimum or a maximum.

Partial miscibility. Only pairs of liquids that are completely miscible have been considered so far. Many pairs of liquid mixtures, however, are only partially miscible in one another, the degree of miscibility often depending strongly on temperature. In most cases, rising temperature produces enhanced solubility, but this is not always so. For example, at 50° C (122° F) the solubility (weight percent) of *n*-butyl alcohol in water is 6.5 percent, whereas that of water in *n*-butyl alcohol is 22.4 percent. At 127° C (261° F), the upper consolute temperature, complete miscibility is attained: above 127° C the two liquids mix in all proportions, but below 127° C they show a miscibility gap. Thus, if *n*-butyl alcohol is added to water at 50° C, there is only one liquid phase until 6.5 weight percent of the mixture is alcohol; when more alcohol is added, a second liquid phase appears the composition of which is 22.4 weight percent water. When sufficient alcohol is present to make the overall composition 77.6 weight percent alcohol, the first phase disappears, and only one liquid phase remains. A qualitatively different example is the system water–triethylamine, which has a lower consolute temperature at 17° C (63° F). Below 17° C, the two liquids are completely miscible, but at higher temperatures they are only partially miscible. Finally, it is possible, although rare, for a binary system to exhibit both upper and lower consolute temperatures. Above 128° C (262° F) and below 49° C (120° F), butyl glycol and water are completely miscible, but between these temperatures they do not mix in all proportions.

Colligative properties. Colligative properties depend only on the concentration of the solute, not on the specific properties of the solute molecules. The concept of an ideal solution, as expressed by Raoult's law (at constant temperature the partial pressure of a component in a liquid mixture is proportional to its mole fraction in the mixture), was already well-known during the last quarter of the 19th century, and it provided the early physical chemists with a powerful technique for measuring molecular weights. (Reliable measurements of molecular weights, in turn, provided important evidence for the modern atomic and molecular theory of matter.)

Rise in boiling point. It was observed that, whenever one component in a binary solution is present in large excess, the partial pressure of that component is correctly predicted by Raoult's law, even though the solution may exhibit departures from ideal behaviour in other respects. When Raoult's law is applied to the solvent of a very dilute solution containing a nonvolatile solute, it is possible to calculate the mole fraction of the solute from an experimental determination of the rise in boiling point that results when the solute is dissolved in the solvent. Since the separate weights of solute and solvent are readily measured, the procedure provides a simple experimental method for the determination of molecular weight. If a weighed amount of a nonvolatile substance, w_2, is dissolved in a weighed amount of a solvent, w_1, at constant pressure, the increase in the boiling temperature, ΔT_{b1}, the gas constant, R (derived from the gas laws), the heat of vaporization of the pure solvent per unit weight, l_1^{vap}, the boiling temperature of pure solvent, T_{b1}, are related in a simple product of ratios to equal the molecular weight of the solute, M_2. The equation is:

$$M_2 = \left(\frac{RT_{b_1}^{\,2}}{l_{\,1}^{\,\text{vap}}}\right)\left(\frac{w_2}{w_1}\right)\left(\frac{1}{\Delta T_{b1}}\right). \qquad (11)$$

The essence of this technique follows from the observation that, in a dilute solution of a nonvolatile solute, the rise in boiling point is proportional to the number of solute molecules, regardless of their size and mass.

Decrease in freezing point. Another colligative property of solutions is the decrease in the freezing temperature of a solvent that is observed when a small amount of solute is dissolved in that solvent. By reasoning similar to that leading to Equation 11, the freezing-point depression, ΔT_f, the freezing temperature of pure solvent, T_{f1}, the heat of fusion (also called the heat of melting) of pure solvent per unit weight, l_1^{fusion}, and the weights of solute and solvent in the solution, w_2 and w_1, respectively, are so related as to equal the molecular weight of solute, M_2, in the equation

$$M_2 = \left(\frac{RT_{f1}^{\,2}}{l_{\,1}^{\,\text{fusion}}}\right)\left(\frac{w_2}{w_1}\right)\left(\frac{1}{\Delta T_f}\right). \qquad (12)$$

A well-known practical application of freezing-point depression is provided by adding antifreeze to the cooling water in an automobile's radiator. Water alone freezes at 0° C (32° F), but the freezing temperature decreases appreciably when ethylene glycol is mixed with water.

Osmotic pressure. A third colligative property, osmotic pressure, helped to establish the fundamentals of modern physical chemistry and played a particularly important role in the early days of solution theory. Osmosis is especially important in medicine and biology, but in recent years it has also been applied industrially to problems such as the concentration of fruit juices, the desalting of seawater, and the purification of municipal sewage. Osmosis occurs whenever a liquid solution is in contact with a

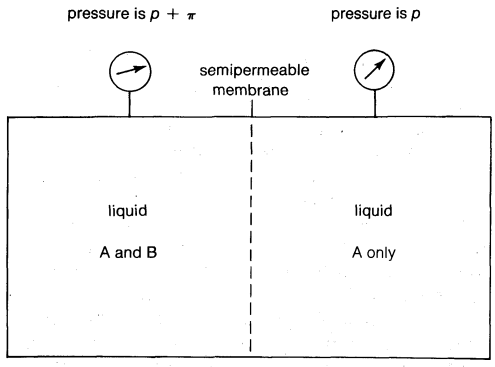


Figure 49: Osmotic pressure π caused by a membrane that allows A to pass, but not B.

semipermeable membrane—*i.e.*, a thin, porous wall whose porosity is such that some, but not all, of the components in the liquid mixture can pass through the wall. A semipermeable membrane is a selective barrier, and many such barriers are found in plants and animals. Osmosis gives rise to what is known as osmotic pressure, as illustrated in Figure 49, which shows a container at uniform temperature divided into two parts by a semipermeable membrane that allows only molecules of component A to pass from the left to the right side; the selective membrane does not allow molecules of component B to pass. Molecules of component A are free to pass back and forth through the membrane, but, at equilibrium, when the fugacity (escaping tendency) of A in the right-hand side is the same as that in the left-hand side, there is no net transfer of A from one side to the other. On the left side, the presence of B molecules lowers the fugacity of A, and, therefore, to achieve equal fugacities for A on both sides, some compensating effect is needed on the left side. This compensating effect is an enhanced pressure, designated by π and called osmotic pressure. At equilibrium, the pressure in the left side of the container is larger than that in the right side; the difference in pressure is π. In the simplest case, when the concentration of B is small (*i.e.*, A is in excess), the osmotic pressure is the product of the gas constant (R), the absolute temperature (T), and the concentration of B (c_B) in the solution expressed in moles of B per unit volume: $\pi = RTc_B$. Since the osmotic pressure for a dilute solution is proportional to the number of solute molecules, it is a colligative property, and, as a result, osmotic-pressure measurements are often used to determine molecular weights, especially for large molecules such as polymers. When w_B grams of solute B are added to a large amount of solvent A at temperature T, and V is the volume of liquid solvent A in the left side of the container, then the molecular weight of B, M_B, is given by

Osmotic equilibrium

$$M_B = \frac{w_B RT}{\pi V}. \qquad (13)$$

Transport properties in solutions. Pure fluids have two transport properties that are of primary importance: viscosity and thermal conductivity. (Transport properties are not equilibrium properties because they reflect not what happens at equilibrium but the speed at which equilibrium is attained.) In solutions these two transport properties are also important. In addition, there is a third one, called diffusivity.

Viscosity. The viscosity of a fluid (pure or not) is a measure of its ability to resist deformation. If water is poured into a thin vertical tube with a funnel at the top, it flows easily through the tube, but salad oil is difficult to force into the tube. If the oil is heated, however, its flow through the tube is much facilitated. The intrinsic property that is responsible for these phenomena is the viscosity (the "thickness") of the fluid, a property which is often strongly affected by temperature. All fluids (liquid or gas) exhibit viscous behaviour (*i.e.*, all fluids resist deformation), but the range of viscosity is enormous: the viscosity of air is extremely small, while that of glass is essentially infinite. The viscosity of a solution depends not only on temperature but also on its composition. By varying the composition of a petroleum mixture, it is possible to attain a desired viscosity at a particular temperature. This is precisely what the oil companies do when they sell oil to a motorist: in winter, they recommend an oil with lower viscosity than that used in summer, because otherwise, on a cold morning, the viscosity of the lubricating oil may be so high that the car's battery will not be powerful enough to move the lubricated piston.

Thermal conductivity. The thermal conductivity of a material reflects its ability to transfer heat by conduction. In practical situations both viscosity and thermal conductivity are important, as is illustrated by the contrast between an air mattress and a water bed. Because of its low viscosity, air yields rapidly to an imposed load, and thus the air mattress responds quickly when someone lying on it changes position. Water, because of its higher viscosity, noticeably resists deformation, and someone lying on a water bed experiences a caressing response whenever position is changed. At the same time, since the thermal conductivity as well as the viscosity of water are larger than those of air, the user of a water bed rapidly gets cold unless a heater keeps the water warm. No heater is required by the user of an air mattress because stagnant air is inefficient in removing heat from a warm body.

Composition and temperature affect the thermal conductivity of a solution but, in typical liquid mixtures, the effect on viscosity is much larger than that on thermal conductivity.

Diffusivity. While viscosity is concerned with the transfer of momentum and thermal conductivity with the transfer of heat, diffusivity is concerned with the transport of molecules in a mixture. If a lump of sugar is put into a cup of coffee, the sugar molecules travel from the surface of the lump into the coffee at a speed determined by the temperature and by the pertinent intermolecular forces. The characteristic property that determines this speed is called diffusivity—*i.e.*, the ability of a molecule to diffuse through a sea of other molecules. Diffusivities in solids are extremely small, and those in liquids are much smaller than those in gases. As a result, a spoon is used to stir the coffee to speed up the motion of the sugar molecules; but if a room smells foul from cigarette smoke, no effort is needed to obtain significant relief—opening the windows for a few minutes is sufficient.

Comparison of viscosity and diffusivity

In order to define diffusivity, it is necessary to consider a binary fluid mixture in which the concentration of solute molecules is c_1 at position 1 and c_2 at position 2, l centimetres from position 1; if c_1 is larger than c_2, then the concentration gradient (change with respect to distance), or $(c_2 - c_1)/l$, is a negative number, indicating that molecules of solute spontaneously diffuse from position 1 to position 2; the number of solute molecules that pass through an area of one square centimetre perpendicular to l, per second, is called the flux J (expressed in molecules per second per square centimetre). The diffusivity D is given by the ratio of the flux to the gradient multiplied by minus one:

$$D = -\frac{J}{\left(\frac{c_2 - c_1}{l}\right)}. \qquad (14)$$

The minus sign is introduced because, when the gradient is positive, J is negative, and, by convention, D is a positive number. In binary gaseous mixtures, diffusivity depends only weakly on the composition, and, therefore, to a good approximation, the diffusivity of gas A in gas B is the same as that of gas B in gas A. In liquid systems, however, the diffusivity of solute A in solvent B may be significantly different from that of solute B in solvent A. In a very viscous fluid, molecules cannot rapidly move from one place to another. Therefore, in liquid systems, the diffusivity of solute A depends strongly on the viscosity of solvent B and vice versa. While the letter D is always used for diffusivity, viscosity is commonly given the symbol η (Greek eta); in many liquid solutions it is observed that, as the composition changes (as long as the temperature remains constant), the product $D\eta$ remains nearly the same.

THERMODYNAMICS AND INTERMOLECULAR FORCES IN SOLUTIONS

The properties of solutions depend, essentially, on two characteristics: first, the manner in which the molecules arrange themselves (that is, the geometrical array in which the molecules occupy space) and, second, the nature and strength of the forces operating between the molecules.

Energy considerations. The first characteristic is reflected primarily in the thermodynamic quantity *S*, called entropy, which is a measure of disorder, and the second characteristic is reflected in the thermodynamic quantity *H*, called enthalpy, which is a measure of potential energy—*i.e.*, the energy that must be supplied to separate all the molecules from one another. Enthalpy minus the product of the absolute temperature *T* and entropy equals a thermodynamic quantity *G*, called Gibbs energy (also called free energy):

Enthalpy

$$G = H - TS. \tag{15}$$

From the second law of thermodynamics, it can be shown that, at constant temperature and pressure, any spontaneous process is accompanied by a decrease in Gibbs energy. The change in *G* that results from mixing is designated by ΔG, which, in turn, is related to changes in *H* and *S* at constant temperature by the equation

$$\Delta G = \Delta H - T\Delta S. \tag{16}$$

At a fixed temperature and pressure, two substances mix spontaneously whenever ΔG is negative; that is, mixing (either partial or complete) occurs whenever the Gibbs energy of the substances after mixing is less than that before mixing.

The two characteristics that determine solution behaviour, structure and intermolecular forces, are, unfortunately, not independent, because the structure is influenced by the intermolecular forces and because the potential energy of the mixture depends on the structure. Only in limiting cases is it possible, on the one hand, to calculate ΔS (the entropy change upon mixing) from structural considerations alone and, on the other, to calculate ΔH (the enthalpy change of mixing) exclusively from relations describing intermolecular forces. Nevertheless, such calculations have proved to be useful for establishing models that approximate solution behaviour and that serve as guides in interpreting experimental measurements. Solutions for which structural considerations are dominant are called athermal solutions, and those for which the effects of intermolecular forces are more important than those of structure are called regular solutions (see below *Regular solutions* and *Athermal solutions*).

Effects of molecular structure. A variety of forces operate between molecules, and there is a qualitative relation between the properties of a solution and the types of intermolecular forces that operate within it. The volume occupied by a solution is determined primarily by repulsive forces. When two molecules are extremely close to one another, they must necessarily exert a repulsive force on each other since two molecules of finite dimensions cannot occupy the same space; two molecules in very close proximity resist attempts to shorten the distance between them just as passengers in a crowded subway car resist entry of an additional passenger.

At larger distances of separation, molecules may attract or repel each other depending on the sign (plus or minus) and distribution of their electrical charge. Two ions (atoms or molecules with net positive or negative electric charges) attract one another if the charge on one is positive and that on the other is negative; they repel when both carry charges of the same sign. Forces between ions are called Coulomb forces and are characterized by their long range; the force (F) between two ions is inversely proportional to the square of the distance between them—*i.e.*, F varies as $1/r^2$. Noncoulombic physical forces between molecules decay more rapidly with distance; *i.e.*, in general F varies as $1/r^n$, n being larger than 2 for intermolecular forces other than those between ions.

The Coulomb force (F) equals the product of the magnitude of the charge on one ion (e_1) and that on the other (e_2) divided by the product of the distance squared (r^2) and the dielectric constant (ε):

$$F = \frac{e_1 e_2}{r^2 \varepsilon}. \tag{17}$$

If both e_1 and e_2 are positive, F is positive and the force is repulsive. If either e_1 or e_2 is positive, while the other is negative, F is negative and the force is attractive. Coulomb forces are dominant in electrolyte solutions.

Molecular structure and charge distribution. If a molecule has no net electrical charge, its negative charge is equal to its positive charge. The forces experienced by such molecules depend on how the positive and negative charges are arranged in space. If the arrangement is spherically symmetric, the molecule is said to be nonpolar; if there is an excess of positive charge on one end of the molecule and an excess of negative charge on the other, the molecule has a dipole moment (*i.e.*, a measurable tendency to rotate in an electric or magnetic field) and is therefore called polar. The dipole moment (μ) is defined as the product of the magnitude of the charge, e, and the distance separating the positive and negative charges, l: $\mu = el$. Electric charge is measured in electrostatic units (esu), and the typical charge at one end of a molecule is of the order of 10^{-10} esu; the distance between charges is of the order of 10^{-8} centimetres. Dipole moments, therefore, usually are measured in debyes (one debye is 10^{-18} esu-cm). For nonpolar molecules $\mu = 0$.

Dipole moment

Polar molecules. The force F between two polar molecules is directly proportional to the product of the two dipole moments (μ_1 and μ_2) and inversely proportional to the fourth power of the distance between them (r^4): F varies as $\mu_1\mu_2/r^4$. The equation for this relationship contains a constant of proportionality ($F = k\mu_1\mu_2/r^4$) the sign and magnitude of which depend on the mutual orientation of the two dipoles; if the positive end of one faces the negative end of the other, the constant of proportionality is negative (meaning that an attractive force exists), while it is positive (meaning that a repulsive force exists) when the positive end of one faces the positive end of the other. When polar molecules are free to rotate, they tend to favour orientations that led to attractive forces. To a first approximation, the force (averaged over all orientations) is inversely proportional to the temperature and to the seventh power of the distance of separation. Mixtures of polar molecules often exhibit only mild deviations from ideality, but mixtures containing polar and nonpolar molecules are frequently strongly nonideal. Because of the qualitative and quantitative differences in intermolecular forces, the molecules segregate: the polar molecules prefer to be with each other, and so do the nonpolar ones. Only at higher temperatures, such that the thermal energy of the molecules offsets the cohesion between identical molecules, do the two liquids mix in all proportions. In mixtures containing polar and nonpolar components, deviations from Raoult's law diminish as temperature rises.

Nonpolar molecules. A nonpolar molecule is one whose charge distribution is spherically symmetric when averaged over time; since the charges oscillate, a temporary dipole moment exists at any given instant in a so-called nonpolar molecule. These temporary dipole moments fluctuate rapidly in magnitude and direction, giving rise to intermolecular forces of attraction called London (or dispersion) forces. All molecules, charged or not, polar or not, interact by London forces. To a first approximation, the London force between two molecules is inversely proportional to the seventh power of the distance of separation; it is therefore short-range, decreasing rapidly as one molecule moves away from the other. The London theory indicates that for simple molecules positive deviations from Raoult's law may be expected (*i.e.*, the activity coefficient γ_i is greater than 1, as explained previously). Since the London theory suggests that the attractive forces between unlike simple molecules are smaller than those corresponding to an ideal solution, the escaping tendency of the molecules in solution is larger than that calculated by Raoult's law. As a result, mixing of small nonpolar molecules is endothermic (absorbing heat from the surroundings) and the volume occupied by the liquid solution often exceeds that of the unmixed components (the components expand on mixing).

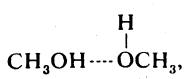
London forces

In addition to the forces listed above, there are so-called

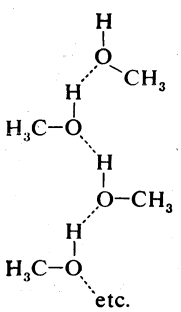
induction forces set up when a charged or polar molecule induces a dipole in another molecule: the electric field of the inducing molecule distorts the charge distribution in the other. When a charged molecule induces a dipole in another, the force is always attractive and is inversely proportional to the fifth power of the distance of separation. When a polar molecule induces a dipole in another molecule, the force is also attractive and is inversely proportional to the seventh power of separation. Induction forces are usually small but may make a significant contribution to the energy of a mixture of molecules that are strongly dissimilar.

Effects of chemical interactions. In many cases the properties of a mixture are determined primarily by forces that are more properly classified as chemical rather than physical. For example, when dinitrogen pentoxide is dissolved in water, a new substance, nitric acid, is formed; and it is necessary to interpret the behaviour of such a solution in terms of its chemical properties, which, in this case, are more important than its physical properties. This example is extreme, and there are many solutions for which the chemical effect is less severe but nevertheless dominant.

Hydrogen bonding: association. This dominance is especially important in those solutions that involve hydrogen bonding. Whenever a solution contains molecules with an electropositive hydrogen atom and with an electronegative atom (such as nitrogen, oxygen, sulfur, or fluorine), hydrogen bonding may occur and, when it does, the properties of the solution are affected profoundly. Hydrogen bonds may form between identical molecules or between dissimilar molecules; for example, methanol (CH_3OH) has an electropositive (electron-attracting) hydrogen atom and also an electronegative (electron-donating) oxygen atom, and therefore two methanol molecules may hydrogen-bond singly to form the structure



or in chains to form



Hydrogen bonding between identical molecules is often called association.

Hydrogen bonding: solvation. In a mixture of methanol and, say, pyridine (C_5H_5N), hydrogen bonds can also form between the electropositive hydrogen atom in methanol and the electronegative nitrogen atom in pyridine. Hydrogen bonding between dissimilar molecules is often called solvation. Since the extent of association or solvation or both depends on the concentrations of the solution's components, the partial pressure of a component is not even approximately proportional to its mole fraction as given by Raoult's law; therefore, large deviations from Raoult's law are commonly observed in solutions in which hydrogen bonding is extensive. Broadly speaking, association of one component, but not the other, tends to produce positive deviations from Raoult's law, because the associating component associates to a smaller extent when it is surrounded by other molecules than it does in the pure state. On the other hand, solvation between dissimilar molecules tends to produce negative deviations from Raoult's law.

Deviations from Raoult's law

THEORIES OF SOLUTIONS

Solutions of nonelectrolytes. *Activity coefficients and excess functions.* As has been explained previously, when actual concentrations do not give simple linear relations for the behaviour of a solution, activity coefficients, symbolized by γ, are used in expressing deviations from Raoult's law. Activity coefficients are directly related to excess functions, and, in attempting to understand solution behaviour, it is convenient to characterize nonelectrolyte solutions in terms of these functions. In particular, it is useful to distinguish between two types of limiting behaviour: one corresponds to that of a regular solution; the other, to that of an athermal solution (*i.e.*, when components are mixed, no heat is generated or absorbed).

In a binary mixture with mole fractions x_1 and x_2 and activity coefficients γ_1 and γ_2, these quantities can be related to a thermodynamic function designated by G^E, called the excess Gibbs (or free) energy. The significance of the word excess lies in the fact that G^E is the Gibbs energy of a solution in excess of what it would be if it were ideal.

In a binary solution the two activity coefficients are not independent but are related by an exact differential equation called the Gibbs-Duhem relation. Thanks to this equation, if experimental data at constant temperature are available for γ_1 and γ_2 as a function of composition, it is possible to check the data for thermodynamic consistency: the data are said to be consistent only if they satisfy the Gibbs-Duhem relation. Experimental data that do not satisfy this relation are thermodynamically inconsistent and therefore must be erroneous.

To establish a theory of solutions, it is necessary to construct a theoretical (or semitheoretical) equation for the excess Gibbs energy as a function of absolute temperature (T) and the mole fractions x_1 and x_2. After such an equation has been established, the individual activity coefficients can readily be calculated.

Gibbs energy, by definition, consists of two parts: one part is the enthalpy, which reflects the intermolecular forces between the molecules, which, in turn, are responsible for the heat effects that accompany the mixing process (enthalpy is, in a general sense, a measure of the heat content of a substance); and the other part is the entropy, which reflects the state of disorder (mixed-upness, or a measure of the random behaviour of particles) in the mixture. The excess Gibbs energy G^E is equal to the excess enthalpy H^E minus the excess entropy S^E times the absolute temperature T:

$$G^E = H^E - TS^E. \tag{18}$$

The word excess means in excess of that which would prevail if the solution were ideal. In the simplest case, both H^E and S^E are zero; in that case the solution is ideal and $\gamma_1 = \gamma_2 = 1$. In the general case, neither H^E nor S^E is zero, but two types of semi-ideal solutions can be designated: in the first, S^E is zero but H^E is not; this is called a regular solution. In the second, H^E is zero but S^E is not; this is called an athermal solution. An ideal solution is both regular and athermal.

Types of semi-ideal solutions

Regular solutions. The word regular implies that the molecules mix in a completely random manner, which means that there is no segregation or preference; a given molecule chooses its neighbours with no regard for chemical identity (species 1 or 2). In a regular solution of composition x_1 and x_2, the probability that the neighbour of a given molecule is of species 1 is given by x_1, and the probability that it is of species 2 is given by x_2.

Two liquids form a solution that is approximately regular when the molecules of the two liquids do not differ appreciably in size and there are no strong orienting forces caused by dipoles or hydrogen bonding. In that event, the mixing process can be represented by the lattice model shown in Figure 50; the left half of the diagram shows pure liquids 1 and 2, and the right half shows the mixture obtained when the central molecule of liquid 1 is interchanged with the central molecule of liquid 2. Before interchange, the potential energy between central molecule 1 and one of its immediate neighbours is Γ_{11} and that between central molecule 2 and one of its immediate neighbours is Γ_{22}. After interchange, the potential energy between molecule 1 and one of its immediate neighbours is Γ_{12}, and that between molecule 2 and one of its immediate neighbours is also Γ_{12}. The change in energy that

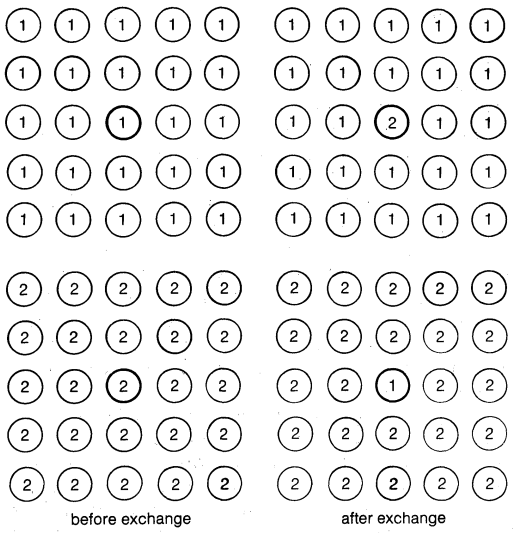


Figure 50: Physical significance of interchange energy. The energy absorbed in the process above is 2ω (see text).

accompanies this mixing process is equal to twice the interchange energy (ω), which is equal to the potential energy after mixing less one-half the sum of the potential energies before mixing, the whole multiplied by the number of immediate neighbours, called the coordination number (z), surrounding the two shifted molecules:

$$\omega = z[\Gamma_{12} - \tfrac{1}{2}(\Gamma_{11} + \Gamma_{22})]. \qquad (19)$$

In the two-dimensional representation (Figure 50), z equals 4; but, in three dimensions, z varies between 6 and 12, depending on the lattice geometry. In this simple lattice model, the interchange process occurs without change of volume; thus, in this particular case, the excess enthalpy is the same as the energy change upon mixing. Assuming regular-solution behaviour (*i.e.*, $S^E = 0$), an equation may be derived relating Gibbs energy, Avogadro's number, interchange energy, and mole fractions. In principle, the interchange energy (ω) may be positive or negative, but, for simple molecules, for which only London forces of attraction are important, ω is positive. The equation obtained from the simple lattice model can be extended semi-empirically to apply to mixtures of molecules whose sizes are not nearly the same by using volume fractions instead of mole fractions to express the effect of composition and by introducing the concept of cohesive energy density, which is defined as the potential energy of a liquid divided by its volume. The adjective cohesive is well chosen because it indicates that this energy is associated with the forces that keep the molecules close together in a condensed state. Again restricting attention to nonpolar molecules and assuming a completely random mixture ($S^E = 0$), an equation may be derived that requires only pure-component properties to predict the excess Gibbs energy (and hence the activity coefficients) of binary mixtures. Because of many simplifying assumptions, this equation does not give consistently accurate results, but in many cases it provides good semiquantitative estimates. The form of the equation is such that the excess Gibbs energy is larger than zero; hence, the equation is not applicable to mixtures that have negative deviations from Raoult's law.

Cohesive energy density

Athermal solutions. In a solution in which the molecules of one component are much larger than those of the other, the assumption that the solution is regular (*i.e.*, that $S^E = 0$) no longer provides a reasonable approximation even if the effect of intermolecular forces is neglected. A large flexible molecule (*e.g.*, a chain molecule such as polyethylene) can attain many more configurations when it is surrounded by small molecules than it can when surrounded by other large flexible molecules; the state of disorder in such a solution is therefore much larger than that of a regular solution in which $S^E = 0$. A solution of very large molecules (*i.e.*, polymers) in an ordinary liquid solvent is similar to a mixture of cooked spaghetti (the polymers) and tomato sauce. When there is a lot of sauce and little spaghetti, each piece of spaghetti is free to exist in many different shapes; this freedom, however, becomes restricted as the number of spaghetti pieces rises and the amount of sauce available for each strand declines. The excess entropy then is determined primarily by the freedom that the spaghetti has in the tomato sauce mixture relative to the freedom it has in the absence of sauce.

Regular solutions and athermal solutions represent limiting cases; real solutions are neither regular nor athermal. For real solutions it has been proposed to calculate G^E by combining the equations derived separately for regular solutions and for athermal solutions, but, in view of the restrictive and mutually inconsistent assumptions that were made in deriving these two equations, the proposal has met with only limited success.

Associated and solvated solutions. For those solutions in which there are strong intermolecular forces due to large dipole moments, hydrogen bonding, or complex formation, equations based on fundamental molecular theory cannot be applied, but it is frequently useful to apply a chemical treatment—*i.e.*, to describe the liquid mixture in terms of association and solvation, by assuming the existence of a variety of distinct chemical species in chemical equilibrium with one another. For example, there is much experimental evidence for association in acetic acid, in which most of the molecules dimerize; *i.e.*, two single acetic acid molecules, called monomers, combine to form a new molecule, called a dimer, through hydrogen bonding. When acetic acid is dissolved in a solvent such as benzene, the extent of dimerization of acetic acid depends on the temperature and on the total concentration of acetic acid in the solution. The escaping tendency (vapour pressure) of a monomer is much greater than that of a dimer, and it is thus possible to explain the variation of activity coefficient with composition for acetic acid in benzene; the activity coefficient of acetic acid in an excess of benzene is large because, under these conditions, acetic acid is primarily in the monomeric state, whereas pure acetic acid is almost completely dimerized. In the acetic acid–benzene system, association of acetic acid molecules produces positive deviations from Raoult's law.

When a solvent and a solute molecule link together with weak bonds, the process is called solvation. For example, in the system acetone–chloroform, a hydrogen bond is formed between the hydrogen atom in chloroform and the oxygen atom in acetone. In this case, hydrogen bonding depresses the escaping tendencies of both components, producing negative deviations from Raoult's law.

Solvation

While hydrogen bonding is frequently encountered in solutions, there are many other examples of weak chemical bond-formation between dissimilar molecules. The formation of such weak bonds is called complex formation—that is, formation of a new chemical species, called a complex, which is held together by weak forces that are chemical in nature rather than physical. Such complexes usually exist only in solution; because of their low stability, they cannot, in general, be isolated. The ability of different molecules to form complexes has a strong effect on solution behaviour. For example, the solubility of a sparingly soluble solute can be very much increased by complex formation: the solubility of silver chloride in water is extremely small since silver chloride dissociates only slightly to silver ion and chloride ion; however, when a small quantity of ammonia is added, solubility rises dramatically because of the reaction of six molecules of ammonia with one silver ion to form the complex ion $Ag(NH_3)_6^+$. By tying up silver ions and thereby forcing extensive dissociation of molecular silver chloride, the ammonia pulls silver chloride into aqueous solution.

In recent years there has been much interest in the use of computers to generate theoretical expressions for the activity coefficients of solutions. In many cases the calculations have been restricted to model systems, in particular to mixtures of hard-sphere (envisioned as billiard balls) molecules—*i.e.*, idealized molecules that have finite size but no forces of attraction. These calculations have produced a better understanding of the structure of simple liquid solutions since the manner in which nonpolar and non-hydrogen-bonding molecules arrange themselves

in space is determined primarily by their size and shape and only secondarily by their attractive intermolecular forces. The results obtained for hard-sphere molecules can be extended to real molecules by applying corrections required for attractive forces and for the "softness" of the molecules—*i.e.*, the ability of molecules to interpenetrate (overlap) at high temperatures. While practical results are still severely limited and while the amount of required computer calculation is large even for simple binary systems, there is good reason to believe that advances in the theory of solution will increasingly depend on computerized, as opposed to analytical, models.

Arrhenius' theory of salts

Solutions of electrolytes. Near the end of the 19th century, the properties of electrolyte solutions were investigated extensively by the early workers in physical chemistry. A suggestion of Svante August Arrhenius, a Swedish chemist, that salts of strong acids and bases (for example, sodium chloride) are completely dissociated into ions when in aqueous solution received strong support from electrical-conductivity measurements and from molecular-weight studies (freezing-point depression, boiling-point elevation, and osmotic pressure). These studies showed that the number of solute particles was larger than it would be if no dissociation occurred. For example, a 0.001 molal solution of a uni-univalent electrolyte (each ion has a valence, or charge, of 1, and, when dissociated, two ions are produced) such as sodium chloride, Na^+Cl^-, exhibits colligative properties (dependent only on the concentration) corresponding to a nonelectrolyte solution whose molality is 0.002; the colligative properties of a 0.001 molal solution of a univalent–divalent electrolyte (yielding three ions) such as magnesium bromide, $Mg^{2+}Br_2^-$, correspond to those of a nonelectrolyte solution with a molality of 0.003. At somewhat higher concentrations the experimental data showed some inconsistencies with Arrhenius' dissociation theory, and initially these were ascribed to incomplete, or partial, dissociation. In the years 1920–30, however, it was shown that these inconsistencies could be explained by electrostatic interactions (called Coulomb forces) of the ions in solution. The current view of electrolyte solutions is that, in water at normal temperatures, the salts of strong acids and strong bases are completely dissociated into ions at all concentrations up to the solubility limit. At high concentrations Coulombic interactions may cause the formation of ion pairs, which implies that the ions are not dispersed uniformly in the solution but have a tendency to form two-ion aggregates in which a positive ion seeks the close proximity of a negative ion and vice versa. While the theory of dilute electrolyte solutions is well advanced, no adequate theory exists for concentrated electrolyte solutions primarily because of the long-range Coulomb forces that dominate in ionic solutions.

Current theory of salts

The equilibrium properties of electrolyte solutions can be studied experimentally by electrochemical measurements, freezing point depressions, solubility determinations, osmotic pressures, or measurements of vapour pressure. Most electrolytes, such as salts, are nonvolatile at ordinary temperature, and, in that event, the vapour pressure exerted by the solution is the same as the partial pressure of the solvent. The activity coefficient of the solvent can, therefore, be found from total-pressure measurements, and, using the Gibbs-Duhem equation, it is then possible to calculate the activity coefficient of the electrolyte solute. This activity coefficient is designated by $\gamma_\pm$ to indicate that it is a mean activity coefficient for the positive and negative ions; since it is impossible to isolate positive ions and negative ions into separate containers, it is not possible to determine individual activity coefficients for the positive ions and for the negative ions. The mean activity coefficient $\gamma_\pm$ is so defined that it approaches a value of unity at very low molality where the ions are so far apart that they exert negligible influence on one another. For small concentrations of electrolyte, the theory of Peter Debye, a Dutch-born U.S. physical chemist, and Erich Hückel, a German chemist, relates $\gamma_\pm$ to the ionic strength, which is the sum of the products of the concentration of each ion (in moles per litre) and the square of its charge; the equation predicts that $\gamma_\pm$ decreases with rising ionic strength in agreement with experiment at very low ionic strength; at higher ionic strength, however, $\gamma_\pm$ rises, and in some cases $\gamma_\pm$ is greater than 1. The derivation of the Debye-Hückel theory clearly shows that it is limited to low concentrations. Many attempts have been made to extend the Debye-Hückel equation to higher electrolyte concentrations. One of the more successful attempts is based on the idea that the ions are solvated, which means that every ion is surrounded by a tight-fitting shell of solvent molecules.

The concept of solvation is often used to explain properties of aqueous solutions; one well-known property is the salting-out effect, in which the solubility of a nonelectrolyte in water is decreased when electrolyte is added. For example, the solubility of ethyl ether in water at 25° C (77° F) is 0.91 mole percent, but, in an aqueous solution containing 15 weight percent sodium chloride, it is only 0.13 mole percent. This decrease in solubility can be explained by postulating that some of the water molecules cannot participate in the dissolution of the ether because they are tightly held (solvated) by sodium and chloride ions.

Electrolyte solutions have long been of interest in industry since many common inorganic chemicals are directly obtained, or else separated, by crystallization from aqueous solution. Further, many important chemical and metallurgical products (*e.g.*, aluminum) are obtained or refined by electrochemical processes that occur in liquid solution. In recent years there has been renewed interest in electrolyte solutions because of their relevance to fuel cells as a possible source of power for automobiles.

Electrolytes in physiology

The properties of electrolyte solutions also have large importance in physiology. Many molecules that occur in biological systems bear electric charges; a large molecule that has a positive electric charge at one end and a negative charge at the other is called a zwitterion. Very large molecules, such as those of proteins, may have numerous positive and negative charges; such molecules are called polyelectrolytes. In solution, the conformation (that is, the three-dimensional structure) of a large, charged molecule is strongly dependent on the ionic strength of the dissolving medium; for example, depending on the nature and concentration of salts present in the solvent, a polyelectrolyte molecule may coagulate into a ball, it may stretch out like a rod, or it may form a coil or helix. The conformation, in turn, is closely related to the molecule's physiological function. As a result, improved understanding of the properties of electrolyte solutions has direct consequences in molecular biology and medicine.

SOLUBILITY OF SOLIDS AND GASES

Since the dissolution of one substance in another can occur only if there is a decrease in the Gibbs energy, it follows that, generally speaking, gases and solids do not dissolve in liquids as readily as do other liquids. To understand this, the dissolution of a solid can be visualized as occurring in two steps: in the first, the pure solid is melted at constant temperature to a pure liquid, and, in the second, that liquid is dissolved at constant temperature in the solvent. Similarly, the dissolution of a gas can be divided at some fixed pressure into two parts, the first corresponding to constant-temperature condensation of the pure gas to a liquid and the second to constant-temperature mixing of that liquid with solvent. In many cases, the pure liquids (obtained by melting or by condensation) may be hypothetical (*i.e.*, unstable and, therefore, physically unobtainable), but usually their properties can be estimated by reasonable extrapolations. It is found that the change in Gibbs energy corresponding to the first step is positive and, hence, in opposition to the change needed for dissolution. For example, at −10° C (14° F), ice is more stable than water, and, at 110° C (230° F) and one atmosphere, steam is more stable than water. Therefore, the Gibbs energy of melting ice at −10° C is positive, and the Gibbs energy of condensing steam at one atmosphere and 110° C is also positive. For the second step, however, the change in Gibbs energy is negative; its magnitude depends on the equilibrium composition of the mixture. Because of the positive Gibbs energy change that accompanies the first step, there is a barrier that makes it more difficult to dissolve solids and gases as compared with liquids.

Dissolution of a solid

For gases at normal pressures, the positive Gibbs energy of condensation increases with rising temperature, but, for solids, the posive Gibbs energy of melting decreases with rising temperature. For example, the change in energy, ΔG, of condensing steam at one atmosphere is larger at 120° C (248° F) than it is at 110° C (230° F), while the change in energy of melting ice at −5° C (23° F) is smaller than it is at −10° C (14° F). Thus, as temperature rises, the barrier becomes larger for gases but lower for solids, and therefore, with few exceptions, the solubility of a solid rises while the solubility of a gas falls as the temperature is raised.

For solids, the positive Gibbs energy "barrier" depends on the melting temperature. If the melting temperature is much higher than the temperature of the solution, the barrier is large, shrinking to zero when the melting temperature and solution temperature become identical.

Table 2 gives the solubilities of some common gases, and Table 3 the solubility of (solid) naphthalene in a few typical solvents. These solubilities illustrate the qualitative rule that "like dissolves like"; thus naphthalene, an aromatic hydrocarbon, dissolves more readily in another aromatic hydrocarbon such as benzene than it does in a chlorinated solvent such as carbon tetrachloride or in a hydrogen-bonded solvent such as methyl alcohol. By similar reasoning, the gas methane, a paraffinic hydrocarbon, dissolves more readily in another paraffin such as hexane than it does in water. In all three solvents, the gas hydrogen (which boils at −252.5° C [−422.5° F]) is less soluble than nitrogen (which boils at a higher temperature, −195.8° C [−320.4° F]).

Table 2: Solubilities of Some Gases*
(mole percent)

	heptane	benzene	water
Hydrogen	0.069	0.026	0.0015
Nitrogen	0.12	0.45	0.0012
Methane	0.47	0.21	0.0024
Carbon dioxide	0.77	0.97	0.0608

*At one atmosphere partial pressure, 25° C.

Table 3: Solubility of Naphthalene in Various Solvents*

solvent	mole percent naphthalene
Benzene	24.1
Carbon tetrachloride	20.5
Hexane	9.0
Methyl alcohol	1.8
Water	0.0004

*At 20° C.

While exceptions may occur at very high pressures, the solubility of a gas in a liquid generally rises as the pressure of that gas increases. When the pressure of the gas is much larger than the vapour pressure of the solvent, the solubility is often proportional to the pressure; this proportionality is consistent with Henry's law, which states that, if the gas phase is ideal, the solubility x_2 of gas 2 in solvent 1 is equal to the partial pressure (the vapour-phase mole fraction y_2 times the total pressure P—*i.e.*, y_2P) divided by a temperature-dependent constant, $H_{2,1}$ (called Henry's constant), which is determined to a large extent by the intermolecular forces between solute 2 and solvent 1; then

Henry's law

$$x_2 = \frac{y_2 P}{H_{2,1}}. \qquad (20)$$

When the vapour pressure of solvent 1 is small compared with the total pressure, then the vapour-phase mole fraction of gas 2 is approximately one, and the solubility of the gas is proportional to the total pressure. (J.M.P.)

GASEOUS STATE

General properties of gases

All matter consists of atoms or molecules or mixtures of these in constant motion, whether it is through space, called translational motion, or motion in one place, called vibrational motion. In the gaseous state, each particle moves about in three dimensions almost independently of the influence of nearby particles, whereas in liquids and solids the motion of the same atoms and molecules is subordinated to the forces exerted on them by their neighbours. The freely moving atoms and molecules of a gas are constantly colliding with one another and rebounding to produce a pressure that is transmitted throughout the bulk of the gas and against the walls of the container; thus, one of the most obvious properties of the gaseous state is that a gas fills completely whatever container it is put in. The Earth's atmosphere, consisting of a mixture of gases, reveals many of the properties of the gaseous state to simple observation. For example, a gas can be compressed, as with a bicycle pump. When a gas is heated, it expands and rises, as over a hot radiator; if a gas is cooled enough, it condenses to a liquid, as with breath on a cold window. The Earth's weather is largely the result of changes in the gaseous properties of air, the most important change being that of temperature, which leads to, among other things, changes in pressure and in the amount of water vapour the mixture can absorb; all clouds, precipitation, storms, winds, and clear skies are phenomena that can be explained in terms of the properties of gases. The compressibility and expansion properties of the gaseous state are utilized in all sorts of ways in industry and in everyday life. For example, whatever fuel is used, the internal-combustion engine, steam engine, jet engine, rocket, and guns are all designed to utilize the properties of the gaseous state. This is also true of the arts of cooking, welding steel, and deep-sea diving, of the performance of aircraft, the operation of windmills, and so on. It is the gaseous properties of oxygen that have determined the evolution of lungs, organs that can handle only gases, not liquids or solids.

It would seem, therefore, that a great deal must have been learned about gases long ago, even by primitive man, but this is not the case. Although gases have been studied experimentally and theoretically since antiquity, any systematic discussion of gases is based upon experimental data collected during the last two centuries and upon theories that began to develop only during the 19th century and that are still evolving. The fact is that it is exceedingly difficult and often impossible to explain the behaviour of gases in terms that allow precise predictions to be made concerning gaseous behaviour under various known conditions.

There are two distinct ways of discussing the gaseous state. First, there is the laboratory description of the behaviour of a bulk volume of gas trapped in a container. This is a macroscopic view, independent of the atomic theory of matter, since it deals with properties, easy to measure, of a vast number of particles, such as the pressure they exert all together on the walls of their container or their temperature as measured by a thermometer. Description in terms of bulk measurable properties is called (phenomenological) thermodynamics and transport theory. Second, there is the particle description of a gas; this is a microscopic view that deals with the properties of the individual particles and how these properties collectively give rise to the observed behaviour of a gas when studied in the laboratory. Description in these terms is called statistical thermodynamics and kinetic theory. During the last 160 years or so, increasingly successful attempts have been made to develop a complete theory that will coordinate these two distinct descriptions.

Macroscopic and microscopic descriptions

Thermodynamic approach to gases

Gases range from simple gases, whose particles are single atoms (monatomic), such as neon or helium, to the complex multi-atom (polyatomic) gases, such as the hy-

drocarbons, obtained in the petroleum industry. The cost of determining all the physical and chemical properties of every gas is prohibitive in terms of time and money, and any procedure that allows predictions to be made of some properties in lieu of actual measurements is of great value. The science of thermodynamics provides general rules that relate some of the behaviour of all gases on the basis of limited data. Of particular importance is the relationship among the pressure, volume, and temperature of a gas, since these are the three properties usually easiest to measure. For example, low pressures are measured by balancing the pressure of the contained gas through a tube against a vertical column of mercury (the height of the column gives an accurate measure of the pressure the mercury exerts on the gas), while for gases at high pressure a free-piston gauge is used, the pressure of the gas being balanced by applying an external force, such as weights of known value, to the piston. Temperature can be measured by a variety of devices, including gas or liquid thermometers, resistance thermometers, and thermocouples. The relationship among pressure, volume, and temperature is called the thermal equation of state. The relationship may be expressed as a precise mathematical equation or as a collection of graphs or tables. The thermodynamics of gases is concerned with the twofold problem of determining the thermal equation of state (see below *Thermal equation of state*) of just a few gases and attempting to predict values for all other gases using the equation; thereafter, the prediction of all of the other properties of a gas can be made in terms of its pressure, volume, and temperature.

THE SIMPLE GAS LAWS

The relationships between the volume of a gas and its temperature and pressure embody easily measured, basic properties of all gases. These relationships have been deduced and put in the form of equations called the gas laws. The first gas law, discovered by the English physicist Robert Boyle and named after him (in France the law is named after Edme Mariotte), states that the volume of a given quantity of gas varies inversely with the pressure exerted on it. That is, if a gas is compressed to half its volume, the pressure it will exert on its container will be twice as much, or, if the pressure on a gas is doubled, the volume will be reduced by one-half; a limiting factor is that the temperature must not change, for, if it does, the pressure-volume relationship will not hold. In symbols, Boyle's law is written: V varies as (or $\propto$) $1/P$, V being the volume and P the pressure. This can be stated as an equation by introducing a constant (depending on the units of volume and pressure used and on the amount of gas) that has been found experimentally: $V = k_1\ 1/P$; that is, the volume equals the product of the constant and the reciprocal of the pressure. The second gas law was worked out (1787) by the French physicist Jacques-Alexandre-César Charles and named after him and also after Joseph-Louis Gay-Lussac, a French chemist who repeated Charles' experiments and published his results in 1802. It states that the volume of gas varies directly as its absolute temperature (see below *Thermal equation of state*), if the pressure is kept constant. That is to say, if the temperature of a gas is increased the volume will increase proportionately, as long as the pressure within the gas does not change. In symbols, V varies as (or $\propto$) T, T being the absolute temperature. In equation form, the relationship is $V = k_2\ T$, k_2 being a constant determined through experiment. In words, then, the volume of a specified amount of any gas expands whether with a rise in temperature or with a reduction in pressure, and the gas contracts if it is cooled or if the pressure is increased, and the changes can be calculated by using one or the other of the two gas laws.

Boyle's law

THERMAL EQUATION OF STATE

If both pressure and temperature are changed at the same time, both equations must be used to calculate the change in volume, pressure, or temperature that is obtained when the other factors are changed. A single equation can be derived from Boyle's and Charles's laws: the product of the pressure and the volume equals the product of the temperature and a constant, R, called the universal gas constant. In symbols the equation is written: $P\overline{V} = RT$, P being the pressure, T the absolute temperature, and $\overline{V}$ the volume of one mole of gas. A mole is a quantity of substance such that its weight in grams is numerically the same as the molecular weight of the substance; *e.g.*, the molecular weight of oxygen (O_2) is 32; therefore, one mole of oxygen is that quantity that weighs 32 grams. The universal gas constant, R, is 1.987 calories per mole degree, a figure found by experiment. The absolute-temperature scale was devised (1848) by the English physicist William Thomson, later Lord Kelvin, on the basis of his work with heat. Charles had already found that all gases shrank $^1/_{273}$ of their volume when they were cooled 1° C (1.8° F) and Kelvin reasoned that if a gas could be cooled 273° C (459.4° F) below 0° C (32° F), the energy of its molecules would become zero. This hypothetical temperature was given the name absolute zero, and the scale of temperatures built on absolute zero, in which each degree was equivalent to a degree on the Celsius thermometer, was called the absolute scale of temperature. On this scale, known as the Kelvin scale, temperatures are designated K, but 0 K, which is equal by definition to −273.15° C (−459.67° F) is still called absolute zero.

The equation is called the thermal equation of state, and it is the most important of the mathematical descriptions worked out for gases, since by simply measuring any two of the unknowns, the third can be calculated. It is applicable to a large variety of gaseous conditions in industry and science, including astronomy.

The thermal equation of state, however, is not applicable to all gases under all conditions. It is an approximation, however, that is certainly accurate for most gases at pressures of less than a few atmospheres and at temperatures between room temperature and several hundred degrees higher.

Ideal gas. Because of this common behaviour pattern, a hypothetical gas, called the ideal, or perfect, gas, has been proposed that obeys the equation exactly under all conditions. The importance of the concept is twofold. First, it provides a criterion of behaviour that most real (as opposed to ideal) gases approach; and, second, there is an exact particle theory for such a gas. The difference between the ideal gas and any real gas is due to several factors, the most obvious being that every molecule has a real volume, and the total volume of the molecules in a container is, therefore, a significant figure when the space occupied by the gas (the $\overline{V}$ in the equation) becomes relatively small; the physical volume of the gas molecules is not taken into account in the equation. Much more important factors that also are excluded concern forces of interaction among atoms and molecules that come into play under various conditions, especially near condensation; at condensation the phenomenon of latent heat (see below *Discrepancies between observed behaviour and the equation of state*) is not taken into account; and, of course, the fact that molecules have different shapes and that their collisions would be affected by these distinct shapes is also not part of the calculations that can be made using $P\overline{V} = RT$. Still another factor enters when the temperature is raised sufficiently to ionize the molecules and then decompose them and, finally, to strip atoms of all their electrons. Such a gas, consisting only of positively charged nuclei and electrons, is called plasma, and its properties, since each particle has an electric charge, differ altogether from those of a normal gas, whose particles are electrically neutral. In passing from normal gaseous state to plasma, therefore, the changes cannot be predicted by the thermal equation of state. The simple gas laws would be absolutely correct for a gas if each molecule were simply a mathematical point in space.

Significance of the concept

In addition to pressure, volume, and temperature, another important thermodynamic property of a gas is its energy. The change in energy with changing volume of any gas that is kept at a constant temperature can be calculated from thermodynamic theory, using the same thermal equation of state. The dependence of the energy on temperature is known as the heat capacity of a gas, and this must be determined experimentally when the gas

is held at constant volume or pressure. A knowledge of the thermal equation of state and the heat capacity allows all of the other thermodynamic properties of many gases to be deduced without further measurement.

It took centuries for these principles to be discovered. Historical highlights were the speculations of Francis Bacon (1620) that heat was not an indestructible fluid called caloric, which everyone believed could enter or leave a gas, but rather that heat was some form of motion of particles. Boyle and Charles developed the concept of the thermal equation of state, while Joseph Black, an English chemist, established (1770) the idea of a heat capacity and energy of a gas in his studies of gas calorimetry. The American Benjamin Thompson, later Count von Rumford (1798), and the Englishman James Prescott Joule (1845) finally disproved the caloric theory of heat. The laws of thermodynamics, which interrelate these properties, were discovered and formulated during the 19th century.

Early speculations

Discrepancies between observed behaviour and the equation of state. If a graph is plotted of the mutual changes in pressure, volume, and temperature of a gas, then a three-dimensional surface is obtained. Such a graph for any laboratory gas being manipulated in a container shows that, for a certain pressure and temperature, the gas will change its phase; *i.e.,* the substance can appear as a gas or a liquid or a solid. It is not easy to picture a three-dimensional surface, and it is usual to make two-dimensional projections of this surface onto a piece of paper. There are three such projections, and the most important one is the graph of pressure against volume. This will be a family of curves such that each member of the family will be a graph of pressure versus volume for a gas kept at constant temperature. Historically, laboratory experiments showed that for some gases, like oxygen and nitrogen, if the temperature were kept at room temperature, then no matter how much the gas pressure was increased (and the volume thereby decreased), the gas never condensed into a liquid; *i.e.,* the pressure-versus-volume curve was a smooth one, with no sharp jumps or kinks in it. On the other hand, for many gases kept at room temperature, as the pressure was increased, a discontinuity in behaviour occurred, and the gas started to condense into a liquid. During condensation, the pressure and the temperature remained constant, while the volume of the mixture of gas and liquid decreased until all of the gas had turned to liquid. Thereafter, a further increase in pressure (still maintaining a constant temperature) resulted in a slight decrease in the volume of the liquid. In addition, as the gas condensed, it released energy, even though its temperature remained constant. This is the phenomenon called the latent heat of condensation, which was discovered by Black (1757) and studied systematically by Black and a colleague (1764), in the case of water vapour condensing into liquid water.

In spite of these early studies of condensation, prior to the work of the Irish physical chemist Thomas Andrews (1869), the conditions necessary to liquefy a gas by compressing it at constant temperature were not clear. Andrews established the existence of a critical-temperature curve, the significance of which was that if the actual temperature of the gas being compressed were greater than this critical temperature, then no amount of compression would liquefy the gas. If, however, the gas were compressed at a constant temperature below the critical temperature, condensation would always occur. If the temperature of the gas is kept just below the critical temperature as the gas is compressed, then the volume change during condensation is small, while, if the gas being compressed is kept at just the critical temperature, then the gas passes directly into the liquid without any change in volume at all. The name critical point is given to that point on the constant critical-temperature curve that represents the termination of any observable distinction between the gaseous and liquid phases. Neither vaporization nor liquefaction, in the sense of boiling or droplet condensation, occurs in a substance having a temperature greater than its critical temperature.

Critical point

An important problem that arises as a gas is compressed at constant temperature is whether the gas obeys the thermally perfect equation of state. In general, this equation is found to hold to within 1 percent as the gas density (which is the mass of gas contained in a fixed volume or, more importantly, the actual number of gas particles in a fixed volume) rises from the density of the gas at room temperature and pressure up to about 100 times this value. Thereafter, deviations from the predictions of the thermally perfect gas equation of state appear, until eventually the equation becomes completely wrong when the gas starts to condense, since during condensation the pressure and temperature remain constant, and the volume decreases.

EMPIRICAL EQUATIONS FOR SPECIFIC GASES

The problem of determining the thermal equations of state of a real gas (as opposed to the hypothetical thermally perfect gas) is an old one. The first serious attempt was made by the Dutch physicist Johannes Diederik van der Waals (1873), who formulated a thermal equation of state containing two adjustable parameters (arbitrary constants that characterize certain variable factors in a system) that could be found by comparing experimental values of pressure, volume, and temperature measured in the laboratory with theoretical values containing the adjustable parameters predicted by his equation of state. There are now well over 100 empirical thermal equations of state available for a whole range of gases. Of the two best known, one contains five adjustable parameters, and the other contains eight adjustable parameters. These complicated equations can reproduce the complete range of observed pressure, volume, and temperature behaviour from the gas phase through condensation into liquid phase for a very wide range of gases. They are very useful for interpolation—*i.e.,* predicting values intermediate between measured values—but they suffer, as do all other empirical thermal equations of state, in that the adjustable parameters are not related directly to any aspect of the particle behaviour of the gas.

The particles of any real gas are surrounded by fields of force, so that the particles interact with each other, even when separated in the gaseous phase. These interparticle forces (called intermolecular forces) are the whole essence of real-gas behaviour, and any thermal equation of state must allow for these forces. Intermolecular forces do not include gravitational force, because, for the small masses of gas that occur in containers of laboratory or industrial size, gravitational force between particles is negligible. Gravitational effects in gases are important only when discussing the structure of the Earth's atmosphere or of interstellar events.

Intermolecular forces

There is only one way of writing the thermal equation of state of a real gas that will both fit laboratory data and also show the effect of intermolecular forces. This is called the virial equation of state, and it is derived from the thermal equation of state for an ideal gas by taking into account the attraction or repulsion between the molecules. It is written as a mathematical series with a very large number of terms:

$$\frac{P\bar{V}}{RT} = 1 + \frac{B(T)}{\bar{V}} + \frac{C(T)}{\bar{V}^2} + \frac{D(T)}{\bar{V}^3} + \dots \qquad (21)$$

The quantities $B(T)$, $C(T)$, $D(T)$, etc., are called the second, third, and fourth virial coefficients, the first virial coefficient being unity. They are all independent of the gas pressure and density and depend only on the temperature. Viewed as a purely empirical thermal equation of state, the virial equation is not particularly convenient, because, at moderately high gas densities, many terms must be added up, and, at very high densities, such as those that occur when the gas is condensing into a liquid, the equation may actually diverge and, therefore, be useless. In spite of this, the virial equation is of the utmost importance, because it forms a link between purely macroscopic (*i.e.,* bulk) properties, such as pressure, volume, and temperature, and microscopic properties, such as the intermolecular forces between the particles.

Because of difficulties associated with the virial equation of state for the case of dense gases during condensation, empirical thermal equations of state are still used today. Of particular value is the principle of corresponding states or phases, a law stating that all substances obey the same

thermal equation of state, provided the equation is written in reduced variables; *i.e.,*

$$\frac{P\bar{V}}{RT} = \text{Function}\left(\frac{P}{P_{\text{crit}}}, \frac{T}{T_{\text{crit}}}\right). \quad (22)$$

This means that the product of pressure and volume, divided by the product of the universal gas constant and the absolute temperature, equals not unity, as would be the case with an ideal gas, but some mathematical function that should be the same for all substances. The reason why the actual pressure and temperature of the gas (on the right-hand side, which corrects the discrepancy between ideal and real gas) are divided by the critical pressure (P_{crit}) and critical temperature (T_{crit}) is that, as already noted, the critical-point values are unique for every substance, and, therefore, they represent a reference point closely related to the condensation of a gas.

The particle-description approach to gases

Although thermodynamics provides a very general and very powerful discipline for determining the interrelationships among the many physical and chemical properties of gases, it is quite unable to predict the actual numerical value of these properties. It can predict, for example, the relationship between the constant-pressure heat capacity and constant-volume heat capacity of all gases in terms of the thermal equation of state, but it cannot predict the numerical value of either heat capacity of any gas. Calculations of the numerical values of the properties of gases require a microscopic theory based on the fact that all gases are composed of particles that may be monatomic (single atoms) or complicated polyatomic molecules. For the normal range of temperatures encountered in industrial applications (but not aerospace applications), all atoms may be regarded as particles possessing no internal structure; *i.e.,* they are considered to be solid particles. When atoms are bound together into polyatomic molecules, however, these molecules do possess internal structure; for example, the simplest polyatomic molecules consist of only two atoms bound together. Such a structure can rotate end over end and, therefore, possesses an internal mode of motion called molecular rotation. In addition, the two atoms may vibrate as if connected by a spring, so that the molecule also possesses another internal mode of motion, called molecular vibration. More complicated molecules can possess more complicated rotations and vibrations.

The vast numbers of gas particles

The use of any laboratory container of gas involves many billions (10^9) of particles; for example, one cubic metre of air at standard conditions contains more than 10^{24} particles. Every one of these particles is moving throughout the container in random fashion, and this is called the external mode of motion, or the random thermal translational motion of the particle. It is called thermal because the intensity of this motion bears a direct relationship to the temperature: the greater the thermal energy of the particle the faster it moves through space. In addition, if the particle is polyatomic, it also possesses internal modes of motion, such as rotational and vibrational motion. In principle, all of these motions, both internal and external, are governed by the laws of quantum mechanics. In practice, however, the translational motion can often be accurately described by the much simpler laws of Newtonian mechanics. Such a simplification does not usually hold for the internal motions, but in some cases it is valid for the rotational motion. Consequently, a discussion of the quantum internal motions may be expected to be more involved than the Newtonian external motion of translation. For gases at densities near atmospheric density, all of the particles spend most of their time moving with a constant speed, called the thermal speed, just as if they alone occupied the container; *i.e.,* any particle behaves as if it were alone in its container, notwithstanding the enormous number of other particles also present. At any instant of time, however, about 1 particle in 100,000 will be colliding with another particle or the container walls, and, in a time interval of one second, each particle of the gas will experience about 1,000,000,000 collisions. A collision between two particles is called a binary collision, and in such a collision both particles may change their speeds and the directions in which they move.

DISTRIBUTION FUNCTIONS

The microscopic picture of a gas, then, is one of billions of particles in incessant thermal motion (a manifestation of its thermal energy), each particle now and again suffering a collision as it wanders randomly throughout its container. Clearly, it would not be sensible to try to follow the behaviour of each particle, even if these could be seen, which they cannot. In fact, if the position in space and the speed of each of a million particles could be written down every second, it would still take the age of the universe to write down the same information for all of the particles in a normal-sized room, and this information would be true only for an instant of time. Actually, this is no loss, because the behaviour of individual particles is not interesting. All that is really of importance is their macroscopic behaviour; *i.e.,* their behaviour averaged out over the vast number of particles in the container. Mathematical statistics is able to handle large numbers, and the procedure for dealing with the problem is to examine the behaviour of groups of particles, where each group consists of a very large number of particles that all possess some common properties. The common properties used are the positions in space and the speeds of particles, and this leads to the concept of what is called a distribution function. The distribution function is a mathematical equation that indicates how many of the vast number of particles in the container have spatial positions within some arbitrarily chosen, exceedingly small region of the container as well as thermal speeds that lie within an exceedingly small, arbitrarily selected velocity range.

More precisely, a distribution function gives the probable number of such particles, the word probable being a reminder that it is impossible to say exactly how many particles even an exceedingly small region of the container actually contains at a given instant; the reason is that the number continually varies about its most probable value as a result of the random arrival and departure of individual particles, so that there is a higher or lower particle density in the minute region than the average density, which can be calculated with great precision by dividing the total number of particles by the total volume of the container. These variations, or fluctuations, in density arise as an inevitable consequence of the particulate nature of a gas and lead to observable effects in the laboratory.

FORCES AFFECTING DISTRIBUTION FUNCTION

A knowledge of the distribution function is essential for any microscopic theory of a gas; however, its determination involves postulating some mathematical description of a gas particle and how a gas-particle collision actually occurs. In any real gas, even a binary collision—one between two particles only—is a complicated process, because each particle is surrounded by a force field that varies in space. This force field is such that it attracts another particle when that other particle is relatively far away but repels it when the other particle comes relatively close. That these intermolecular forces exist is evident from the fact that gases do condense into liquids, a phenomenon that cannot be explained except as a consequence of attractive forces at work between the particles. In addition, liquids strongly resist being compressed, which must be evidence of the repulsive forces. The long-range attractive forces, called van der Waals forces, arise from electrostatic interactions between the whole particles. Short-range forces arise when the electron clouds that surround the nuclei of atoms in a molecule overlap. These short-range forces are called valence forces, and they become stronger than the van der Waals forces when the two particles come sufficiently close together in a binary collision.

Van der Waals forces

Multiparticle collisions. So far, the discussion has been concerned with gases at pressures near atmospheric pressure and density near that of normal atmosphere, a condition in which binary collisions alone are significant. As the density of the gas is increased, there is the possibility of three particles coming together at the same place and colliding in what is called a ternary collision. Perhaps even

four particles and, in dense gases, an even higher number of particles may come together in a collision. To include ternary and higher order collisions in the discussion, the nature of the intermolecular forces among the various colliding particles must be known. It is just this specification of the forces among collision partners that gives rise to the various mathematical theories of a gas. It has been shown that the most complete and satisfactory theory is the one designed for a dilute gas, in which ternary and higher order collisions can be completely ignored. The dilute gas is not the same as the thermally perfect gas, a distinction that will be discussed later. If a dense gas is under consideration, then ternary and higher order collisions must be allowed for; the theories for dense gases are still incomplete.

Theory of binary collisions in dilute gas. It has been stated that every particle spends most of its time moving freely with a constant speed between collisions throughout its container. Of course, any such collision may, and usually does, change the speed of each particle; each retains this new speed undiminished until the next collision. Consequently, only the single-particle distribution function is important, since the laboratory properties of the gas will not depend upon the relative spatial positions and speeds of clusters of two or more particles but only upon the spatial position and speed of particles taken one at a time. Although collisions between particles in a dilute gas are infrequent as far as any particle is concerned, they must be allowed for. If the gas consists of monatomic, structureless particles, then, in a binary collision, the mass, momentum, and thermal kinetic energy (energy due to random motion) of the collison partners remain constant. This is known as an elastic binary collision. If the particles are polyatomic molecules possessing internal structure and internal motions, then, in a binary collision, although mass and momentum remain constant, it is possible that some of the thermal kinetic energy of the collision pair may be transformed into internal energy of the molecules by increasing the amount of energy already involved in the internal motions of rotation and vibration. This is called an inelastic binary collision.

Boltzmann transport equation

It is not difficult to derive the mathematical equation, called the Boltzmann transport equation, that describes the way the single-particle distribution function changes with time as a result of binary collisions, but it is difficult to solve this equation. Although the Boltzmann equation lies at the very heart of transport theory, it cannot be solved in general terms but only under special conditions, which, fortunately, are not too restricted. This discussion is concerned with two solutions of the Boltzmann equation, the equilibrium solution and the near-equilibrium solution.

EQUILIBRIUM SOLUTION FOR THE BOLTZMANN EQUATION

The equilibrium solution is considered first. Mathematical analysis shows that if no external forces act on the gas, then the distribution function tends toward a value that is independent of spatial position and time, a value called the Maxwellian translational distribution function of thermal speeds. The existence of this value means that the gas reaches a uniform steady state known as a state of thermodynamic equilibrium. In this equilibrium state, all of the laboratory properties of the gas—pressure, density, temperature, energy, heat capacity, and entropy—are constant, and, in addition, the pressure, temperature, and density have the same values throughout the entire volume of the gas. Entropy is a measure of the disorganization of a gas. In any gas there is always some of the energy that can be converted into work (which is usually what an engineer wants); the amount that can be converted is determined by the entropy of the gas (see ENERGY, THE CONCEPT OF). The equilibrium state is achieved because there is an overall balance in the collision processes suffered by the particles; *i.e.,* the number of particles with any particular spread, or range, of thermal speeds remains the same anywhere in the container because of a balance between the numbers of particles gaining and losing a given change of thermal speed due to binary collisions. In addition, further analysis (known as Boltzmann's H-theorem) shows that there is a detailed balance in the binary collisions, so that the number of particles in a particular thermal-speed spread, which are changed by binary collisions of a specified type, is exactly balanced by the same number of inverse collisions; if, for example, a collision is speed depleting for one of the particles, then the inverse collision is speed replenishing for the same particle. An inverse collision is distinguished from a reverse collision, in which the two particles merely reverse their original paths in space. The existence of the Maxwellian translational distribution function of speeds has been confirmed by direct measurement using a molecular beam. (Such a beam is produced by allowing gas particles to escape from a minuscule hole in the walls of a heated tube into a space kept at very low pressure by vacuum pumps; the individual particles fall in the earth's gravitational field and are separated out according to their thermal speeds.)

Maxwellian speed distribution function. The Maxwellian speed distribution for a dilute gas is, in fact, a general result and holds for dense gases as well. Once the translational distribution function (Maxwellian) of a gas has been obtained, it is a simple procedure to calculate the many equilibrium properties of the gas, such as the number of particles colliding with the container walls every second, by allowing for particles coming from all directions and with all thermal speeds. The result can be used to calculate the average force that the particles exert on the walls as they rebound from it, and that force is a measure of the pressure of the gas, since the collisions are sufficiently numerous to be thought of as a continuous force on the walls. Intermolecular forces affect the calculated pressure, because the attractive van der Waals forces exerted by the particles within the body of the gas restrain particles that are just about to collide with the wall. In addition, the mean free path of a particle can be calculated as the average distance travelled by a particle between binary collisions. A particle interpretation of the temperature of a gas can also be obtained in terms of the average kinetic energy of the particles as they move randomly throughout their container.

Equilibrium distribution function for internal energies. The Maxwellian distribution function refers to the external motions of the particles but does not refer to their internal motions of rotation and vibration. It might be expected that just as there is an equilibrium distribution function for the translational random kinetic energy, so there would be an equilibrium distribution function for the internal energies as well. In an evaluation of this equilibrium internal energy distribution function, a mathematical term arises that is called the internal partition function of the gas. The partition function, which is a mathematical series, has no immediate direct physical meaning but is intimately related to the manner in which the total energy of the gas particles, due to their structure together with the energy involved in the effects of intermolecular forces, is divided up (partitioned) among the particles. It is here that kinetic molecular theory merges into statistical thermodynamics, because the basic problem of the latter is to decide just how the total energy of the gas in equilibrium is to be divided up among the various energy modes of the particles. The analysis of the partitioning of energy examines the infinite number of ways the energy can be distributed and seeks to find if there is a preferred, or most probable, partitioning that is overwhelmingly likely to occur. The results of the analysis are expressed in terms of the partition functions of the various energy modes of the gas particles, which, as has been said, are very closely related to the equilibrium distribution functions of kinetic theory. The evaluation of the internal partition functions for thermally perfect gases is relatively easy, and it is possible to calculate all of the thermodynamic properties (*e.g.,* energy, heat capacity, entropy, etc.) of a thermally perfect dilute gas of particles possessing internal structure. If, however, the gas is one of interacting particles, then the evaluation of the internal partition function can be difficult, especially if the gas is dense, requiring an evaluation of the effect of ternary and higher order collisions.

Partition function

LINK BETWEEN THERMODYNAMIC AND MICROSCOPIC PROPERTIES

The partition function provides a direct link between the microscopic properties of the particles—their quantum energy states, their moments of inertia, their intermolecular potentials, etc.—and the laboratory properties of the gas—the thermal equation of state, the thermodynamic energy, heat capacity, and entropy. If the thermal equation of state of a gas of interacting particles is evaluated from the partition functions, then it turns out to be the virial (average kinetic energy) equation of state already introduced above. Each virial coefficient involves a mathematical quantity called the cluster integral. The second virial coefficient involves that cluster integral relating to the binary collision, while the third virial coefficient involves the cluster integral relating to ternary collisions, etc. Consequently, to evaluate the cluster integral, the dynamics of the collision must be calculated for a known intermolecular potential. For a dilute gas, where binary collisions alone are involved, the second virial coefficient can be evaluated for a variety of intermolecular potentials (see below *Mathematical models: virial coefficients*). To calculate the higher order cluster integrals, it is customary, because of ignorance of the true behaviour of three or more mutually interacting particles (called the many-body problem), to make the assumption of pairwise additivity: this states that the intermolecular forces involved in a many-particle collision are made up of the sum of two particle forces; for example, if three or more particles are interacting, then their interaction energy is the sum of their energies obtained by taking them in pairs. Although the assumption must be made in order to evaluate the cluster integrals, it is only approximate. Consequently, the third virial coefficient can be evaluated only approximately, which is the case for higher virial coefficients as well.

The many-body problem

The ideal gas versus real gas. The relationship between the hypothetical thermally perfect gas, discussed earlier, and a real gas is that the former is a special type of the latter. The thermally perfect gas has particles that are not surrounded by attractive intermolecular force fields, so that its particles do not interact when separated. They can be thought of as rigid particles that interact with other particles only at the instant of collision, when, on touching, they repel each other with infinite force. Any gas of rigid particles will show thermally perfect behaviour provided its density is sufficiently low so that the second and higher virial coefficients are all extremely small; if the particles are "point" particles of vanishingly small volume, then all of the virial coefficients are zero, and the gas is thermally perfect at all densities. It is for this reason that the internal-partition functions for thermally perfect gases can readily be evaluated.

Origins of particle theory of gases. The particle description of gases grew from ideas in the 17th and 18th centuries concerning special aspects of the collective behaviour of structureless gas particles that were supposed to have no surrounding force fields. The pressure exerted by such a gas on the walls of its containers was calculated, and the idea developed that temperature was related to particle motion. In the 19th century, the heat capacity of a gas kept at constant volume was calculated and found not to agree with experimentally measured values. When the heat capacity at constant pressure was calculated for the same gas, agreement was obtained with experimental values. The ratio of the two heat capacities was also calculated, and the idea was introduced that gas particles that possessed internal structure also had internal energy modes, such as molecular rotation and vibration. The idea of mean free path and the concept of binary collision were introduced. The first formal mathematical attempt to calculate the single-particle distribution function of a dilute gas was made by Maxwell in 1859 and refined by Boltzmann in 1868. The work of Maxwell and Boltzmann led directly to the classical (Newtonian) statistical thermodynamic theory of dilute gases of structureless particles in thermodynamic equilibrium in terms of the external, or translational, partition function. The idea of imperfect gases, real gases whose particles are surrounded by force fields, goes back to Newton. The first correct attempt to analyze them was made in the 18th century, but it was van der Waals (1873) who first deduced the thermal equation of state of a real gas. The German physicist Rudolf Clausius developed (1875) a general theorem that gives the relationship between the intermolecular forces and the thermal equation of state, but it was not until 1902 that the virial equation of state and the virial coefficients were introduced. The statistical thermodynamic theory of imperfect gases was formulated in general terms by the American physicist J. Willard Gibbs (1901) using classical Newtonian mechanics. The previous year, the German scientist Max Planck discovered energy quantization, but it was not until later that the quantum statistical thermodynamics of real gases was developed by Satyendra Nath Bose and, independently, by Albert Einstein in 1924 and Enrico Fermi and Paul A.M. Dirac in 1926.

Mathematical models: virial coefficients. Various mathematical models have been suggested for the intermolecular force between two particles. The commonest assumption is one that is made for the sake of mathematical simplicity, and it assumes that the particles are both essentially spherical, so that the force between them does not depend upon their orientation in space (the intermolecular force has no angular dependence and is spherically symmetrical) but only upon their distance apart. Intermolecular forces are more often discussed in terms of the intermolecular potential energy between the particles. The simplest spherically symmetrical intermolecular potential is the rigid-sphere potential, which states that the particle is a solid sphere of prescribed size but not surrounded by an attractive force field. Such an intermolecular potential belongs to the thermally perfect gas. Among the many spherically symmetrical intermolecular potentials that have been proposed, the commonest and most useful is called the Lennard-Jones potential, which accurately describes the attractive van der Waals forces and, in addition, includes a term that gives the repulsive valence force, chosen for mathematical convenience and also because it is in good agreement with experiment. Virial coefficients have been calculated for many such intermolecular potentials, and it is found that they are more sensitive to the attractive force than to the repulsive one, so that any comparison between measured and calculated virial coefficients should yield information on the attractive van der Waals forces. The results of all of these calculations show that all virial coefficients should be negative at low temperatures, becoming zero at some intermediate temperature and rising to a positive maximum value at high temperature. Consequently, a determination of the pressure, volume, and temperature behaviour of a gas over a wide range of values of these properties should yield experimental data that can be interpreted in terms of the values of the virial coefficients. Experimentally, only the second and third virial coefficients can be measured, since the others are too small. In addition, the intermolecular potential cannot be determined uniquely merely by making laboratory measurements on pressure, volume, and temperature. An indirect procedure must be used, by assuming a mathematical equation for the intermolecular potential that contains unspecified constants, then calculating the virial coefficients and their dependence on temperature and then comparing experimental and theoretical results in order to determine the unspecified constants. Thereafter, the intermolecular potential, now with known constants, can be used to calculate other measurable gas properties that also depend upon the virial coefficients, such as the Joule-Thomson coefficient, the heat of sublimation of crystals, and the transport coefficients (see below *Transport coefficients*).

Method of determination

The ability of the intermolecular potential to predict the experimentally observed values of these properties is a good test of its accuracy. Unfortunately, the largest virial coefficient (the second) is not very sensitive to the precise form of the attractive intermolecular force, while the third virial coefficient suffers from uncertainties about the effects of assuming pairwise additivity and also the effects of experimental error, since it is so small. Because of these limitations, any comparison of theoretically calculated and

experimentally observed virial coefficients seldom eliminates an assumed form of intermolecular potential; rather, it gives the best form from a group of assumed forms on the basis that the best form fits the widest range of properties.

Virial equation of state and dense gas. It had been hoped that the virial equation of state would be applicable to dense gases and even to condensing gases, provided that a sufficient number of virial coefficients were taken into consideration. Several difficulties have arisen, however. For example, for ternary and higher order collisions, the intermolecular potential must be guessed on the basis of pairwise additivity, and this means that the cluster integrals for many-particle collisions are only very approximately known and may be seriously in error. Also, during condensation, when an enormously large number of particles are mutually interacting, the cluster integrals become dependent on the volume of the gas as the clusters get bigger and bigger. This is a branch of statistical thermodynamics that is not well understood and is still under active investigation. It seems clear that a virial expansion up to the fifth virial coefficient is mathematically convergent for densities as high as the density of the saturated vapour, that portion of the gas still uncondensed; however, the virial expansion is mathematically divergent and, therefore, useless for the density of the liquid. Consequently, the virial expansion cannot give a theory that goes right through the condensation transition from entire gas to entire liquid. There is no complete theory of condensation available at the present time. Ultimately, the virial expansion must fail altogether at very high densities, when the concept of the existence of individual interacting particles loses its meaning. This has led to another approach to the thermal equation of state of extremely dense gases and of liquids where a structure is assumed and where the question of how this structure arose is ignored.

Latent-heat theory. Another phenomenon that a theory of condensation should predict is the latent heat; namely, the release of energy from a gas that is condensing as its temperature remains constant. There is no theory for the latent heat, because it is intimately related to the condensation process, and, at the present time, knowledge of latent heats is dependent on thermodynamics. An equation called the Clausius-Clapeyron equation enables the latent heat of condensation to be deduced from knowledge of the rate of change of vapour pressure with temperature, together with knowledge of the densities of the liquid and vapour in the mixture. The latent heat depends upon the pressure and temperature at which condensation occurs, and it becomes zero at the critical point. The Clausius-Clapeyron equation, being a thermodynamic equation, is independent of the particle theory of matter.

Historically, specific forms for the intermolecular potential were studied by Maxwell (1866) and others, but it was not until 1924 that the effects of both attractive and repulsive forces were considered. The physical explanation of why repulsive forces arise was given in 1927, and in 1930 the van der Waals attractive forces were explained. A virial-equation approach to the problem of the condensation of a gas into a liquid was pioneered in 1927 and considerably extended in 1937.

NONEQUILIBRIUM SOLUTION FOR THE BOLTZMANN EQUATION

Turning aside from the equilibrium solution of the Boltzmann transport equation, which describes the state of thermodynamic equilibrium of any gas, be it dilute or dense, the case can now be considered where the gas is in a nonuniform state. If there are variations (*i.e.*, gradients) of density, gas velocity, or temperature in the gas, then transport processes arise. This means that some physical property of the gas flows through the gas down the gradients. If the property transported is mass, then the transport process is called diffusion, and it arises because of a gradient in the density of particles between any two regions of the gas. If the property transported is momentum, then it is viscosity that arises from gradients of gas velocity in the gas, such as will occur if the gas is in motion along a wall. If the property transported is thermal energy, then the transport process is called thermal conduction, and it arises if there are regions of different temperature in the gas. The three transport processes are characterized by three transport coefficients.

Properties transported

For a dilute gas in which only binary collisions need be considered, the governing equation is the Boltzmann transport equation; it gives the space and time behaviour of the single-particle distribution function. The entire equation can be mathematically manipulated to give certain mathematical relationships called the flux vectors, which contain the transport coefficients. To evaluate the flux vectors and, therefore, the transport coefficients, two things must be known; namely, the single-particle distribution function of the nonuniform gas and, because the transport processes are related to binary collisions between particles, details of the dynamics of a binary collision, which, as has been pointed out, require a knowledge of the intermolecular potential.

TRANSPORT COEFFICIENTS

It has been remarked that the Boltzmann transport equation for binary collisions can be solved only under special conditions: the effective diameters of the particles must be much less than their mean free path, and the changes in density, gas velocity, and temperature over a mean free path must be extremely small. Under these conditions, the Boltzmann equation can be solved (the Chapman-Enskog solution) by using the method of successive approximations; it yields a solution for the nonequilibrium distribution function in the form of a mathematical series. Each term in the series represents an increasing departure from the Maxwellian equilibrium distribution function, so that the nonequilibrium distribution function is expressed in terms of the localized density, gas velocity, and temperature values, all of which vary throughout the volume of the gas and also vary with time. The first approximation to the solution is one that is valid for small spatial gradients and gives a measure of the number of particles transferred, or the flux vectors, that depend on these gradients, so that the transport coefficients can be calculated after the intermolecular potential has been specified. In this respect, the calculation of the three transport coefficients is completely analogous to a calculation of the virial coefficients of the equilibrium gas; that is, the intermolecular potential is specified; the collision integral, a mathematical quantity relating the angle of deflection of one particle with respect to the other involved in the binary collision, is calculated; and, thereafter, the flux vectors are calculated, and the transport coefficients are obtained in terms of the assumed intermolecular potential.

Chapman-Enskog solution

Transport coefficients have been calculated for a variety of spherically symmetrical and other types of intermolecular potentials. Such calculations show that, although the transport coefficients are more sensitive to the repulsive than to the attractive part of the force between two particles (whereas the virial coefficient is just the converse), they depend essentially upon the fact that there is a repulsive force and not much on the precise form of the force. As with the virial coefficients, the form of the intermolecular potential that agrees with laboratory measurements of the transport coefficients cannot be deduced.

Transport coefficient for dilute, polyatomic gas. The Chapman-Enskog solution of the Boltzmann equation was originally framed to deal with dilute, monatomic gases of structureless particles where all binary collisions are elastic. For the commoner case of polyatomic molecules, allowance must be made for inelastic collisions between nonspherical particles, when random kinetic energy is converted into internal energy of rotation and vibration. A Chapman-Enskog type of solution has been given to the Boltzmann equation for a dilute, polyatomic gas, but it involves details of the intermolecular potential in an inelastic binary collision that are still unknown. A simplified version of this theory in which the inelastic collisions are dealt with in terms of rotational and vibrational relaxation times (see below *Thermal relaxation stage*) has been given, and studies show that inelastic binary collisions do not appreciably affect the diffusion and viscosity coefficients but do significantly affect the thermal-conduction coeffi-

cient. Such a result would be expected, since the thermal-conduction coefficient deals with energy transfer through the gas, and inelastic collisions are concerned with energy-transfer processes between particles. The problem of calculating the transport coefficients of dilute, polyatomic gases is still under investigation, and much remains to be done.

Transport coefficient for nonchaotic gas. When the gas density is more than about 100 times atmospheric density, ternary and higher order collisions must be considered in addition to binary collisions, and this consideration affects the transport coefficients profoundly. The Boltzmann transport equation and its Chapman-Enskog solution are not applicable to the dense gas, because what may be called molecular chaos may not exist. Molecular chaos, which must hold if the Boltzmann equation is to be applicable, requires at all times that the spatial positions and speeds of the particles involved in collisions be in no way related to each other (they are not correlated), so that there is no effect of previous collisions on any collision occurring, which means that the results of a collision must depend upon the immediate state of the particles during the collision and not on their previous history. Molecular chaos is readily achieved in a dilute gas, because particle correlation is negligible, since a recollision between the same two particles is extremely unlikely. In a dense gas, a substantial proportion of all particles are suffering collisions at any instant, and such collisions may be recollisions, because the mean free path becomes comparable with the particle diameter (in a dilute gas, the mean free path is very large compared with the particle diameter, and most of the time particles in a dilute gas are travelling freely). In such a recollision, spatial positions and speeds may be correlated, and, thus, molecular chaos—and, therefore, the Boltzmann equation—no longer holds. What is required is a multiparticle distribution function that allows for particle correlation (*i.e.,* the relative spatial positions and speeds of clusters of two or more particles), recognizing the fact that any one particle can affect the spatial positions and speeds of all of the other particles in its vicinity.

Recollisions

DISTRIBUTION FUNCTION FOR N PARTICLES

It is possible to write down the equation (the Liouville equation) that gives the changes in space and time of the N-particle distribution function, N being the total number of gas particles in the container, a total that is, as has been discussed, an enormous number. From the equation it is possible to derive, by mathematical methods, a chain of equations in which each equation in the chain involves the changes in space and time of two consecutive lower order (*i.e.,* fewer particles) distribution functions.

At the end of the chain, there will be an equation that gives the single-particle distribution function used in the Boltzmann equation, in terms of the two-particle distribution function. The chain of equations is called the BBGKY hierarchy of equations. For purposes of calculation, the chain of equations must be broken, since it is impossible to handle a vast number of equations at the same time, by making some assumption about the dependence of some higher-order distribution function on the lower-order distribution functions and then dealing with a finite, as opposed to an effectively infinite, set of chain equations.

One method of doing this is based on a realization that if a gas achieves a state of thermodynamic equilibrium, or a nonequilibrium steady state with unchanging gradients of properties, solely by collisions between particles—be they collisions of second or of higher order, elastic or inelastic—then the process of achieving that state will require a finite time because such collisions do not occur infinitely fast. The problem of calculating the uniform state of thermodynamic equilibrium or nonuniform steady state of a gas is still unsolved and represents a great challenge to present-day research workers. Any progress that has been made has been based on the recognition that, during this approach to a final state, several quite different time scales are involved. The first time scale is that; appropriate to the initial stage. This time scale is the duration of a collision; that is, the quotient of the effective range of the intermolecular force field between two colliding particles and their average random thermal relative speed; it has a value of about 10^{-13} second. Analysis of the initial stage is a dynamic problem dealing with the detailed motions of individual particles, and the gas cannot be described in terms of its laboratory properties, such as density, gas velocity, and temperature. The next characteristic time interval is that between collisions, about 10^{-11} of a second. In between the two times, the gas is being brought into a state where a statistical description is possible in terms of distribution functions but where a laboratory description is still not yet possible in terms of density, gas velocity, and temperature. This may be called the kinetic stage; and the gas must be described by its multiparticle distribution function. After a time equal to the duration of a collision—*i.e.,* after the initial stage—all the multiparticle distribution functions have changed so rapidly that they can all be described in terms of the single-particle distribution function, which effectively has not changed at all. Assuming such a simplification has enabled researchers to solve the BBGKY chain of equations that describe the kinetic stage and that allow, in principle, for multiparticle collisions before molecular chaos is established. Unfortunately, the solution is a mathematical series that diverges and, therefore, is useless when four-particle and higher-order collisions are considered. It would appear that the method (named Bogolyubov) is valid for gases having densities such that binary and ternary, but no higher order, collisions occur.

Time scales

This is as far as the kinetic theory of dense gases has been developed; a completely different approach may be required, one that does not treat a dense gas as a gas in which multiparticle collisions occur in terms of the binary, ternary, and higher-order collisions.

Special conditions

HYDRODYNAMIC PROPERTIES OF A GAS

The initial and kinetic stages in the process by which a gas approaches an equilibrium state or a steady nonuniform state take place so rapidly that they are, to all intents and purposes, beyond laboratory measurement. In view of the great difficulties involved in developing a theory of these stages, the discussion will revert to a consideration of the behaviour of a gas after they are complete. The next stage in the approach of the gas to its final state, one that does lie within the capabilities of laboratory techniques, is that describable by the Boltzmann equation, involving the single-particle distribution function and its solutions. This is called the hydrodynamic stage, because it can be described by mathematical equations that involve properties such as density, gas velocity, and temperature, rather than the statistical concept of a distribution function. The equations are called the Navier-Stokes and the Burnett equations of hydrodynamics. If the gas is polyatomic, there is another stage, which is the final approach to thermodynamic equilibrium or to a steady nonuniform state; this is the thermal relaxation stage and involves energy exchanges between the external (translational) energy mode and the internal energy modes of the polyatomic molecules.

Hydrodynamic stage. Rather than continue the discussion in general terms, it is easier to consider an important application of the hydrodynamic and thermal-relaxation stages: the passage of a plane sound wave through a gas. Such a sound wave causes any randomly chosen, small volume of gas to be alternately compressed and expanded, so that the density, gas velocity, and temperature of the small volume will vary cyclically with time as the sound wave continues to pass through the gas. The compressions heat the gas, and the expansions, called rarefactions, cool it. Transport processes, especially the transfer of heat and of momentum, occur between the regions of compression and rarefaction, and in each cycle some of the forward-directed energy of the sound wave is converted into random thermal energy of the gas. The energy of the sound wave suffers classical (*i.e.,* Newtonian) absorption and also classical dispersion; and the speed of passage of the wave through the gas depends upon the frequency of the alternations of compression and expansion. The frequencies for which significant sound-wave absorption and dispersion occur are usually higher than the frequencies

Sound waves

of sound waves audible to the human ear, and the waves are called ultrasonic. Such waves disturb the state of thermodynamic equilibrium of the gas, and the rate of return to a state of thermodynamic equilibrium can be measured in the laboratory. This is the hydrodynamic stage governed by the macroscopic equations of Navier and Stokes and of Burnett.

Classical absorption and dispersion (the type just described) is the only type of energy-loss process occurring when a sound wave passes through a monatomic gas, whose particles possess no internal structure. If, however, the gas is composed of polyatomic molecules, much larger absorption and dispersion result from thermal-relaxation processes. Thermal relaxation is a consequence of the fact that the internal energy modes of polyatomic molecules cannot change their energies as quickly as the external (translational) energy mode.

The physical picture is one in which, before passage of the sound wave, the gas is in a state of thermodynamic equilibrium with the total energy partitioned out, in accordance with the partition functions already referred to, between the various energy modes. The sound wave comes along and disturbs this equilibrium, because the compressions increase the total energy of the gas, while the rarefactions decrease the total energy. The change in energy must be distributed between the various energy modes (external and internal) of the particles to allow a new state of equilibrium to be achieved. The redistribution (or repartitioning) requires inelastic collisions to transfer the energy from the external translational mode into the internal energy modes, and such inelastic collisions take a finite time to occur. Evidence obtained from experimental studies of the absorption and dispersion of ultrasonic waves by polyatomic gases suggests that the energy is transferred between the various energy modes by binary inelastic collisions, even for exceptionally dense gases having densities approaching those of liquids. It is found that the translational energy relaxation time (the time taken for the translational energy distribution function to change to its new equilibrium value) is about equal to the time between binary collisions. In addition, the rotational energy relaxation time for all polyatomic molecules (except such simple diatomic molecules as oxygen, nitrogen, and hydrogen) is only slightly longer. The vibrational energy relaxation time, however, may lie between 10^{-8} and 10^{-5} of a second, depending on the molecule concerned. Consequently, the vibrational energy changes in polyatomic molecules cannot keep in step, or in equilibrium, with the very rapid translational and rotational energy changes that occur as the sound wave passes through the gas. To calculate the amount of absorption and dispersion that occurs requires a calculation of the number of binary inelastic collisions occurring in the polyatomic gas every second and knowledge of how this number depends upon the density and temperature of the gas. It is a problem in quantum mechanics involving two polyatomic molecules, not necessarily possessing spherical symmetry, interacting with some complex intermolecular potential that depends upon the spatial orientation of the two molecules, as well as on their separation distance. Such a complicated collision problem can be solved only in rather approximate terms; it is a much more difficult calculation than that needed to evaluate the virial and transport coefficients of monatomic gases. The relative simplicity of the treatment of monatomic gases derives from the applicability of Newtonian mechanics to elastic collisions uncomplicated by the participation of internal energy modes and thermal-relaxation effects.

Energy transfer via binary inelastic collisions

Thermal relaxation stage. Approximate calculations of inelastic collisions show that at any given temperature the relaxation time depends, roughly speaking, on the amount of energy to be exchanged between the translational energy mode and the internal energy mode under discussion. It does not take much energy to set molecular rotations into motion, and many of the binary collisions that occur in a room-temperature polyatomic gas are, in fact, sufficiently energetic to cause molecular rotation (they are inelastic collisions), and this is why rotational relaxation times are not much greater than the time between binary elastic collisions. It usually requires, however, a considerably larger amount of energy to start off molecular vibrations, and exceedingly few of the binary collisions in a room-temperature polyatomic gas are sufficiently energetic to pass translational energy into the vibrational energy modes of the molecules; hence, the vibrational relaxation time is long.

HIGH TEMPERATURES

Since internal energy-relaxation-time measurements are capable of yielding valuable information on intermolecular forces, considerable effort is being given to studies of the temperature dependence of relaxation times. Ultrasonic absorption studies are limited to gases at room temperature and a few hundred degrees above it, so that another technique must be found to study high-temperature gas relaxation behaviour that is becoming so important in modern aerospace technology, as in the flow of a hot gas through a rocket nozzle. Such a technique is the shock-wave method, in which relaxation processes can be studied up to temperatures as high as 1,000,000°. The shock-wave method has been used for studying not only rotational and vibrational relaxation processes in polyatomic gases but also dissociation relaxation processes, in which the molecular vibrations have become so energetic that the atoms of the molecule fly apart and the molecule breaks up. In addition, at higher temperatures, the individual atoms, which previously had behaved as structureless particles, show that they are formed from electrons and nuclei, because electronic excitation of the electrons occurs, eventually resulting, perhaps, in ionization of the atoms, whereby they lose one or more of their electrons.

Shock-wave method

CHEMICAL REACTIONS IN GASES

Dissociation and ionization processes are more commonly thought of as chemical reactions. Polyatomic gases are capable of undergoing a vast number of different chemical reactions where the chemical bonds that bind atoms into molecules and electrons into atoms are broken and the molecules or atoms thereby change their chemical nature. Among the striking features of chemical reactions is the strong dependence of their velocities on density and temperature, indicating that reactions proceed because of inelastic collisions between molecules. Chemical-reaction-rate theory is a branch of kinetic theory, and a quantum kinetic theory of chemical reactions exists that is very similar in description to the kinetic theory of the transport processes in dilute, polyatomic gases and which suffers from all the limitations of that theory. (M.McC.)

PLASMA STATE

The plasma state of matter is unique like the solid, liquid, and gaseous states, and is often considered the fourth state of matter. A plasma is an electrically conducting medium, in general consisting of negatively charged electrons, positively charged ions, and atoms or molecules or both. Its properties depend on the collective behaviour of the particles, as distinct from the individual; and the number of positive charges is almost equal to the number of negative charges. Nearly all the matter in the universe exists in the plasma state, occurring predominantly in this form in the Sun and stars and in interstellar space. Auroras, lightning, and welding arcs are also plasmas; plasmas exist in neon and fluorescent tubes, the crystal structure of metallic solids, and in many other phenomena and objects.

A plasma may be produced in the laboratory by heating a gas to an extremely high temperature, which causes such vigorous collisions between its atoms and molecules that they dissociate into electrons and ions. Plasmas may be partially or completely ionized. A completely ionized hydrogen plasma, consisting solely of electrons and protons

(hydrogen nuclei), is less complex than a mixture of atoms or molecules and so is the most elementary aggregation of matter known. In certain circumstances, a plasma obeys the laws of magnetohydrodynamics, the study of the behaviour of electrically conducting fluids in the presence of electric and magnetic fields.

Relation to laws of magneto-hydro-dynamics

The modern concept of the plasma state is of recent origin, dating back only to the early 1950s. Its history is interwoven with many disciplines. Three basic fields of study have made unique contributions: electrical discharges, magnetohydrodynamics, and kinetic theory.

Interest in electrical-discharge phenomena may be traced back to the beginning of the 18th century, with three English physicists, Michael Faraday in the 1830s and Joseph John Thomson and John Sealy Edward Townsend at the turn of the 19th century, laying the foundations of the present understanding of the phenomena. Irving Langmuir introduced the terminology plasma in 1923, while investigating electrical discharges. In 1929 he and Lewi Tonks, another physicist working in the United States, used the term to designate those regions of such a discharge in which certain oscillations of the negatively charged electrons could occur. They called these oscillations plasma oscillations, their behaviour suggesting that of a jelly-like substance. Not until 1952, however, when two more United States physicists, David Bohm and David Pines, first considered the collective behaviour of electrons in metals as distinct from that in ionized gases was the general applicability of the concept of a plasma fully appreciated.

The collective behaviour of charged particles in magnetic fields is implicit in magnetohydrodynamic studies, the foundations of which were laid in the early and middle 1800s by Faraday and André-Marie Ampère of France. Not until the 1930s, however, when new solar and geophysical phenomena were being discovered, were many of the basic problems of the mutual interaction between ionized gases and magnetic fields considered. In 1942 Hannes Alfvén, a Swedish physicist, introduced the concept of magnetohydrodynamic waves.

Kinetic theory

These two separate approaches—the study of electric discharges and the study of the behaviour of charged particles within magnetic fields—were unified by the introduction of the kinetic theory of the plasma state, which holds that plasma, like gas, consists of particles all in random motion. In 1905 the Dutch physicist Hendrik Antoon Lorentz applied the kinetic equation for atoms (the formulation by the Austrian physicist Ludwig Eduard Boltzmann) to the behaviour of electrons in metals. Various physicists and mathematicians in the 1930s and 1940s developed the plasma kinetic theory further to a high degree of sophistication, and since the early 1950s interest has increasingly focussed on the plasma state itself. Space exploration, the development of electronic devices, a growing awareness of the importance of magnetic fields in astrophysical phenomena, and the quest for controlled thermonuclear (nuclear-fusion) power reactors all have stimulated such interest. Problems remain for the theory of plasma state, especially involving the simultaneous solution of James Clerk Maxwell's electromagnetic equations determining electric and magnetic fields and kinetic equations determining statistical behaviour of plasma particles.

General concepts

Although a gas normally acts as an electrical insulator, at sufficiently high temperatures it becomes a conductor. Increased thermal agitation of the molecules and atoms causes them to become ionized as they collide, disintegrating into electrons and positively charged entities called ions. At sufficiently high electron densities an ionized gas becomes a plasma. A basic property of a plasma is its electrical quasi-neutrality; that is, the numbers of negative charges and of positive charges per unit volume of plasma are approximately equal. At sufficiently high electron densities such quasi-neutrality is inevitable. Any excess concentration of either charge would lead to the development of large electrostatic restoring forces resulting from the attraction between opposite charges. Just as a weight on a spring will bob up and down about its rest position, any general displacement of the electrons as a group with respect to the positive ions leads to oscillation of the electrons as a whole about an equilibrium state. These movements were the plasma oscillations studied by Langmuir and Tonks.

PLASMA PROPERTIES AND PARAMETERS

The basic properties of plasma

In general, a plasma is a complex mixture of many different particle types, each distinguishable by its mass and charge and with all particles in rapid thermal motion. The basic properties of a plasma as matter are essentially independent of its chemical properties; they are determined primarily by conservation of energy and momentum laws and by the behaviour of the electrons. The study of the extremely complex interactions among the vast numbers of particles requires the use of statistical methods. Number densities or number per unit volume (n) and kinetic temperatures (T) are defined for each type of particle present. The average kinetic energy per particle is 1.5 times Boltzmann's constant (k) and the kinetic temperature, or $\frac{3}{2}kT$. As in a gas, kinetic pressures are defined as equal to the number density times Boltzmann's constant times temperature, or nkT. In thermodynamic equilibrium, all temperatures are the same as that of the enclosure, and (ignoring particle interactions) the sum of the kinetic pressures is the force per unit area exerted on the enclosure. A basic classification of plasmas can be made in terms of the electron-number density (n_e, the number of electrons per unit volume), the electron temperature (T_e), and the degree of ionization, or the fraction of the plasma ionized. One of the most important parameters associated with a plasma is the Debye length, a concept similar to that introduced in 1923 by the chemists Peter W. Debye and Walter K.F. Hückel in their theory of strong electrolytes (compounds that ionize easily and conduct electricity). In a plasma the Debye length is the maximum distance over which a large difference in the number of positive charges and negative charges can occur. It is equal to a constant times the square root of the ratio of the electron temperature to the electron number density; that is, $h = 69(T_e/n_e)^{1/2}$, in which the Debye length h is given in metres, n_e is the number per cubic metre, and T_e is given in degrees Kelvin. For distances appreciably greater than the Debye length, quasi-neutrality must exist, whereas for distances less than the Debye length an individual positively charged ion can exert a force on an electron. This phenomenon leads to a quantitative definition of a plasma: an ionized gas is a plasma provided the physical dimensions of the gas are much greater than the Debye length. For disturbances on a scale greater than the Debye length, the plasma behaviour is macroscopic, or collective; whereas, on a scale smaller than the Debye length, individual particle behaviour can occur.

When the plasma is not in equilibrium and, especially, when it contains electrical currents (*i.e.*, general motion of the electrons relative to the ions), the particle interactions within a sphere having a Debye radius are best accounted for in terms of equivalent binary collisions. That is, it is possible to add together the many simultaneous interactions between particles within such a sphere in such a way that they may be represented as a series of discrete events between two particles only. The number of such events is called the collision frequency.

Another important measure used in plasma physics is the angular frequency (ω_p) of the plasma oscillations noted by Langmuir and Tonks; it is equal to the average thermal speed of the electron divided by the Debye length, giving a value proportional to the square root of the electron density number, or $\omega_p = 56\sqrt{n_e}$ radians per second.

ELASTIC AND INELASTIC COLLISIONS

Collisions defined

In any gas, separate collision frequencies are defined for collisions between all different particle types and all different types of collisions. The total collision frequency for a particular particle type is the sum of these separate frequencies. Two basic types of collisions exist: elastic and inelastic; in an elastic collision, the kinetic energy of colliding particles remains the same after a collision as

before; in an inelastic collision, a fraction of the kinetic energy is transferred to the internal energy states of the colliding particles. In an atom, for example, the electrons have certain allowed (discrete) energies and are said to be bound. In a collision a bound electron may be raised from a low to a high energy state, or excited. This, however, can occur only by the expenditure of kinetic energy and only if the kinetic energy exceeds the difference between the two energy states. If the energy is sufficient, a bound electron may become a free electron, and the atom is said to be ionized; the minimum, or threshold, energy required to free an electron is called the ionization energy. Inelastic collisions may also occur with positive ions if electrons are still attached to them. In general, only collisions of electrons and photons (electromagnetic radiation particles) with atoms and ions are significant in these inelastic collisions; ionization by a photon is called photo-ionization.

A molecule also has discrete energy states, which may be excited by particle collisions. At sufficiently high energies of interaction, the molecule can dissociate into atoms or into atoms and atomic ions. Again, collision of electrons and photons with molecules may cause ionization, producing molecular ions. In general, the reaction rate for inelastic collisions is similar to that of chemical reactions. At sufficiently high temperatures, the atoms are stripped of all electrons and are bare atomic nuclei, and at temperatures of about 1,000,000 K or greater, nuclear reactions can occur—another form of inelastic collisions. When such reactions lead to the formation of heavier elements the process is called thermonuclear fusion; excess mass is transmuted into kinetic energy (that of motion), and kinetic energy is gained instead of lost.

RADIATION FROM PLASMAS

Elastic collisions occur between all particles, a fraction of the energy and momentum of one particle being exchanged with the other. The amount of energy exchanged depends on the mass ratio of the colliding particles. On the average, the energy exchanged between one particle type and another is proportional to the temperature difference and the collision frequency. In basically elastic collisions of electrons with ions, however, all of the energy is not exchanged; some is radiated; such radiation is called *bremsstrahlung* (German: "braking radiation") and is associated with the deceleration of electrons in a collision. The power (energy per unit of time) radiated is proportional to the square root of the electron temperature, the radiation being emitted as a continuum of wavelengths.

A similar process occurs in a magnetic field. In such a field an electron gyrates about a magnetic line of force in a helix, a configuration similar to a so-called spiral staircase, the frequency of rotation being called the cyclotron frequency. Charged particles moving in an arc or circle radiate energy, and when an electron moves in a helix about a magnetic line of force, an electromagnetic wave is radiated. At low electron energies an electron radiates at integral multiples of the cyclotron frequency, and at the highest energies the radiation becomes a continuum of frequencies, called synchrotron radiation, a phenomenon associated with similar processes in a synchrotron particle accelerator.

Synchrotron radiation

Other forms of radiation met with in plasma physics include line and recombination radiation. The lifetime of an atom or ion in an excited state is only about 1/100,000,000 of a second. When an electron goes from a high to a low energy state, a photon is emitted, with its frequency proportional to the energy difference between the two states. This emission is called line radiation because the photon would be observed as a line in a spectrometer. In another process, a free electron can be captured by an ion, followed by the emission of a photon; this is called radiative recombination, the associated radiation being a continuum of frequencies with a low frequency cutoff.

Emitted radiation may be absorbed and re-emitted from one atom or molecule to another, and depending on the processes, frequency changes can occur. If the distance travelled after emission and before absorption is small, compared with the dimensions of the plasma, the plasma is said to be optically thick. In this case, the spectrum of the radiation emitted from the surface is characteristic of an ideal blackbody (which emits radiation at the same rate as it absorbs it) and a blackbody temperature may be assigned; the power radiated is proportional to the fourth power of the blackbody temperature. If, however, the absorption distance is large compared with the plasma dimensions, the plasma is said to be optically thin, and the spectrum of the radiation emitted is characteristic of the individual types of radiation—in general, a function of the electron temperature.

Basic plasma physics

PLASMA FORMATION

Apart from solid-state plasmas, such as those in metallic crystals, plasmas do not usually occur naturally at the surface of the Earth. For laboratory experiments and technological applications, therefore, plasmas must be produced artificially. Because the atoms of such alkalies as potassium, sodium, and cesium possess low ionization energies, plasmas may be produced from these by the direct application of heat at temperatures around 3,000 K. In most gases, however, before any significant degree of ionization is achieved, temperatures in the neighbourhood of 10,000 K are required. Because all substances melt at temperatures far below that level, so far no container can withstand an external application of the heat necessary; therefore, any heating must be supplied internally by applying an electric field to the gas to accelerate and scatter all of the free electrons, thereby heating the plasma. Because of their small energy exchange in elastic collisions, electrons can be raised to much higher temperatures than other particles. As long as the rate of loss of electrons exceeds their rate of production, however, a plasma cannot take form.

Artificial production of plasmas

For plasma formation a sufficiently high electric field must be applied, its exact value depending on geometry and the gas pressure. An equilibrium state is reached when the electrical power input is balanced by energy losses to the outside. The electric field may be set up via electrodes or by transformer action, which is induced by a changing magnetic field. Laboratory temperatures of about 10,000,000 K with electron densities of about 10^{19} per cubic metre have been achieved by the transformer method. Extremely high temperatures, but relatively low-density plasmas, have been produced by the separate injection of ions and electrons into a mirror system (a plasma device using a particular arrangement of magnetic fields for containment). Other methods have used the high temperatures that develop behind a wave that is moving much faster than sound to produce what is called a shock front; lasers have also been employed.

EFFECTS OF MAGNETIC AND ELECTRIC FIELDS

Individual particle description. The behaviour of a plasma may be described at different levels. If collisions are relatively infrequent, it is useful to consider the motions of individual particles. A magnetic field exerts a force on a charged particle only if the particle is moving, the force being at right angles to both the direction of the field and the direction of particle motion. In a uniform magnetic field, (B), a charged particle gyrates about a line of force with an angular frequency equal to the magnetic field strength times the ratio of the charge of a particle (q) to its mass (m), or, $\omega_c = Bq/m$, charges of different sign rotating in opposite directions; in this equation, ω_c is the rate of gyration, or the cyclotron frequency.

The radius of the orbit is called the Larmor radius (r_L) and is equal to the product of the particle mass and the component (v) of the particle velocity in a plane perpendicular to the magnetic field divided by the product of its charge and the magnetic field—*i.e.*, $r_L = mv/B$. The centre of the orbit is called the guiding centre. The particle may also have a component of velocity parallel to the magnetic field and so traces out a helix when the motion of the guiding centre is uniform. If a uniform electric field (E) is applied at right angles to the direction of the magnetic field, the guiding centre drifts with a uniform velocity of magnitude equal to the ratio of the electric to the magnetic field (E/B), at right angles to both the electric and magnetic fields.

Should the electric field alternate in time, the problem is even more complex. If, however, such an alternating electric field varies at the same frequency as the cyclotron frequency, the guiding centre will remain stationary and the particle will be forced to travel in an ever-expanding orbit. This phenomenon is called cyclotron resonance and is the basis of the cyclotron particle accelerator.

Cyclotron resonance

The motion of a particle about its guiding centre constitutes a circular current. As such, the motion produces a dipole magnetic field not unlike that produced by a simple bar magnet. Thus, a moving charge not only interacts with magnetic fields but also produces them. Therefore, interactions between particles caused by their motions must also be considered. The direction of the magnetic field produced by a moving particle, however, depends both on whether the particle is positively or negatively charged and on the direction of its motion. If the motion of the charged particles is completely random, the net associated magnetic field is zero. On the other hand, if charges of different sign have an average relative velocity (*i.e.*, if an electrical current flows) then a net magnetic field over and above any externally applied field exists. The magnetic interaction between charged particles is therefore of a collective, rather than of an individual, particle nature.

Statistical and fluid descriptions. At a higher level of description than that of the single particle, kinetic equations of the Boltzmann type are used. Such equations essentially describe the behaviour of those particles about a point in a small-volume element, the particle velocities lying within a small range about a given value. The interactions with all other velocity groups, volume elements, and any externally applied electric and magnetic fields are taken into account.

At a still higher level of description, equations of a fluid type may be derived from the kinetic equations; they express the conservation of mass, momentum, and energy per unit volume, with one such set of equations for each particle type. Total conservation equations may be obtained from these by a process of summation from the equations of each particle type. For instance, the derived thermal-energy equation states that the rate gain of thermal energy is equal to the work done by pressure forces plus heating from electrical currents less losses from radiation and heat conduction. The total heat flux—that is, the thermal energy transported across a unit area of the plasma in a unit of time—is the sum of each term for each particle type present.

DETERMINATION OF PLASMA VARIABLES

The basic variables useful in the study of plasma are number densities, temperatures, electric and magnetic field strengths, and particle velocities. In the laboratory and in interplanetary space, both electrostatic (charged) and magnetic types of sensory devices called probes help determine the magnitudes of such variables. With the electrostatic probe, ion densities, electron and ion temperatures, and electrostatic potential differences can be determined. Small search coils and other types of magnetic probes yield values for the magnetic field; and from Maxwell's electromagnetic equations, the current and charge densities and the induced component of the electric field may be found.

Furthermore, in the laboratory, the absorption, scattering, and excitation of neutral and high-energy ion beams are helpful in determining electron temperatures and densities; in general, the refraction, reflection, absorption, scattering, and interference of electromagnetic waves also provide ways to determine these same variables. As an example, a powerful but difficult technique for such measurements is based on the scattering of a laser beam by plasma electrons. Aside from the above methods, much can be learned from the radiation generated and emitted by the plasma itself; in fact, this is the only means of studying cosmic plasma beyond the solar system. The various spectroscopic techniques covering the entire continuous radiation spectrum determine temperatures and identify such nonthermal sources as those producing synchrotron radiations.

Ion temperatures can be determined by measuring the widths of spectral lines on a spectrographic plate; their widths vary with thermal motions, a phenomenon called Doppler broadening. Velocities of plasmas in space are measured by Doppler shifts; that is, the spectral lines of radiation from receding bodies shift to lower frequencies, and vice versa. From the frequency of line radiation, researchers can identify elements, map the distribution of particles in plasmas, and determine their electron temperatures. Spectral lines are affected also by magnetic fields, and thus magneto-optical effects can determine such fields.

Doppler measurements of plasmas

FLUID-LIKE BEHAVIOUR

Plasma waves. An important parameter in plasma physics is the magnetic Reynolds number, the value of which depends on electrical resistivity of the plasma and the velocity of the particles. The Reynolds number is named for the British physicist Osborne Reynolds, who devised a measure of the critical velocity at which a flowing fluid changes its character. When the magnetic Reynolds number is much greater than one, resistance effects can be ignored and the magnetic lines of force are said to be frozen to (or to move with) the plasma. Because of this phenomenon certain interesting types of waves, called magnetohydrodynamic waves, occur at low frequencies. Generally, in a wave, the plasma particles oscillate about an equilibrium position, and their energy and momentum are transferred from one to another either by collisions or by interaction with electric and magnetic fields. Two fundamental types of wave motion can occur: longitudinal, like a sound wave, in which particle oscillation is in a direction parallel to the direction of wave propagation; and transverse, like a surface water wave, in which particle oscillation is in a plane perpendicular to the direction of wave propagation. In both types a wave may be characterized by a speed of propagation (u), a wavelength (λ), and a frequency (ν) related by an expression in which the velocity is equal to the product of the wavelength and frequency, namely, $u = \lambda\nu$. At low frequencies, on twanging or disturbing a line of force, a transverse wave called an Alfvén wave propagates parallel to the magnetic field with a velocity that depends on the particle density and the magnetic field strength. It is equal to the magnetic flux density (B) divided by the square root of the particle density (ρ) times the permeability of free space (μ_0); that is to say, $B/\sqrt{\mu_0\rho}$. A sonic wave that is an ordinary longitudinal sound wave can also propagate parallel to the magnetic field. Perpendicular to the magnetic field a different type of longitudinal wave called a magneto-sonic wave can occur.

Alfvén waves

In these magnetohydrodynamic transverse and longitudinal waves the plasma behaves as a whole, and the speeds are independent of wave frequency when the frequency is low; as in any fluid, excessive speeds can produce shock waves. At higher frequencies, however, the separate behaviour of ions and electrons causes the wave velocities to vary with frequency. The Alfvén wave splits into two components, referred to as the fast and slow Alfvén waves, which propagate at different frequency dependent speeds. At still higher frequencies these two waves (called the electron cyclotron and ion cyclotron waves, respectively) cause electron and cyclotron resonances (synchronization) at the appropriate resonance frequencies. Beyond these resonances, transverse wave propagation does not occur until frequencies, called cutoffs, comparable with the plasma frequency, are reached, when the usual kind of electromagnetic wave may propagate.

For propagation of a transverse wave perpendicular to a magnetic field, two transverse modes of wave vibration exist. One corresponds to the electric field oscillating parallel to the magnetic field. This wave behaves as though there were no magnetic field present and is referred to as the ordinary wave. Propagation occurs only if the frequency is greater than the plasma frequency (ω_p); an electromagnetic wave with a lower frequency is reflected by the plasma. The other wave, in which the electric field is perpendicular to the magnetic field, is affected by the magnetic field and is called the extraordinary wave. It has several resonance and cutoff frequencies. At low frequencies it is a fast Alfvén wave, and at high frequencies it is a normal electromagnetic wave.

The speed of an ordinary longitudinal sonic wave at high frequencies also becomes frequency dependent and is referred to as the ion-acoustic wave, a resonance occurring at a frequency determined by electrostatic oscillations of the ions. Beyond this frequency no sonic wave propagates parallel to a magnetic field until the frequency reaches the plasma frequency, above which electro-acoustic waves occur. The wavelength of these waves at the critical frequency (ω_p) is infinite, the electron behaviour at this frequency being the plasma oscillations of Langmuir and Tonks. Even without particle collisions, waves shorter than the Debye length are heavily damped; that is, their amplitude is rapidly decreased with time. This phenomenon, called Landau damping, arises from the fact that some electrons always have the same velocity as the wave. As they move with the wave, they are accelerated and thus extract energy from the wave, damping it in the process. In the absence of truly energy-dissipative mechanisms, such as collisions, and in suitable circumstances, after the initial wave has damped away, an echo having related wave characteristics can be detected in the plasma.

Landau damping of wave amplitude

Containment and instabilities. Magnetic fields are used to contain high-density, high-temperature plasmas, such fields exerting pressures and tensions on the plasma. An equilibrium configuration is reached only when at all points in the plasma these pressures and tensions exactly balance the pressure from the motion of the particles. A well-known example of this is the pinch effect observed in specially designed equipment. If an external electric current is imposed on a cylindrically shaped plasma and flows in parallel to the plasma axis, the magnetic forces cause the plasma to constrict, or pinch. An equilibrium condition is reached for a temperature proportional to the square of the electric current. This phenomenon implies that any temperature may be achieved by making the electric current sufficiently large, the heating resulting from currents and compression. In practice, however, since no plasma can be infinitely long, serious energy losses occur at the ends of the cylinder; and also, major instabilities develop in such a simple configuration.

In general, there are two basic methods of eliminating or minimizing end losses from an artificially created plasma: the production of toroidal plasmas and the use of magnetic mirrors (see ATOMS: *Nuclear fusion*). A toroidal plasma is essentially one in which a plasma of cylindrical cross section is bent in a circle so as to close on itself. For such plasmas to be in equilibrium and stable, however, requires special magnetic fields, the largest component of which is a circular field parallel to the axis of the plasma; any currents must be produced by transformer action.

A useful way of describing the confinement of a plasma by a magnetic field is by measuring containment time (τ_c), or the average time for a charged particle to diffuse out of the plasma; such time is different for each type of configuration. Various types of instabilities can occur in plasma. These lead to a loss of plasma and a catastrophic decrease in containment time. The most important of these instabilities is called magnetohydrodynamic. Although an equilibrium state may exist, it may not correspond to the lowest possible energy. The plasma, therefore, seeks a state of lower potential energy, just as a ball at rest on top of a hill (representing an equilibrium state) rolls down to the bottom if perturbed; the lower energy state of the plasma corresponds to a ball at the bottom of a valley. In seeking the lower energy state, turbulence develops, leading to enhanced diffusion, increased electrical resistivity, and large heat losses. In toroidal geometry, circular plasma currents must be kept below a critical value called the Kruskal–Shafranov limit, otherwise a particularly violent instability consisting of a series of kinks may occur.

Kruskal–Shafranov limit

Another class of instabilities is called drift wave, or universal. Spatial changes in density, temperature, or magnetic fields permit an abnormal number of particles with speeds equal to those of certain plasma waves. In an inverse manner to Landau damping, such waves grow by absorbing energy from the particles, and thus instabilities develop. A third basic type, dissipative instability, is associated with collisions between particles. Although a completely stable system appears to be virtually impossible, considerable progress has been made in devising systems that eliminate the major instabilities. Temperatures of the order of 10,000,000 K at densities of 10^{19} particles per cubic metre and containment times as high as 1/50 second have been achieved. The containment, however, is basically inefficient for practical use.

APPLICATIONS OF PLASMAS

The most important practical applications of plasmas lie in the future, largely in the field of power production. The major method of generating electrical power has been to use heat sources to convert water to steam, which drives turbogenerators. Such heat sources depend on the combustion of fossilized fuels, such as coal, oil, and natural gas, and fission processes in nuclear reactors. A potential source of heat might be supplied by a fusion reactor, with a basic element of deuterium–tritium plasma; nuclear fusion collisions between those isotopes of hydrogen would release large amounts of energy to the kinetic energy of the reaction products (the neutrons and the nuclei of hydrogen and helium atoms). By absorbing those products in a surrounding medium, a powerful heat source could be created. To realize a net power output from such a generating station—allowing for plasma radiation and particle losses and for the somewhat inefficient conversion of heat to electricity—plasma temperatures of about 100,000,000 K and a product of particle density times containment time ($n\tau_c$) of about 10^{20} seconds per cubic metre are necessary. For example, at a density of 10^{20} particles per metre cubed, the containment time must be one second. Such figures have still to be reached, although there has been much progress.

In addition to its generation of power, a fusion reactor might desalinate seawater. Approximately two-thirds of the world's land surface is uninhabited, but only one-half of this because it is arid. The use of both giant fission and fusion reactors in the large-scale evaporation of seawater could make irrigation of such wastes economically feasible.

Another possibility in power production is to eliminate the heat–steam–mechanical energy chain. One suggestion depends on the dynamo effect. If a plasma moves perpendicular to a magnetic field, an electromotive force, according to Faraday's law, is generated in a direction perpendicular to both the direction of flow of the plasma and the magnetic field. This dynamo effect can drive a current in an external circuit connected to electrodes in the plasma, and thus electrical power may be produced without the need for steam-driven rotating machinery. This process is referred to as MHD (magnetohydrodynamic) power generation, and has been proposed as a method of extracting power from certain types of fission reactors.

Current from dynamo effect

The inverse of the dynamo effect, called the motor effect, may be used to accelerate plasma. By pulsing cusp-shaped magnetic fields in a plasma, for example, thrusts proportional to the square of the magnetic field can be achieved. Motors based on such a technique have been proposed for the propulsion of craft in deep space. They have the advantage of being capable of achieving large exhaust velocities, thus minimizing the amount of fuel that must be carried.

Provided the current is not too large, a glow discharge occurs between two electrodes at pressures of one thousandth of an atmosphere or thereabouts. Such glow discharges are responsible for the light given off by neon tubes and such other light sources as fluorescent lamps that operate by virtue of the plasmas that they produce in electric discharge. The degree of ionization in such plasmas is usually low, but electron densities of 10^{16} to 10^{18} electrons per cubic metre can be achieved with an electron temperature of 100,000 K. The electrons responsible for current flow are produced by ionization in a region near the cathode, most of the potential difference between the two electrodes occurring there. This region does not contain a plasma, but the region between it and the anode, or positive electrode, does.

Other applications of the glow discharge include electronic switching devices; and it and similar plasmas produced by radio frequency techniques can be used to provide ions for particle accelerators and act as generators

Use in radio switching devices

of laser beams. In increasing the current through a glow discharge, a stage is reached when the energy generated at the cathode is sufficient to provide all the conduction electrons directly from the cathode surface, rather than from gas between the electrodes. Under this condition the large cathode potential difference disappears and the plasma column contracts. This new state of electric discharge is called an arc. Compared with the glow discharge, it is a high-density plasma and will operate over a large range of pressures. Arcs are used as light sources for welding, in electronic switching, for rectification of alternating currents, and in high-temperature chemistry. By running an arc between concentric electrodes and injecting gas into such a region, a hot, high-density plasma mixture is ejected, called a plasma jet; it has many chemical and metallurgical applications.

Natural plasmas

EXTRATERRESTRIAL FORMS

It has been suggested that the universe originated as a violent explosion about 10,000,000,000 years ago and initially consisted of a fireball of completely ionized hydrogen plasma. Irrespective of the truth of this, there is little matter in the universe now that does not exist in the plasma state. The observed stars are composed of plasmas, as are interstellar and interplanetary media and the outer atmospheres of planets. Man's knowledge of the universe has come primarily from studies of electromagnetic radiation emitted by plasmas and transmitted through them; and, since the 1960s, from space probes within the solar system.

In a star the plasma is bound together by gravitational forces, and the enormous energy that it emits originates in thermonuclear fusion reactions within the interior. Heat is transferred from the interior to the exterior by radiation; in the outer layers, convection or circulation of the plasma is important. In the vicinity of a hot star the interstellar medium consists almost entirely of completely ionized hydrogen, ionized by the star's ultraviolet radiation. Such regions are referred to as H II regions. The greater proportion by far of interstellar medium, however, exists in the form of neutral hydrogen clouds, referred to as H I regions. Because the heavy atoms in such clouds are ionized by ultraviolet radiation (or photoionized), they also are considered to be plasmas, although degree of ionization is probably only one part in 10,000. Other components of the interstellar medium are grains of dust and cosmic rays, the latter consisting of very high-energy atomic nuclei completely stripped of electrons. The almost isotropic velocity distribution of the cosmic rays is believed to stem from interactions with waves of the background thermal plasma.

Throughout this universe of plasma there are magnetic fields. In interstellar space magnetic fields are about 5×10^{-6} gauss (a unit of magnetic field strength), in interplanetary space 5×10^{-5} gauss, whereas in intergalactic space they could be as low as 10^{-9} gauss. These values are exceedingly small compared with the Earth's surface field of about one-half gauss. Although small in an absolute sense, these fields are nevertheless gigantic, considering the scales involved. For example, to simulate interstellar phenomena in the laboratory, fields of about 10^{15} gauss would be necessary. Thus, these fields play a major role in nearly all astrophysical phenomena. On the Sun the average surface field is something like one to two gauss, but magnetic disturbances arise, such as sunspots, in which fields of between 10 gauss and 1,000 gauss occur. Many other stars are also known to have magnetic fields, whereas such fields as 10^{-3} gauss are associated with various extragalactic nebulae, synchrotron radiation being observed.

Magnetic fields in space

SOLAR–TERRESTRIAL FORMS

The visible region of the Sun is the photosphere (see Table 4), with its radiation about the same as the continuum radiation from a blackbody. Lying above the photosphere, the chromosphere is characterized by the emission of line radiation from atoms and ions. Outside the chromosphere, the corona expands into the everblowing solar wind (see below), which on passing through the planetary system eventually encounters the interstellar magnetic field.

Table 4: Various Natural Plasmas, Their Electron Densities and Temperatures

plasma	n_e (per cu m)	T_e (°K)
Sun		
Centre	10^{31}	1.5×10^7
Photosphere	10^{20}	4,200
Chromosphere	10^{17}–10^{20}	5×10^5
Corona	10^{13}	1.5×10^6
Solar wind (near Earth)	5×10^6	4×10^5
Interstellar space		
H II regions	10^6	10^4
H I regions	10^2	100–125
Intergalactic space	1	3 (?)
Earth		
Outer magnetosphere	10^7–10^6	10^4
Plasmasphere	10^{10}–10^9	10^4
Ionosphere	10^{11}–10^{12}	250–3,000
Metals	10^{28}	10^4

The Sun's corona is believed to be heated by magnetohydrodynamic or acoustic waves generated in the photosphere. The solar wind, a collisionless plasma made up primarily of electrons and protons, represents an outflow of matter moving at supersonic speed. The wind takes with it frozen-in extensions of the Sun's magnetic field. In the region of the Earth, the wind has an average speed of 400 kilometres (128.5 miles) per second; and on encountering the Earth's magnetic field, a shock front develops, the pressures acting to compress the field on the side toward the Sun and elongate it on the night side (in the Earth's lee away from the Sun). The Earth's magnetic field is, therefore, confined to a cavity called the magnetosphere, into which the direct entry of the solar wind is prohibited. This cavity extends for about 10 Earth radii on the Sun's side and about 1,000 Earth radii on the night side.

Solar wind

The magnetosphere contains plasma. At about four Earth radii from the centre of the Earth, electron density markedly increases. Within this region, plasma particles are confined in closed orbits by the Earth's magnetic field, whereas those outside the region may drift to the edge of the magnetosphere. The inner region is called the plasmasphere. To within about 2,500 kilometres (1,550 miles) of the Earth's surface, the plasma throughout the magnetosphere consists of completely ionized hydrogen. In a somewhat ill-defined region below 2,500 kilometres and above 350 kilometres (217 miles), oxygen and helium ions predominate. The degree of ionization decreases but the electron density increases with decreasing height. The degree of ionization at 350 kilometres is about 5×10^{-3} per cubic metre, whereas the electron density reaches a maximum of about 10^{12} cubic metres. Below this level, the electron density slowly decreases, falling away rapidly beneath 100 kilometres (62 miles).

The region between about 100 and 350 kilometres is normally referred to as the ionosphere, although this term has also been used to describe the entire magnetospheric plasma. In the ionosphere the dominant ions are those of atomic and molecular oxygen and nitric oxide. These are produced by photo-ionization of the upper atmosphere by solar ultraviolet radiation.

High-energy particles also exist in the magnetosphere. At about 1.5 and 3.5 Earth radii from the centre of the Earth, two regions contain high-energy protons. These regions are the Van Allen radiation belts; the protons in them are trapped in mirror systems formed by the Earth's magnetic dipole field. In general, the inner regions of the plasmasphere are stable, but in the outer regions disturbances can lead to the rejection and ejection of high-energy particles. These precipitate into the upper atmosphere in the region of the poles, causing auroral displays (see ATMOSPHERE: *Auroras*).

TERRESTRIAL PLASMA PHENOMENA

Plasma can exist also in the lowest regions of the Earth's atmosphere. A lightning stroke is actually an oxygen–nitrogen plasma heated to approximately 20,000 K, with an ionization of about 20 percent, similar to that of a laboratory arc. Although the bolt or stroke is only a few centimetres thick and lasts only a fraction of a second, tremendous

Lightning flashes and ball lightning

energies are dissipated. A lightning flash between ground and a cloud, on the average, consists of four such strokes. An associated phenomenon is ball lightning. There are authenticated reports of glowing, floating, stable balls of light several tens of centimetres in diameter occurring at times of intense electrical activity in the atmosphere. On contact with an object these balls release large amounts of energy. Although lightning balls are probably plasmas, so far no adequate explanation of them has been given.

Considering the origins of plasma physics and the fact that the universe is little more than a vast sea of plasma, it is ironic that the only naturally occurring plasmas at the surface of the Earth besides lightning are those to be found in ordinary matter. The free electrons responsible for electrical conduction in a metal constitute a plasma. Ions are fixed in position at lattice points, and so plasma behaviour in metals is limited to such phenomena as plasma oscillations and electron cyclotron waves (called helicon waves) in which the electron component behaves separately from the ion component. In semiconductors, on the other hand, the current carriers are electrons and positive holes, the latter behaving in the material as free positive charges of finite mass. By proper preparation, the number of electrons and holes can be made approximately equal so that the full range of plasma behaviour can be observed.

The importance of magnetic fields in astrophysical phenomena has already been emphasized. It is believed that these fields are produced by self-generating dynamos, although the exact details are still not fully understood. On Earth, differential rotation in its liquid core causes the external magnetic dipole field (manifest as the North and South Poles). Fields inside the core are twisted into circular loops; cyclonic turbulence in the liquid, generated by heat conduction and Coriolis forces (apparent forces accompanying all rotating systems, including the heavenly bodies) then regenerate the original dipole field from these loops.

Similar processes are believed to occur both in the Sun and the Galaxy. In the Sun the circular internal magnetic field is observed by lines of force apparently breaking the solar surface to form exposed loops; entry and departure points are what are observed as sunspots. Although the exterior magnetic field of the Earth is that of a dipole, this is further modified by currents in both the ionosphere and magnetosphere. Lunar and solar tides in the ionosphere lead to motions across the Earth's field that produce currents, like a dynamo, which modify the initial field. Again, changes in the intensity of the solar wind induce or produce other currents in the magnetosphere; such currents constitute the essence of a magnetic storm. (B.S.L.)

COLLOIDS

A colloid is any substance composed of fine particles whose dimensions are much larger than those of atoms or ordinary molecules but much smaller than those of particles visible to the unaided eye; colloids cannot be detected by optical microscopes in transmitted light. Usually colloidal particles exist as a dispersion in some medium, but they may exist purely—*i.e.,* without a medium. Some colloids are vital to everyday life—cheese, dough, paints, silk, rubber, proteins, plastics, clay, foam, smoke, and gelatin, for example. Colloids can be generated and also eliminated by nature as well as by industrial or technological processes. Because colloidal systems have certain properties in common and because their existence is limited by a range of factors, such as temperature, they were once classified as another state of matter. This classification is no longer adequate. A classification according to the particular states of matter involved—*e.g.,* solid in liquid, liquid in liquid, solid in gas—although not exhaustive is still useful.

Colloidal particles consist of aggregates of molecules, sometimes only of atoms, linked or bonded together in a variety of ways. The production of colloidal particles or their elimination, either by further clotting or by breaking them up into molecules, is based on standard principles of chemical and physical bonding, but still contains elements of empiricism, especially in obtaining the desired size range (between 10^{-7} and 10^{-3} centimetres). It is safe to say that the colloids prepared in living organisms by biological processes are absolutely vital to the existence of the organism and that the colloids generated with inorganic compounds in the earth, waters, atmosphere, and in technological processes, though not nearly so important at first examination, are just as vital to the existence of lifeforms.

History of colloid studies

The art of preparing and processing colloidal materials goes back as far as civilization itself. The scientific study of colloids, however, dates only from the beginning of the 19th century, one of the first important experiments being with a lump of clay in a U-tube filled with water. When an electric potential difference was applied to the two ends of the tube, the water was seen to move toward the negative electrode, while particles of clay moved toward the positive electrode. Such a movement to either positive or negative electrode of suspended, now called colloidal, particles under an applied electrical field is termed electrophoresis. Another early discovery was that particles in suspension when viewed under a microscope are in ceaseless, irregular, or random motion, called Brownian movement in honour of the British botanist Robert Brown, who in 1827 first recorded the phenomenon using aqueous suspensions of pollen. Brownian movement is now known to result from the irregular bombardment of colloidal particles by molecules of the surrounding liquid.

Francesco Selmi, an Italian chemist, published (1845–50) the first systematic study of inorganic colloids, in particular silver chloride, Prussian blue, and sulfur. Selmi showed that these colloidal bodies would coagulate (group together into large aggregates that separated from the liquid, a process similar to the precipitation of a substance from solution into solid particles that then settle out) if salts were added to the system and that the salts differed in their coagulating or precipitating power. From the absence of temperature changes upon coagulation, or upon another process called peptization (see below), he concluded that the particles were not simple molecules distributed as molecules throughout the medium but some kind of combination of molecules.

Graham's definitions

Thomas Graham, a Scottish chemist, is usually regarded as the founder of modern colloid science. During the 1860s, Graham clearly recognized those properties regarded as characteristic of the colloidal state, such as absence of crystallinity and of ordinary chemical relations and an exceedingly low rate of diffusion, which arises from the large size of colloidal particles (about 10^{-7} to about 10^{-3} centimetre). He also recognized the biological significance of colloids.

The term colloid (from the Greek word for "glue") was coined by Graham in 1861 and applied by him to such substances as gelatin, albumin, and gums, which are retained by membranes such as parchment paper when subjected to dialysis (a technique of separating substances by their different rates of diffusion through semipermeable membranes). Such substances as salts and sugars, which readily pass through the membrane, he termed crystalloids, because of the comparative ease with which they could be obtained in crystalline form from their solutions. Of Graham's other contributions to the nomenclature of colloids, those still in use are the terms sol, gel, peptization, and syneresis. A sol is a dispersion of a solid in discrete units (*i.e.,* a solid broken up into independent particles and distributed through a liquid or gaseous medium), the type of dispersion medium being shown by a prefix—*e.g.,* hydrosol (for dispersions in water), aerosol (for dispersions in air). A gel is a dispersion of the solid particles linked together to form a structure with some mechanical strength. Peptization, from analogy with peptic digestion,

referred to the spontaneous dispersion of a precipitate to form a colloid as, for example, the dispersion of a precipitate of Prussian blue when an attempt is made to wash it. The term syneresis (from a Greek word meaning "to contract") was suggested by the study of silicic acid (obtained from water glass and dilute acids) for the phenomenon of spontaneous shrinkage of a gel with the exudation of the dispersion medium.

An interesting example is provided by gold sols, colloids that are prepared by reducing (freeing the elemental metal from its compound) dilute solutions of gold chloride and that are usually strikingly coloured (ruby red, blue, green, and other colours). Although gold sols had been known to the alchemists of the 17th century, the English physicist and chemist Michael Faraday, in 1857, was the first to make a scientific study of their preparation and properties. He showed that addition of salts turned the ruby-red sols blue and then coagulated them and that these effects could be prevented by the addition of gelatin and other hydrophilic (easily dissolved in water as compared with hydrophobic, difficult to dissolve in water) colloids. The ruby sols, which had all the appearances of solutions, were shown by him to contain particles of gold because, unlike true solutions, a cone of light passing through them became visible to an observer situated at right angles to the beam. This phenomenon, later studied in more detail by John Tyndall, a British physicist, is usually referred to as the Tyndall effect. Of the other metal sols, those of silver were much studied, particularly in the United States, in connection with the photographic process. The preparation and purification of sols were facilitated by electrodialysis (separation by applying an electrical field, and allowing the dissolved impurities to diffuse through a membrane) and by ultrafiltration—*i.e.,* filtration through filters of very fine pore size.

Stability of sols

The stability of hydrophobic sols to salts that were added to the system was quantitatively examined after electrophoresis had shown that all colloids carry an electric charge, the same kind of charge on each particle in the system, and that it is to this charge that hydrophobic colloids owe their stability; because like charges repel one another, the particles cannot come into collision and, therefore, cannot coagulate and group together into larger than colloidal size. Coagulation is brought about, however, by the addition of an ion carrying the charge opposite to that on the colloidal particle, and the coagulating power of an ion increases rapidly with its charge (*e.g.,* to coagulate negatively charged sols such as gold or arsenic sulfide, the requisite concentration of sodium, calcium, and aluminum ions, with single, double, and triple positive charges, respectively, are approximately in the ratio 1:1/70:1/900). Addition of a hydrophilic colloid to a hydrophobic one protects the latter from the coagulating action of salts, a phenomenon termed protection. Hydrophilic colloids such as the proteins, the condensation products of amino acids that are the essential constituents of all living cells, are only precipitated by high concentrations of a salt. For a series of salts, the comparative efficacy runs approximately parallel to their solubility in water, indicating that precipitation of the colloid is caused by dehydration, or removal of water molecules, from the colloid particle.

In 1903 the invention of the ultramicroscope brought a notable advance in techniques for examining colloidal systems. The chief difference from the usual method of microscopic examination was the mounting of the microscope at right angles to the incident light beam, which strikes the object to be examined; the colloidal particles, because of the light they scatter laterally, then stand out as bright points of light on a dark ground. This apparatus can detect particles down to about 100 angstroms in size, compared with about 2000 angstroms for the ordinary microscope.

In 1905 the terms hydrophilic and hydrophobic were introduced to differentiate aqueous suspensions of markedly differing properties, as typified by gelatin and the colloidal dispersion of metals, respectively. These types differ in many ways, the most striking being in their stability to salts added to their systems and their reversibility after precipitation. A gelatin solution, for example, is only precipitated by high concentrations of salts and, after drying, will readily take up water again. It "likes" water; it is hydrophilic. A sol of a metal, or of an insoluble salt, is precipitated readily by traces of salts and does not return to the colloidal state, even if the salts are eliminated by washing the precipitate. It "abhors" water: it is hydrophobic. These phenomena reveal the reversible property of hydrophilic systems and the irreversible property of hydrophobic ones. For dispersion media other than water, the parallel terms lyophilic and lyophobic (*lyo-* from a Greek word meaning "to loosen, to dissolve") were proposed.

Many striking developments in physics and chemistry in the early years of the 20th century were relevant to colloids, including, for example, the increase in knowledge of the electronic structure of the atom, of molecular size and shape, and of the nature of solutions. In addition, powerful new methods for studying the size and shape of colloidal particles were developed, such as ultracentrifugation, electrophoresis, diffusion, and the scattering of light and of X-rays. The study of interfaces (surfaces of discontinuity between substances or states of matter), clearly of great relevance to colloids, developed as a major field, surface chemistry. The investigation of colloids in industry and biology also led to major developments in the knowledge of detergents, dyes, proteins, polymers, and other substances.

Irreversible systems

The main classification of colloids is into reversible and irreversible systems. (Any system is irreversible if, after its components have been induced to interact, the products of their reaction are so stable, or are removed so effectively from the system, that no manipulation of the system will bring about a reaction to reproduce the original components. In a reversible system the products of a reaction may be induced to react in a way that will reproduce the original components. The reactions may be physical or chemical.) In the irreversible systems, to which belong sols (dilute suspensions), emulsions, foams, pastes (concentrated suspensions), and certain types of gels, the size of the particles is greatly dependent upon the method of preparation. In the reversible systems particle size is either determined by the molecular size of the colloidal material, as in polymers, polyelectrolytes, and proteins, or the particles are formed from small molecules by a reversible association, as in soaps, detergents, and certain dyes. The most important aspects of these diverse systems are outlined below.

LYOPHOBIC SOLS

Lyophobic sols are dilute systems, some typical examples of which are colloidal metals, silver chloride, and Prussian blue, which already have been mentioned. Of suspensions in gaseous media (aerosols), some, such as fogs, mists, and smokes, are well-known; others, such as airborne suspensions of bacteria, viruses, and molds, are less obvious but of equal importance. Various kinds of smoke have found considerable use in war for camouflage and antipersonnel purposes, and peacetime uses of aerosols include insecticidal sprays and frost prevention in orchards by means of smudge pots. The removal of smokes still presents a major industrial problem. In an electrostatic precipitator, the smoke passes between oppositely charged wires (charged to a high potential difference); because of the charges that smoke particles carry or that are induced upon them by the electrical field, the particles are attracted to one or other of the wires, which then neutralizes them and thereby enables the particles to cluster into aggregates too large to be carried upward by the rising gases; these larger masses can then be collected by mechanical means from the precipitators.

Electrostatic precipitators

Dilute suspensions are known in all three dispersion media, gas, liquid, and solid. Only those in air and in aqueous solutions will be considered here. The dispersed particles are almost invariably composed of a large number of atoms or molecules and are usually visible in the ordinary microscopes or ultramicroscopes (about $^{1}/_{1,000,000}$ to about $^{1}/_{1,000}$ centimetre). The amount of particles in the

dispersed phase is always small, usually much less than 1 percent. Suspensions can be prepared by breaking down particles of macroscopic dimensions, a process called dispersion, or by growth from molecular or atomic units, a process called aggregation, until particles of the requisite colloidal dimensions are obtained.

Dispersion is usually carried out in the laboratory by grinding the substance in an agate mortar and, commercially, by means of a colloid mill in which the mixture of coarse particles and dispersion medium is subjected to intense shearing forces; a protective colloid is frequently added to prevent the colloidal particles produced by the mill from reaggregating. Metals can be dispersed in liquids by striking an electric arc between them, usually in the presence of a trace of sodium hydroxide to stabilize (*i.e.*, prevent from forming agglomerates) the sol formed. In some cases dispersion can be brought about by adding a small amount of a third substance called a peptizing agent; clays, for example, are peptized by alkalies, and many finely divided precipitates are peptized by soaps or other hydrophilic colloids.

In aggregation processes the first step is to form a supersaturated solution (in which more solute is dissolved than the solvent can contain normally, and any disturbance of the system will force the excess of dissolved solute out of solution) by chemical or physical means. This supersaturation leads to formation of solute nuclei that grow to colloidal size and then have to be stabilized in order to prevent flocculation—*i.e.*, coagulation of the particles into large flakes. An example of such a chemical process is the reaction between arsenic oxide and hydrogen sulfide in very dilute solution; in this reaction the insoluble arsenic sulfide appears in the form of minute particles dispersed in the aqueous solution. A common physical method is typified by the formation of sulfur sols by pouring a solution of sulfur in alcohol into water. Mist, fog, and clouds are aerosols formed by physical means in the atmosphere—the rapid cooling of air that is saturated with water vapour to a temperature at which the vapour condenses into minute droplets of water.

Importance of large numbers of nuclei

The size of the particles formed in aggregation processes is determined by the rates of nucleation (formation of nuclei) and by the growth of the nuclei into crystals or crystal-like growths. A high rate of nucleation is imperative for the formation of small particles, because the fewer the nuclei are, the larger they will grow, the total mass of substance precipitating being the same whatever the number of nuclei.

Liesegang rings. Another phenomenon in which nucleation plays an important role, the Liesegang rings (named after the German physicist and chemist R.E. Liesegang), is observed if an insoluble precipitate is formed during the diffusion of one reactant into the other. If, for example, a thin layer of dilute potassium dichromate in a gelatin gel is placed on a glass plate and a crystal of silver nitrate is placed at the centre, a precipitate of insoluble silver chromate is eventually produced in the form of concentric rings as shown in Figure 51. Each band is formed when a certain supersaturation has been reached locally, so that nucleation occurs. By subsequent growth of the nuclei, the concentration is lowered in the neighbourhood of the bands, and the diffusion of silver nitrate can proceed a certain distance before a sufficient degree of supersaturation is again reached and the next band precipitates. These bands are very similar to the banding in some minerals, such as agate, and are believed to explain such formations.

Stability. The stability of sols arises chiefly from the electrical charges carried by the particles (all being negative or all being positive), thus preventing contact under conditions such that without charges collisions would occur. The charges are usually obtained by the preferential adsorption of one type of ion, often a negative ion, because most sols are negatively charged. When the sol particles are close together, however, attractive forces (of the same nature as those between gas molecules, which cause gases to condense into liquids) make themselves felt. The addition of inorganic salts, which effectively screen the charges on the colloidal particles, leads to a reduction of the range of the electrical repulsive forces, so that the attraction becomes predominant and coagulation follows.

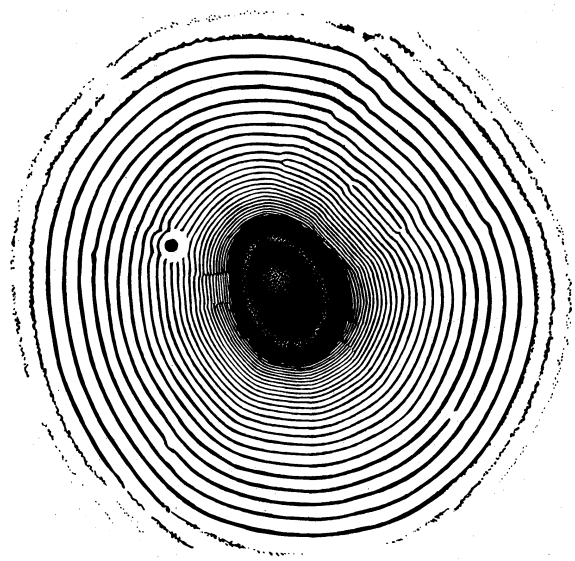


Figure 51: Liesegang rings of silver chromate. These resemble the banding in many minerals and are believed to explain the mineral formation (see text).

EMULSIONS

The term emulsion is used to denote any colloidal dispersion of one liquid in another, but in practice only oil and water (aqueous) solutions need to be considered, because biological emulsions, as well as those of most domestic and industrial importance, contain an oil and an aqueous solution as the two immiscible phases. In order to attain the requisite degree of stability, a third component, called the emulsifying agent, is necessary and is usually present in amounts of about 1 to 5 percent. Emulsions can exist in two types, oil-in-water (indicated by O/W) and water-in-oil (W/O), depending upon whether the aqueous or the oil phase is the continuous one (*i.e.*, one in which the other exists in dispersed form). Two fundamental problems arise: the origin of the stability induced by the emulsifying agent and the reason some agents promote oil-in-water and others water-in-oil.

Importance of emulsions

The technical and biological aspects of emulsions are vast. The principal technical uses are in pharmacy and cosmetics (creams and ointments are usually emulsions), in food preparations (*e.g.*, salad dressings, mayonnaise, margarine), and in many industries, particularly leather, textiles, and paper. Margarine, for example, is a water-in-oil emulsion stabilized by about 1 percent of oxidized soybean oil; mayonnaise and other salad dressings are emulsions of edible vegetable oils in aqueous solutions of edible acids stabilized by egg yolk.

Commercially, oils are often emulsified with water to achieve three main goals: dilution of the oil, increase in area of interface or overall surface contact between the components, and modification of physical properties. Dilution effects economy by spreading the oil over a larger surface, as, for example, in polishes and oil-soluble insecticides. An increase in area of interface may accelerate some chemical reactions (*e.g.*, polymerization of vinyl compounds, hydrolysis of fats in soap manufacture). The modification of physical properties includes the removal of objectionable characteristics, as, for example, emulsification diminishes the taste or smell of pharmaceutical preparations such as cod-liver oil and increases the fluidity of heavy oils such as tar. On the other hand, crude petroleum is frequently produced as a water-in-oil emulsion. Breaking up this emulsion is often difficult, but it is a necessary step before further processing of the petroleum can take place.

In evaluating the stability of emulsions, creaming, coagulation, and coalescence should be distinguished. Creaming is the accumulation of the droplets, usually at the top of the emulsion, under the influence of gravity. Small drops

cream much more slowly than large ones. In coagulation the droplets come together to form aggregates in which the original drops are still separated. Coagulation enhances creaming. Coalescence means that the drops unite to form a continuous phase; fresh milk creams markedly in a few hours, and by acidification it can be made to coagulate, after which coalescence is strongly increased by churning.

Emulsification is achieved by intense agitation of the mixture. On a small scale this can sometimes be achieved merely by shaking the mixture by hand, but it is more usual to force the coarse emulsion under pressure through some type of valve, as in the farmer's "cream machine." Industrially, various types of colloid mills are used in which the emulsion is subjected to intense shearing forces, finally passing through a homogenizer. These treatments decrease the particle size and so increase the stability. (Milk is frequently homogenized in order to diminish creaming.)

The principal emulsifying agents for oil-in-water emulsions are: proteins, gums, carbohydrates, natural and synthetic soaps, clays, and certain hydrated oxides, particularly of silicon or aluminum; for water-in-oil emulsions: heavy metal salts of the fatty acids (*e.g.,* zinc stearate, nickel oleate), long-chain alcohols such as cetyl alcohol, long-chain esters such as glyceryl monooleate, oxidized oils, soot, lampblack, graphite, and other substances.

Various theories have been advanced to explain emulsifiers, the most simple being that originally suggested for solid powders. As Figure 52 shows, for idealized spherical particles, if the solid is preferentially wetted by one component or phase, then more particles can be accommodated if the interface between the two components is convex to the wetting phase; and, in an encounter between two droplets, actual contact between the liquid in the droplets is prevented by the particles clinging to them, so that coalescence of the droplets cannot occur. In other words, preferential wetting by water should tend to give oil-in-water emulsions as in Figure 52B, while preferential wetting by oil should give the inverse type, as in Figure 52A. This difference is in agreement with the above list of emulsifiers, those promoting oil-in-water emulsions being all hydrophilic, whereas those that favour water-in-oil are more oleophilic (readily soluble in oil). Other conditions are also involved, however, in particular the relative amounts of oil and water and even the order in which the phases are mixed.

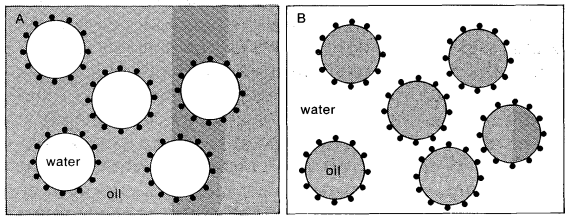


Figure 52: Solid powder at an oil–water interface, showing (A) how preferential wetting by oil tends to give water–oil emulsions and (B) how preferential wetting by water tends to give oil–water emulsions.

Natural and synthetic soaps soluble in water are among the most widely used oil-in-water emulsifying agents. Certain synthetic soaps (see below *Association colloids*) are often preferable to ordinary (fatty acid) soaps, because they are not precipitated by calcium ions (which do form a precipitate with ordinary soaps), and so the synthetic soaps can be used with hard water (high in calcium ion). The electrical charge acquired by the oil drops when they adsorb the soap ions is one factor in stability; the mechanical properties of the adsorbed film seem to be another one. The strong reduction of the interfacial tension facilitates emulsification greatly.

As mentioned before, the salts of divalent and trivalent metals with fatty acids (*e.g.,* nickel oleate) are efficient emulsifiers for the water-in-oil type. In these systems, the interfacial film consists of solid particles or of a polymolecular skin rather than a film one molecule deep, a monolayer. A simple experiment to illustrate phase inversion (reversal of the roles components play in a system) is to prepare an oil-in-water emulsion stabilized by sodium oleate (soluble in water) and to add a few drops of saturated magnesium chloride; on being shaken, the system inverts to the water-in-oil type because of the magnesium oleate that is formed. Proteins and gums as emulsifiers normally promote only the oil-in-water type of emulsion. A great variety of proteins are used commercially, such as egg albumin, dried blood, casein (from milk); and proteins are wholly or partly responsible for the stability of many natural emulsions such as milk and rubber latex. Food emulsions are frequently stabilized by proteins, gelatin, and other substances. The stabilizing action of these arises from the mechanical protection given by the protein monolayers, or films, that form spontaneously and that are relatively impenetrable for oil.

Phase inversion

FOAMS

A foam is a gaseous dispersion, usually of air, in a liquid. Pure liquids do not give stable foams, and, for any reasonable stability, a third component, the foaming agent, is essential. Of the great variety of substances that act as foam promoters, the best known are the natural and synthetic soaps, proteins including gelatin and other compounds obtained from proteins, certain finely divided solid powders, certain polymers (large molecules formed from smaller ones), and the saponins (naturally occurring substances that produce a soapy lather).

Household soap, which consists chiefly of a mixture of sodium stearate and sodium palmitate, is well-known for its bubble-blowing capacity. This property is greatly improved by the addition of glycerol, which probably acts by increasing the viscosity of the film. Many synthetic soaps, sodium cetyl sulfate, for example, are equally effective and have an added advantage in that their lathering power is not readily diminished by hard water (see below *Association colloids*).

Stabilization of foams arises mainly from the same factors that govern the stability of emulsions. Indeed, a suspension of air bubbles in water (a foam) is very similar to a suspension of oil droplets in water. In both cases the same types of stabilizers are used. Proteins as in whipped cream, marshmallow (made from gelatin and sugar), and meringue (from egg white) find a wide use as foaming agents in foodstuffs, because of their edible or nontoxic nature. The fire foam used to combat oil fires consists of bubbles of carbon dioxide (liberated from sodium bicarbonate and aluminum sulfate) stabilized by dried blood, glue, or any other cheap protein-containing materials. Beer froth is believed to be stabilized by the colloidal constituents present in the brew, which include proteins and carbohydrates. Saponins, noted for their foaming capacity, are obtained commercially from various plants and are complex mixtures. They form rather solid surface films with a mechanical strength to which their foams owe their stability, just as in the case of proteins. Of the polymers with foaming power for aqueous solutions, the most important are methyl cellulose (water-soluble type) and polyvinyl alcohol (see below *Polymers*). Certain finely divided solids, particularly if they have a waxy or hydrophobic (not soluble in water) surface, such as aluminum stearate or heavy metal soaps insoluble in water, remain at the air–water interface on agitation and stabilize the foam so formed. By suitable treatment many other materials can be made to behave similarly, and this behaviour forms the basis of what is called the froth-flotation method for separating minerals. The structure of a single bubble in a foam stabilized by a fine powder is quite similar to that of the emulsion droplet represented in Figure 52B.

Foaming may often be an undesirable property, as in lubricating oils, and its prevention is not always easy. Aqueous foams can usually be broken by spraying with small amounts of certain alcohols (*e.g.,* amyl, octyl) or fatty acids.

PASTES

Pastes are concentrated dispersions of fine solid particles in a liquid continuum. From the practical point of view they are of extreme importance, covering such diverse materials as paints, putty, clays, dough, drilling muds,

Importance of pastes

toothpaste, and others. These materials are all noted for their peculiar and characteristic mechanical properties, as is indicated by such terms applied to them as plasticity, ductility, and moldability. The study of the flow and deformation properties of pastes, which are so vital in their practical uses, belongs to the science called rheology.

The flow behaviour of a paste is determined not only by the volume concentration of the solid particles but also by their size and shape, by their size distribution, and by the colloidal stability of the suspended particles. At a given solid concentration, a flocculated paste (one in which particles are held together loosely in amorphous, cloudy flakes) may be stiff and plastic, whereas a stable, or deflocculated, one may be easily pourable. Examples of flocculated systems are very polar compounds (in a polar molecule the distribution of electrons is not symmetrical and causes the molecule to behave as though it had a positive end and a negative end, or pole) in organic media (*e.g.,* silica, calcium carbonate, and other inorganic salts in such oils as benzene) or nonpolar substances such as carbon black in polar media such as water. Such systems can be rendered deflocculated by suitable additives, a practical example being the addition of aluminum stearate to the oil vehicle of inorganic paints. By such means a higher concentration of the disperse phase can be obtained without increasing the consistency of the material; or, alternatively, an increase in pourability can be obtained at the same solid concentration. It is for higher concentration that materials such as sodium silicate are added to potter's clay, because the water present, which ultimately will have to be removed, should be kept to a minimum. Putty, another example of a paste, consists of finely ground whiting (natural calcium carbonate) in a medium of linseed oil.

Pastes, particularly those with the higher concentrations of disperse phase, usually show some sort of yield value (give) upon deformation. This property is clearly essential if the material is to exhibit plastic properties (*e.g.,* potter's clay), because it has to retain its shape against the forces of gravity. Certain pastes exhibit the phenomenon known as thixotropy; *i.e.,* after stirring or shaking they become much thinner in consistency, returning to their original state on being allowed to stand. A well-known example is quicksand, which consists of a paste of sand and water made thixotropic by the presence of certain clays. Paint should be thixotropic in order to flow easily under the brush and to allow the brush streaks to disappear. After a short time the applied paint should solidify to prevent the paint from flowing away from vertical planes. The reverse phenomenon, termed dilatancy, in which a suspension that flows freely under a small force turns rigid when larger forces are applied, is shown by very concentrated, highly stabilized suspensions—*e.g.,* by pastes of pure silica in water.

Thixotropic pastes

GEL STRUCTURES

A flocculated paste, or suspension of very fine particles, often behaves as a gel—*i.e.,* a quasi-homogeneous solid-like system, but rich in liquid. During the flocculation a scaffolding structure is formed by the molecules, preventing sedimentation of the floc in the structure and enmeshing the dispersion medium. Such structures are preferentially formed by long or flat particles. The systems are often thixotropic; the bonds between the particles can be broken by mechanical action. They show syneresis; *i.e.,* in the long run the scaffolding retracts by further flocculation, and liquid is exuded. Examples of gels are: certain clays, particularly bentonite (flat particles), iron hydroxide (flat), and vanadium pentoxide (needles). Those gels formed from ferric hydroxide are reddish brown and become more transparent when liquefied. Fatty-acid salts such as aluminum, calcium, and zinc oleates are the thickening agents in many greases, which are examples of gels with a nonaqueous liquid phase.

In some cases gels are formed directly by chemical reaction. Substances that tend to separate as amorphous precipitates frequently give colloids that are gels when first formed, the hydroxides of aluminum, chromium, and silicon being good examples. Of these, the formation of silica gel by reaction between sodium silicate (water glass) and acids or between silicon tetrachloride and water has been much studied. The colloid, a sol, first formed slowly transforms spontaneously into another type of colloid, a translucent gel, that shows bluish opalescence, and this gel may later shrink, exuding the aqueous medium (syneresis). The changes are believed to arise from polymerization (the process of linking together small molecules, called monomers, into chains to form large molecules—the polymer—consisting of as many as hundreds of thousands of monomers) of the monosilicic acid first formed. The process can be repeated in one, two, or three dimensions, leading to a colloid composed of silicic acids with high molecular weight that form a rather rigid three-dimensional gel.

A property of clays of great agricultural importance is called base exchange or, more correctly, ion or cation exchange (ions are atoms or groups of atoms with electric charges; cations have positive, anions have negative charge). If, for example, a salt solution such as potassium sulfate percolates through a soil, the potassium ions (with a single positive charge) are removed and replaced by an equivalent amount of calcium ions (with two positive charges) originally present in the clay. The percentages and types of clay in soil determine its physical structure as well as its chemistry, both of which are of primary importance to the growth and health of plants. Such ion-exchange processes form the basis of the zeolite process and of similar processes. In the above case, calcium ions in the water are replaced by sodium ions, which have one less positive charge, and the water is thus softened because sodium salts are soluble whereas calcium salts are not.

Ion-exchange in clays

Clays, the finest mineral constituents of the soil, consist of layer structures of aluminosilicates (compounds of aluminum, silicon, and an alkali or alkaline-earth metal) in which some cations are easily accessible and exchangeable. The nature of these ions is not critical for maintaining the clay structure; only their total electric charge is important.

Zeolites are essentially infinite three-dimensional frameworks, potentially infinite in size, as in the various forms of silica (silicon dioxide), except that some of the silicon atoms (often about one-half) are replaced by aluminum atoms. This replacement makes the framework of the colloid negative (silicon has four positive while aluminum has three positive charges), and equivalent cations are taken up in the interstices to maintain electrical neutrality. Synthetic zeolites have been manufactured in various ways, such as by fusing a mixture of sodium carbonate, china clay or alumina, and silica (sand, quartz, etc.), the resulting glassy (vitreous) mass being then washed out (leached) with water.

Reversible systems

Reversible colloids are characterized by the fact that colloidal solutions or gels can be formed spontaneously when the dry colloid and the dispersion medium are brought together. The size of the particles in reversible colloids is determined by the molecular weight of the colloidal material in the case of polymers or by reversible association (as explained below) in the case of detergents, certain dyes, and a few other substances.

POLYMERS

Polymers are high-molecular-weight substances built up of a large number of identical (or practically identical) repeating units (generally, simple molecules), the single conglomerate molecule being of colloidal dimensions. Familiar polymeric materials that are colloids are cellulose in wood and plants and proteins in horn, hair, and wool. In nature, three different functions of polymers can be distinguished: they serve as building materials (cellulose, proteins of skin and muscle); they are storage substances (starch, glycogen); and they play a fundamental role in biochemical reactions (enzymes, nucleic acids, genes). These functions are related to the high molecular weight of the substances. The bonds between atoms holding them together in organic molecules (called primary bonds) are quite strong, not inferior to the bond strength between atoms in inorganic molecules. The bonds, however, be-

tween organic molecules (called secondary bonds) are weak and can be easily disrupted. Crystals of organic substances of low molecular weight, as a rule, are soft, indicating weak intermolecular bonds. Consequently, a structure built of organic materials in order to be strong should contain a high proportion of primary bonds, a condition that can be achieved by building the structure with polymers. Another important property of organic materials is that they can be processed at much lower temperatures than are needed to process inorganic materials, such as metals or ceramics. The intricate functions necessary to regulate biochemical reactions including reproduction can, of course, be more easily incorporated in large than in small molecules, because the large molecules allow more diversification. The commercial synthesis of polymers—broadly speaking, of plastics—seeks to mimic nature in many respects. Strong, tough, and insoluble materials are manufactured at much lower temperatures than metals or glass for an almost endless list of purposes.

Polymer chains

Most polymers, both natural and synthetic, are linearly built; *i.e.,* they are long chains. In solution, and often even in the dry state, the colloid composed of this long chain is folded or coiled. In proteins and nucleic acids the coiling is regular. In synthetic polymers and in many natural ones such as rubber and gums, the coiling is irregular and determined by chance rather than by some building principle, so that the term statistical coils is used. The large extension in solution of polymer molecules is the cause of a high viscosity (sluggish flow), which can be used for determining the molecular weight of the polymer. When linear polymers are interconnected locally, a cross-linked polymer network is obtained. Such a network, when deformed, tends to return to its original shape because that was the situation of most probable coiling. A network, therefore, cannot flow at all but behaves as a gel. The behaviour of coils cross-linked in such networks also explains both the elasticity of rubber and the swelling of dried polymer gels.

Polymers are usually subdivided, according to their origin, into natural and synthetic, although an intermediate class, man-made derivatives from natural polymers, is also a convenient classification. Some of the more important members of these three subgroups are as follows:

1. Natural polymers: cellulose, starch, rubber, proteins, nucleic acids, and the products from linseed and other drying oils.
2. Derivatives from natural polymers: nitrocellulose, cellulose acetate, regenerated cellulose (viscose, cellophane), and vulcanized rubber.
3. Purely synthetic polymers: nylon, polystyrene, polythene, synthetic rubbers, silicones, phenol–formaldehyde and urea–formaldehyde resins.

Natural polymers. Cellulose is a colloid that is the principal building constituent of the plant kingdom. Paper is derived from wood pulp. Cotton and flax, widely used in the textile industries, are almost wholly cellulose. The cellulose molecule consists of a linear array of many thousands of molecules of glucose, linked end to end in a process that eliminates one water molecule (a reaction called condensation) at each link. The chain length, or degree of polymerization, depends upon the source and the method of extraction of the cellulose.

Starch is another colloid of great importance in the plant world, not as a building material but as a convenient means of storing glucose and, hence, energy and food; wheat, maize (corn), and other grains consist largely of it. It resembles cellulose in being a product of glucose but differs in having a branched-chain structure rather than a straight-chain one. Starch can be dispersed in hot water and is then used for various purposes.

Structure of rubber

Natural rubber is a linear polymer built up from units of the compound isoprene and is a colloid with characteristic properties. The rubber is obtained in the form of a milky suspension called latex from a wide variety of plants and trees. After the latex has been coagulated by adding acetic acid, or in other ways, the resulting so-called crepe rubber is usually compounded with other materials to make it suitable for industrial use. For example, in the manufacture of tires it is compounded with carbon black and certain inorganic fillers such as zinc oxide and sulfur, and then subjected to heat treatment. The action of sulfur leads to vulcanization, as it is termed, which is the formation of cross-links between the chains through one or more sulfur atoms. Vulcanization makes the product colloid more resilient and less liable to plastic flow.

Paints usually consist of a colloidal dispersion of an inorganic coloured powder (*e.g.,* lead chromate, zinc oxide) in an oily medium containing unsaturated compounds, such as linseed oil, which, when exposed as a thin film to the action of air, rapidly oxidize and polymerize, forming a protective skin for the underlying material and the pigment. In modern paints the linseed oil is often replaced by rubber or synthetic polymers.

Derivatives from natural polymers. Cellulose forms the basis from which several important colloidal substances are derived. A group of compounds called nitrocellulose, or cellulose nitrate, made by treating cotton or other cellulose materials with a mixture of nitric acid and sulfuric acid, have properties that vary considerably, the members with high nitrate content being used as propellants (*e.g.,* cordite), the ones with less nitrate as lacquers, etc. Celluloid, the oldest commercially made plastic, consists of a mixture of cellulose nitrate with camphor. Cellulose acetate, obtained from cellulose and acetic acid, is manufactured for artificial silk, for lacquers, for safety film, and for water-desalting membranes. Cellulose can also be dissolved and reprecipitated to form the colloids cellophane and rayon.

Purely synthetic polymers. The science and the production of purely synthetic polymers were greatly developed during World War II. Synthetic rubber was and is still made on a large scale by copolymerizing a mixture of butadiene and styrene, both of synthetic origin, to give a linear polymer resembling in many ways natural rubber. Nylon is made by condensing diamines with dibasic acids, the resulting structure having a close resemblance to that of natural silks and other fibrous proteins.

Many other types of colloidal polymers are made by polymerization of vinyl compounds, for example, polythene, polyvinyl chloride, polymethyl methacrylate (Perspex or Plexiglas), and polyacrylonitrile (Orlon).

The silicones are more recently developed polymers based upon a silicon–oxygen chain. They are much more stable to heat, organic solvents, and chemical reaction than are the older plastics and thus find many uses.

Bakelite

Another well-known type of colloid, the polymer Bakelite, is made by heating phenols with formaldehyde. These are cross-linked polymers, termed thermosetting, because once formed they do not soften when warmed, in contrast with the linear polymers noted above that are thermoplastic; *i.e.,* they soften upon warming and harden again upon cooling.

Polyelectrolytes. Polymers carrying a large number of ionizable groups deserve special attention. An example of these so-called polyelectrolytes is polyacrylic acid, a long-chain polymer with a carboxylic acid group on every other carbon atom of the chain. Many natural polymers, such as gum arabic, agar, proteins, and nucleic acids, are also polyelectrolytes. Because of the mutual repulsion between the ionized groups linked to the polymer chain, the form of the coil of a linear polyelectrolyte depends greatly upon its degree of ionization. Viscosity and swelling increase strongly with ionization.

PROTEINS AND NUCLEIC ACIDS

The most important classes of hydrophilic colloids in living organisms are the proteins and the nucleic acids. The complete genetic information of the cell is laid down in the molecular structure of the nucleic acids, in particular that of deoxyribonucleic acid (DNA). This information is transported from the cell nucleus to the rest of the cell in the form of ribonucleic acid (RNA), which in its turn directs the synthesis of the proteins. Simple viruses are nucleoproteins (combinations of nucleic acids and proteins) that have the ability to reproduce themselves when injected into an uninfected host organism. Enzymes are proteins that regulate and catalyze all of the complex chemical reactions that occur in the cell and are essential

for life. Nature uses proteins also as protective colloids, for example for fat particles in milk or rubber particles in latex, and as a means of building rigid structures in which mechanical strength is required. Materials such as horn, hair, skin, cuticle, connective tissue, and muscle are largely protein in nature. Gelatin and glue are breakdown products from proteins such as collagen. Biological membranes are colloidal and consist of lipids (fatty material) associated with proteins and nucleic acids.

The nucleic acid molecules are very long chains, consisting of alternating phosphate and sugar groups, each sugar group carrying one of four organic bases as a side group. The sugar in RNA is ribose, in DNA it is deoxyribose, containing one oxygen atom less than ribose. The great variety necessary for carrying all the genetic information is obtained by variations in the order in which the four different bases are arranged along the chain. The molecular weight of the DNA is in the many millions, that of the RNA varies with their specific role in the cell and ranges from about 25,000 to a few millions.

Protein molecules are chains called polypeptides, built up by condensing a long series of α-amino acids, the simplest of which are glycine and alanine.

Classification of proteins. Proteins can be classified in various ways—for example, according to their degree of complexity, their size and shape, their solubility, their origin, and in other ways. Classified on the basis of complexity, proteins are divided into those that contain only α-amino acids (*e.g.,* egg albumin, the chief constituent of egg white) and those that contain additional groups (*e.g.,* hemoglobin, the red pigment of blood responsible for oxygen transport, in which the protein part of the molecule is linked to a substance called heme, which is one of a group of compounds called the porphyrins). On the basis of shape, proteins are frequently classed as globular or fibrous, the former having a corpuscular shape (*e.g.,* hemoglobin, egg albumin), sometimes approximately spherical, the latter being very much more elongated (*e.g.,* myosin, an important constituent of muscle). Two other important fibrous proteins—both insoluble—are keratin, of which hair and wool are largely composed, and fibroin, the chief constituent of natural silk. Under some conditions (*e.g.,* strong urea solutions), globular proteins can be unrolled, or denatured, forming fibrous proteins that are much less soluble than the native proteins.

In the early days, proteins were largely separated by means of repeated (fractional) precipitation with salts—those precipitated by half-saturated ammonium sulfate being termed globulins, and those that required fully saturated ammonium sulfate being called albumins. Alcohol can also precipitate proteins; added in less than precipitating amounts, it renders them more readily coagulated by salts. In all cases precipitation occurs most readily at the isoelectric point (see below).

Size and shape of proteins. The size and shape of soluble proteins (and of other polymers in solution) have been found chiefly by measuring rates of diffusion and their rates of movement in very strong centrifugal fields. In the ultracentrifuge, the protein molecules are subjected to intense forces up to 1,000,000 times that of gravity. By such means it has been determined that the molecular weight of many proteins lies in the range from about 10,000 to several million. The single molecules are thus of sufficient size to bring them into the colloidal range. For example, hemoglobin has a molecular weight of about 68,000; hemocyanin, the respiratory pigment of certain snails, has a molecular weight of about 5×10^6. In both these cases, the molecules are approximately spherical in shape, showing that in the colloidal particle the polypeptide chain must be folded in some way, whereas others, such as myosin, are quite elongated, with the polypeptide chain more or less fully extended.

Because of the presence of ionizable groups, protein molecules in solution normally carry a charge. In very acid solutions they are positively charged; in very alkaline ones, negatively charged. At some value of acidity the net charge is zero. This value is termed the isoelectric point and is of considerable importance. It can be determined in various ways, usually by applying the process of electrophoresis.

Isolectric point

GELS

Several lyophilic colloids can form gels, which can be considered as network structures with enmeshed liquid. The jelling component may be gelatin, agar, starch, pectin, or some other substance, which is usually present in concentrations of less than 10 percent. Table jellies are usually made from gelatin, a protein; jams are jelled by pectin, a carbohydrate derivative. Agar plates, widely employed in bacteriological work, consist of a suitable nutrient medium jelled by the addition of 1–2 percent agar, a polymer of carbohydrate prepared from certain seaweeds.

Some materials such as gelatin and agar set to a gel on cooling, whereas in other cases gelation is brought about by heating as in egg white and blancmange. In both cases, however, gelation arises from a process akin to a diminution of the solubility of the jelling agent leading to the formation of interconnections between the individual molecules.

ASSOCIATION COLLOIDS

Soaps and other detergents and a number of dyes constitute by far the most important substances classed as association colloids—*i.e.,* substances of relatively small molecular weight that associate spontaneously to form particles of a type called micelle (defined below) of colloidal size in certain concentration ranges.

Soaps. Soaps were originally limited to colloidal substances of natural occurrence, but a great number of compounds have been prepared synthetically in view of their importance as wetting agents, emulsifiers, and detergents. A few typical soaps, classified as anionic (negative), cationic (positive), and neutral according to the charge carried by the organic part of the molecule, are: sodium palmitate, anionic; sodium dodecyl sulfate, anionic; cetyltrimethylammonium bromide, cationic; polyethylene oxide derivatives, neutral. In addition, there are naturally occurring colloids that act as soaps, such as the bile salts and lecithin, substances of great biological importance.

All these colloids, as well as the colloids that are dyes, despite their diverse chemical types, have as common features a large hydrocarbon portion (also called paraffin or alkane) and a small polar (charge-carrying) group. These two parts of the molecule differ radically in their affinity for water, and it is this two-sided contradictory, or amphipathic, nature that leads to the formation of colloidal aggregates.

With aqueous solutions of the sodium and potassium salts of the long chain fatty acids, certain physical properties such as a high electrical conductivity suggest that the ions on the paraffin chain are aggregated, forming a colloidal particle—*i.e.,* an ionic micelle (defined below), as shown in Figure 53. In the case of sodium palmitate, for

Micelles

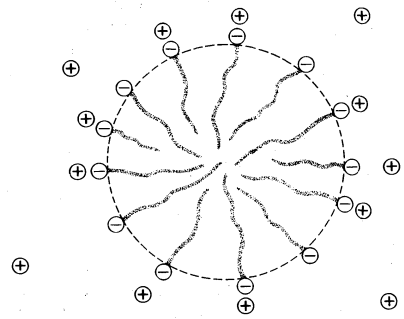


Figure 53: Structure of a micelle formed by the aggregation of paraffin-chain ions in a soap solution.

example, the diameter of the micelle, as determined from diffusion measurements, is about 40 angstroms and thus contains about 50 molecules, the molecular weight being about 14,000. In relatively dilute solutions the shape of micelles is spherical or slightly elongated. In concentrated solutions they may be spherical, elongated rods or threads, or flat structures.

Aggregation to micelles is caused by the tendency of the water to squeeze out the paraffin chains, the same factor that is largely responsible for the marked surface

activity of soap solutions. The characteristic features of soap solutions arise from the combination of surface activity and the peculiar properties of micelles in solution. The power of soap solutions to dissolve organic compounds that are insoluble or only slightly soluble in water arises directly from the presence of micelles. Micelles, as shown in Figure 53, have an interior closely resembling a droplet of liquid hydrocarbon, in which hydrocarbons and other compounds can dissolve. One important application of this phenomenon is in connection with synthetic polymers, for it enables polymerization of relatively insoluble monomers such as styrene (see above *Polymers*) to be carried out in aqueous solution or emulsion.

The soaping process

The characteristic and most widely used colloidal property of soap solutions is, of course, their detergent action. What is generally called dirt consists either of greasy materials or of particles with a greasy surface. Its removal involves three separate stages—access of the detergent to the dirty surface, the loosening or peptizing of the dirt, and finally its removal into the bulk of the solution. Because of its greasy surface, dirt is hydrophobic and not wetted by water, but adsorption of the surface-active soap molecules makes its outer surface hydrophilic and therefore greatly increases its affinity for water. In addition, the detergent, because of its powerful tendency to adsorb on, or cling to, all surfaces, will tend to displace the grease from the surface. The dirt thus loosened is detached by mechanical agitation and carried away into the solution as an emulsion or suspension, or solubilized in the interior of the micelles.

The chief drawback of fatty-acid soaps (ordinary household soaps) arises from their property of having insoluble calcium or magnesium salts. This characteristic explains the impossibility of getting a lather in seawater with ordinary soap and the necessity to soften hard water; *i.e.,* the calcium and magnesium ions must be removed in order to keep the soap in solution. Many of the synthetic detergents do not suffer from this disadvantage and thus are used in seawater soaps.

Ordinary household soap is prepared by boiling natural fats (compounds that are glycerides of the higher fatty acids) with strong caustic soda solution. After hydrolysis is complete, the soap is salted out by addition of common salt. The potassium salts, which make soft soap, are preferable to sodium salts for some purposes, because the potassium soaps have a higher solubility in cold water.

Dyes. Dye molecules, although resembling the soaps in possessing a hydrophobic part and one or more polar groups, lack their comparative uniformity as regards the relative disposition of these constituent parts. This situation results in the aggregation being much more specific and hence in a more varied colloidal behaviour. Dyestuffs are usually classed as acidic or basic (corresponding to anionic or cationic soaps) according to whether the organic ion carries a negative or a positive charge. The question of aggregation is of considerable importance in connection with the colloidal aspect of the dyeing process. The size of the aggregates, or micelles, can be determined by the same methods as for soaps, diffusion being regarded as the best. The degree of association is increased by increasing salt content, which explains the influence of salts in dyeing.

The dyeing process

The dyeing process has naturally excited considerable interest throughout history, although it is by no means fully understood. The materials to be dyed are usually colloids of a fibrous nature, consisting of macromolecules of vegetable, animal, or purely synthetic origin. Most vegetable fibres, such as linen and cotton, have the colloidal structure of cellulose as the fundamental constituent; wool and silk are the principal animal fibres, and, of the purely synthetic materials, nylon is without doubt the best known. In dyeing a fibre, the coloured molecules have to diffuse into the interior, in which they are held in position by precipitation or by some other reaction. The behaviour of wool and silk is probably the simplest, for the major factor in bringing about combination is believed to be salt formation between the dye anions on the fibre. With cellulose (cotton and rayon) the affinity of the dye for the fibre is probably caused by van der Waals forces and hydrogen bonding.

Dyes also find an important use in the identification of biological materials (*e.g.,* bacteria) by staining methods. These depend upon the different affinities of the enveloping constituents for various dyes, the principles involved being closely related to those in ordinary dyeing. Dyestuffs of the basic type (*e.g.,* Gentian Violet) find use as antiseptics, resembling the cationic soaps in their bactericidal powers. (J.T.G.O.)

LOW-TEMPERATURE PHENOMENA

The term low-temperature phenomena refers to the behaviour of matter at temperatures closer to absolute zero (−273.15° C [−459.67° F]) than to room temperature. At such temperatures the thermal, electric, and magnetic properties of many substances undergo great change, and, indeed, the behaviour of matter may seem strange when compared to that at room temperature. Superconductivity and superfluidity can be cited as two such phenomena that occur below certain critical temperatures; in the former, many chemical elements, compounds, and alloys show no resistance whatsoever to the flow of electricity, and, in the latter, liquid helium can flow through tiny holes impervious to any other liquid.

Superconductivity

Superconductivity was discovered in 1911 by a Dutch physicist, Heike Kamerlingh Onnes, who was awarded the Nobel Prize for Physics for 1913 for his low-temperature research. Kamerlingh Onnes found that the electrical resistivity of a mercury wire disappears suddenly when it is cooled below a temperature of about 4 K (−269° C [−452° F]); absolute zero is 0 K, the temperature at which all matter loses its disorder. He soon discovered that a superconducting material can be returned to the normal state either by passing a sufficiently large current through it or by applying a sufficiently strong magnetic field to it.

For many years it was believed that, except for the fact that they had no electrical resistance (*i.e.,* that they had infinite electrical conductivity), superconductors had the same properties as normal materials. This belief was shattered in 1933 by the discovery that a superconductor is highly diamagnetic; that is, it is strongly repelled by and tends to expel a magnetic field. This phenomenon, which is very strong in superconductors, is called the Meissner effect for one of the two men who discovered it. Its discovery made it possible to formulate, in 1934, a theory of the electromagnetic properties of superconductors that predicted the existence of an electromagnetic penetration depth (see below *The Meissner effect*), which was first confirmed experimentally in 1939. In 1950 it was clearly shown for the first time that a theory of superconductivity must take into account the fact that free electrons in a crystal are influenced by the vibrations of atoms that define the crystal structure, called the lattice vibrations. In 1953, in an analysis of the thermal conductivity of superconductors, it was recognized that the distribution of energies of the free electrons in a superconductor is not uniform but has a separation called the energy gap.

Meissner effect

The theories referred to thus far served to show some of the interrelationships between observed phenomena but did not explain them as consequences of the fundamental laws of physics. For almost 50 years after Kamerlingh Onnes' discovery, theorists were unable to develop a fundamental theory of superconductivity. Finally, in 1957, such a theory was presented by the physicists John Bardeen, Leon N. Cooper, and John Robert Schrieffer of the United States; it won for them the Nobel Prize for Physics for 1972. It is now called (for them) the BCS theory, and all later theoretical work is based on it. The BCS theory also

The BCS theory

provided a foundation for an earlier model that had been introduced by the Soviet physicists Lev Davidovich Landau and Vitaly Lazarevich Ginzburg (1950). This model has been useful in understanding electromagnetic properties, including the fact that any internal magnetic flux in superconductors exists only in discrete amounts (instead of in a gradual and continuously changing spectrum of values), an effect called the quantization of magnetic flux. This flux quantization, which had been predicted from quantum-mechanical principles, was first observed experimentally in 1961.

In 1962 the British physicist Brian D. Josephson predicted that two superconducting objects placed in electrical contact would display certain remarkable electromagnetic properties. These properties have since been observed in a wide variety of experiments, demonstrating quantum-mechanical effects on a macroscopic scale.

The theory of superconductivity has been tested in a wide range of experiments, involving, for example, ultrasonic absorption studies, nuclear-spin phenomena, low-frequency infrared absorption, and electron-tunnelling experiments (see below *Energy gaps*). The results of these measurements have brought understanding to many of the detailed properties of various superconductors.

THERMAL PROPERTIES OF SUPERCONDUCTORS

Superconductivity is a startling departure from the properties of normal (that is, non-superconducting) conductors of electricity. In materials that are electrical conductors, some electrons are not bound to individual atoms but are free to move through the material; their motion constitutes the electric current. In normal conductors these so-called conduction electrons are scattered by impurities, dislocations, grain boundaries, and lattice vibrations (phonons). In a superconductor, however, there is an ordering among the conduction electrons that prevents this scattering. Consequently, electric current can flow with no resistance at all. This ordering of the electrons, called Cooper pairing, involves the momenta of the electrons rather than their positions. The energy per electron that is associated with this ordering is extremely small, typically about a thousandth of the amount by which the energy per electron changes when a chemical reaction takes place. One reason that superconductivity remained unexplained for so long was the smallness of the energy changes that attend the transition between normal and superconducting states. In fact, many incorrect theories of superconductivity were advanced before the BCS theory was proposed. For fuller details on electrical conduction in metals, and the effects of temperature and other influences, see the article ELECTRICITY AND MAGNETISM.

Superconducting materials

Hundreds of materials are known to become superconducting at low temperatures. Twenty-seven of the chemical elements, all of them metals, are superconductors in their usual crystallographic forms at atmospheric pressure. Among these are commonly known metals such as aluminum, tin, lead, and mercury, and less common ones such as rhenium, lanthanum, and protactinium. In addition, 11 chemical elements that are metals, semimetals, or semiconductors (see above) are superconductors at low temperatures and high pressures. Among these are uranium, cerium, silicon, and selenium. Bismuth and five other elements, though not superconducting in their usual crystallographic form, can be made superconducting by preparing them in a highly disordered form, which is stable at extremely low temperatures. Superconductivity is not exhibited by any of the magnetic elements chromium, manganese, iron, cobalt, and nickel.

Most of the known superconductors are alloys or compounds. It is possible for a compound to be superconducting even if the chemical elements constituting it are not; examples are disilver fluoride (Ag_2F) and a compound of carbon and potassium (C_8K). Some semiconducting compounds, such as tin telluride (SnTe), become superconducting if they are properly doped with impurities.

Transition temperatures. The temperature below which a substance is superconducting is called the transition temperature (T_c). The vast majority of the known superconductors have transition temperatures that lie between 1 K and 10 K. Of the chemical elements, tungsten has the lowest transition temperature, 0.015 K, and niobium the highest, 9.2 K. The transition temperatures of some of the commonly known superconducting elements are indicated in Table 5.

Table 5: Transition Temperatures and Low-Temperature Values of Critical Magnetic Field of Some Superconducting Elements

	Tc (Kelvin)	*Ho* (oersted)
Zinc	0.88	53
Aluminum	1.20	99
Indium	3.41	282
Tin	3.72	306
Mercury	4.15	411
Lead	7.19	803

The transition temperature is usually very sensitive to the presence of magnetic impurities. A few parts per million of manganese in zinc, for example, lowers the transition temperature considerably.

Specific heat and thermal conductivity. The thermal properties of a superconductor can be compared with those of the same material at the same temperature in a normal state. (The material can be forced into a normal state at the temperature by a large enough magnetic field.)

When a small amount of heat is put into a system, some of the energy is used to increase the lattice vibration (an amount that is the same for a system in the normal and in the superconducting state) and the remainder is used to increase the energy of the conduction electrons. The electronic specific heat (C_e) of the electrons is defined as the ratio of that portion of the heat used by the electrons to the rise in temperature of the system. Figure 54 shows how the specific heat of the electrons in a superconductor varies with the absolute temperature (T) in the normal and in the superconducting state. It is evident from the figure that the electronic specific heat in the superconducting state, designated C_{es}, is smaller than in the normal state, designated C_{en}, at low enough temperatures, but that C_{es} becomes larger than C_{en} as the transition temperature T_c is approached, at which point it drops abruptly to C_{en}. Precise measurements have indicated that, at temperatures considerably below the transition temperature, the logarithm of the electronic specific heat is inversely proportional to the temperature. This temperature dependence, together with the principles of statistical mechanics, strongly suggests that there is a gap in the distribution of energy levels available to the electrons in a superconductor, so that a minimum energy is required for the

Temperature dependence of electronic specific heat

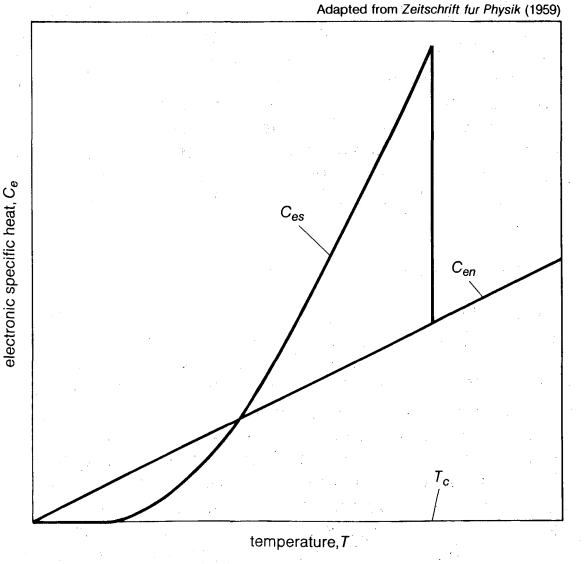


Figure 54: Specific heat in the normal (C_{en}) and superconducting (C_{es}) states as a function of absolute temperature. The two functions are identical at the transition temperature (T_c) and higher.

excitation of each electron from a state below the gap to a state above the gap.

The heat flow per unit area of a sample equals the product of the thermal conductivity (K) and the temperature gradient ∇T. This may be expressed as $J_Q = -K \nabla T$, the minus sign indicating that heat always flows from a warmer to a colder region of a substance.

It is a fact that thermal conductivity in the normal state (K_n) approaches thermal conductivity in the superconducting state (K_s) as the temperature (T) approaches the transition temperature (T_c) for all materials, whether they are pure or impure. This suggests that the energy gap (Δ) for each electron approaches zero as the temperature (T) approaches the transition temperature (T_c). This would also account for the fact that the electronic specific heat in the superconducting state (C_{es}) is higher than in the normal state (C_{en}) near the critical temperature: as the temperature is raised toward the transition temperature (T_c), the energy gap in the superconducting state decreases, the number of thermally excited electrons increases, and this requires the absorption of heat.

Energy gaps. As stated above, the thermal properties of superconductors indicate that there is a gap in the distribution of energy levels available to the electrons, and so a finite amount of energy, designated as delta (Δ), must be supplied to an electron to excite it. This energy is maximum at absolute zero (designated Δ_0) and changes little with increase of temperature until the transition temperature is approached, where Δ decreases to zero, its value in the normal state. The BCS theory predicts an energy gap with just this type of temperature dependence.

Cooper electron pairs

According to the BCS theory, there is a type of electron pairing (electrons of opposite spin acting in unison) in the superconductor that is important in interpreting many superconducting phenomena. The electron pairs, called Cooper pairs, are broken up as the superconductor is heated. Each time a pair is broken, an amount of energy that is at least as much as the energy gap (Δ) must be supplied to each of the two electrons in the pair, so an energy at least twice as much (2Δ) must be supplied to the superconductor. The value of twice the energy gap at 0 K (which is $2\Delta_0$) might be assumed to be higher when the transition temperature of the superconductor is higher. In fact, the BCS theory predicts a relation of this type; namely, that the energy supplied to the superconductor at absolute zero would be equal to 3.53 times the product of the Boltzmann's constant, k (1.38×10^{-23} joule per degree Kelvin), and the transition temperature, T_c, in degrees Kelvin; *i.e.*, $2\Delta_0 = 3.53\, kT_c$.

The energy gap (Δ) can be measured most precisely in a tunnelling experiment (a process in quantum mechanics that allows an electron to escape from a metal without acquiring the energy required along the way according to the laws of classical physics). In this experiment, a thin insulating junction is prepared between a superconductor and another metal, assumed here to be in the normal state. In this situation, electrons can quantum-mechanically tunnel from the normal metal to the superconductor if they have sufficient energy. This energy can be supplied by applying a negative voltage (V) to the normal metal, with respect to the voltage of the superconductor.

Tunnelling will occur if eV, the product of the electron charge, e (-1.60×10^{-19} coulomb), and voltage, is at least as large as the energy gap Δ. The current flowing between the two sides of the junction is small up to a voltage equal to the energy gap divided by the electron charge ($V = \Delta/e$), but then it rises sharply. This provides an experimental determination of the energy gap (Δ). In describing this experiment it is assumed here that the tunnelling electrons must get their energy from the applied voltage, rather than from thermal excitation.

MAGNETIC AND ELECTROMAGNETIC PROPERTIES OF SUPERCONDUCTORS

Critical field. One of the ways in which a superconductor can be forced into the normal state is by applying a magnetic field. The weakest magnetic field that will cause this transition is called the critical field (H_c) if the sample is in the form of a long, thin cylinder or ellipsoid and the field is oriented parallel to the long axis of the sample. (In other configurations the sample goes from the superconducting state into an intermediate state, in which some regions are normal and others are superconducting, and finally into the normal state.) The critical field increases with decreasing temperature. For the superconducting elements, its values (H_0) at absolute zero range from 1.1 oersted for tungsten to 830 oersteds for tantalum. Values of H_0 are listed in Table 5 for some common superconducting elements.

These remarks about the critical field apply to ordinary (so-called type I) superconductors. In the following section the behaviour of other (type II) superconductors is examined.

The Meissner effect. As was stated above, a type I superconductor in the form of a long, thin cylinder or ellipsoid remains superconducting at a fixed temperature as an axially oriented magnetic field is applied, provided the applied field does not exceed a critical value (H_c). Under these conditions, superconductors exclude the magnetic field from their interior, as could be predicted from the laws of electromagnetism and the fact that the superconductor has no electrical resistance. A more astonishing effect occurs if the magnetic field is applied in the same way to the same type of sample at a temperature above the transition temperature and is then held at a fixed value while the sample is cooled. It is found that the sample expels the magnetic flux as it becomes superconducting. This is called the Meissner effect. Complete expulsion of the magnetic flux (a complete Meissner effect) occurs in this way for certain superconductors, called type I superconductors, but only for samples that have the described geometry. For samples of other shapes, including hollow structures, some of the magnetic flux can be trapped, producing an incomplete or partial Meissner effect. Figure 55 shows graphs of the magnetization of a type I sample (with a complete Meissner effect) and of a type II sample (with an incomplete Meissner effect).

Behaviour of type II superconductors

Type II superconductors have a different magnetic behaviour. Examples of materials of this type are niobium and vanadium (the only type II superconductors among the chemical elements) and some alloys and compounds. As a sample of this type, in the form of a long, thin cylinder or ellipsoid, is exposed to a decreasing magnetic field that is axially oriented with the sample, the increase of magnetization, instead of occurring suddenly at the critical field (H_c), sets in gradually. Beginning at the upper critical field (H_{c2}), it is completed at a lower critical field (H_{c1}; see Figure 55). If the sample is of some other shape, is hollow, or is inhomogeneous or strained, some magnetic flux frequently remains trapped, and some magnetization of the sample remains after the applied field is completely removed. Known values of the upper critical field extend all the way up to 6×10^5 oersteds, the value for the compound of lead, molybdenum, and sulfur with formula $PbMo_6S_8$.

Reprinted with permission from K. Maki, *Physics* (1964); Pergamon Press

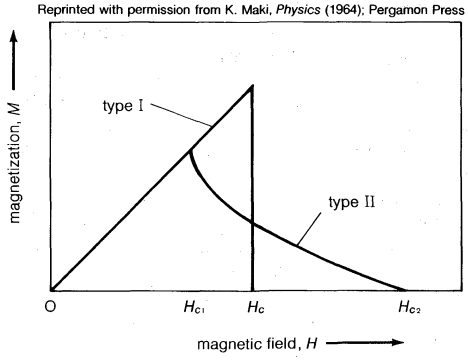


Figure 55: Magnetization as a function of magnetic field for a type I superconductor and a type II superconductor. For type I, magnetic flux is expelled, producing a magnetization (M) that increases with magnetic field (H) until a critical field (H_c) is reached, at which it falls to zero as with a normal conductor. A type II superconductor has two critical magnetic fields (H_{c1} and H_{c2}); below H_{c1}, type II behaves as type I, and above H_{c2} it becomes normal.

Electromagnetic penetration depth

The expulsion of magnetic flux by type I superconductors in fields below the critical field (H_c) or by type II superconductors in fields below H_{c1} is never quite as complete as has been stated in this simplified presentation, because the field always penetrates into a sample for a small distance, known as the electromagnetic penetration depth. Values of the penetration depth for the superconducting elements at low temperature lie in the range from about 390 to 1,300 angstroms. As the temperature approaches the critical temperature, the penetration depth becomes extremely large.

High-frequency electromagnetic properties. The foregoing descriptions have pertained to the behaviour of superconductors in the absence of electromagnetic fields or in the presence of steady or slowly varying fields; the properties of superconductors in the presence of high-frequency electromagnetic fields, however, have also been studied.

The energy gap in a superconductor has a direct effect on the absorption of electromagnetic radiation. At low temperatures, at which a negligible fraction of the electrons are thermally excited to states above the gap, the superconductor can absorb energy only in a quantized amount that is at least twice the gap energy (at absolute zero, $2\Delta_0$). In the absorption process, a photon (a quantum of electromagnetic energy) is absorbed, and a Cooper pair is broken; both electrons in the pair become excited. The photon's energy (E) is related to its frequency (ν) by the Planck relation, $E = h\nu$, in which h is Planck's constant (6.63×10^{-34} joule second). Hence the superconductor can absorb electromagnetic energy only for frequencies at least as large as twice the energy gap divided by Planck's constant, or $2\Delta_0/h$.

Magnetic-flux quantization. The laws of quantum mechanics dictate that electrons have the properties of waves and that the properties of an electron can be summed up in what is called a wave function. If several wave functions are in phase (*i.e.*, act in unison), they are said to be coherent. The theory of superconductivity indicates that there is a single, coherent, quantum-mechanical wave function that determines the behaviour of all of the superconducting electrons. As a consequence, a direct relationship can be shown to exist between the velocity of these electrons and the magnetic flux (Φ) enclosed within any closed path inside the superconductor. Indeed, inasmuch as the magnetic flux arises because of the motion of the electrons, the magnetic flux can be shown to be quantized; *i.e.*, the intensity of this trapped flux can change only by units of Planck's constant divided by twice the electron charge.

Fluxoids

When a magnetic field enters a type II superconductor (in an applied field between the lower and upper critical fields, H_{c1} and H_{c2}), it does so in the form of quantized fluxoids, each carrying one quantum of flux. These fluxoids tend to arrange themselves in regular patterns that have been detected by electron microscopy and by neutron diffraction. If a current is passed through the superconductor, the fluxoids move. This motion leads to energy dissipation that can heat the superconductor and drive it into the normal state. The maximum current per unit area that a superconductor can carry without being forced into the normal state is called the critical current density (J_c). In making wire for superconducting high-field magnets, manufacturers try to fix the positions of the fluxoids by deliberately making the wire inhomogeneous in composition.

Josephson currents. If two superconductors are separated by an insulating film that forms a low-resistance junction between them, it is found that Cooper pairs can tunnel from one side of the junction to the other. (This process occurs in addition to the single-particle tunnelling already described.) Thus a flow of electrons, called the Josephson current, is generated and is intimately related to the phases of the coherent quantum-mechanical wave function for all of the superconducting electrons on the two sides of the junction. It was predicted that several novel phenomena should be observable, and experiments have demonstrated them. These are collectively called the Josephson effect or effects.

The first of these phenomena is the passage of current through the junction in the absence of a voltage across the junction. The maximum current that can flow at zero voltage depends on the magnetic flux (Φ) passing through the junction as a result of the magnetic field generated by currents in the junction and elsewhere. The dependence of the maximum zero-voltage current on the magnetic field applied to a junction between two superconductors is shown in Figure 56.

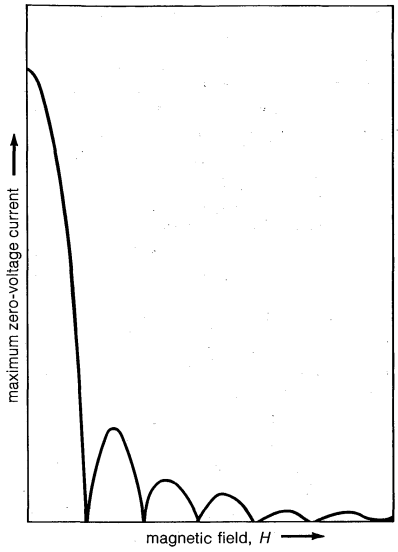


Figure 56: Maximum zero-voltage (Josephson) current passing through a junction by Cooper-pair tunnelling as a function of magnetic field.

Oscillating currents

A second type of Josephson effect is an oscillating current resulting from a relation between the voltage across the junction and the frequency (ν) of the currents associated with Cooper pairs passing through the junction. The frequency (ν) of this Josephson current is equal to two times the product of the electron charge (e) and the impressed voltage (V) divided by Planck's constant (h); that is, $\nu = 2eV/h$. Thus the frequency increases by 4.84×10^{14} hertz (cycles per second) for each additional volt applied to the junction. This can be demonstrated in various ways. The voltage can be established with a source of direct current (dc) power, for instance, and the oscillating current be detected by the electromagnetic radiation of frequency (ν) that it generates. Another method is to expose the junction to radiation of another frequency (ν') generated externally. It is found that a graph of the dc current versus voltage has current steps at values of the voltage corresponding to Josephson frequencies that are integral multiples (n) of the external frequency ($\nu = n\nu'$); that is, the voltage is equal to an integer times Planck's constant times the external frequency divided by twice the electron charge, or $V = nh\nu'/2e$. The observation of current steps of this type has made it possible to measure the ratio of Planck's constant to the charge on the electron (h/e) with far greater precision than by any other method and has therefore contributed to a knowledge of the fundamental constants of nature.

The Josephson effect has been used in the invention of novel devices for extremely high-sensitivity measurements of currents, voltages, and magnetic fields.

HIGHER TEMPERATURE SUPERCONDUCTIVITY

Ever since Kamerlingh Onnes discovered that mercury becomes superconducting at temperatures less than 4 K, scientists have been searching for superconducting materials with higher transition temperatures. Until 1986, a compound of niobium and germanium (Nb_3Ge) had the highest known transition temperature, 23 K, less than a 20-degree increase in 75 years. Most researchers expected that the next increase in transition temperature would be found in a similar metallic alloy and that the rise would be only one or two degrees. In 1986, however, the Swiss physicist Karl Alex Müller and his West German associ-

ate, Johannes Georg Bednorz, discovered, after a three-year search among metal oxides, a material that had an unprecedentedly high transition temperature of about 30 K. They were awarded the Nobel Prize for Physics in 1987, and their discovery immediately stimulated groups of investigators in China, Japan, and the United States (especially the American physicists Paul C.W. Chu and Maw-Kuen Wu and their colleagues) to produce superconducting oxides with even higher transition temperatures.

These high-temperature superconductors are ceramics. They contain lanthanum, yttrium, or another of the rare-earth elements (with the exception of cerium, praseodymium, promethium, and terbium); barium or strontium (both alkaline-earth elements); copper; and oxygen. The compounds have been divided into two classes. Materials in class 1 have transition temperatures around 40 K and chemical formulas of the type $La_{2-x}A_xCuO_{4-y}$, where A represents the alkaline-earth element. The general formula for class-2 materials, which have transition temperatures around 90 K, is $RBa_2Cu_3O_{7-y}$, where R is the rare-earth element and y is a number between 0.1 and 0.5.

The compounds in both classes have crystal structures containing planes of Cu and O atoms, and the class-2 superconductors also have chains of Cu and O atoms. The roles played by these planes and chains have come under intense investigation. Varying the oxygen content of the materials dramatically changes their transition temperatures, critical magnetic fields, and other properties. Single crystals of the high-temperature superconductors are anisotropic; *i.e.,* their properties associated with a direction, such as the critical fields or the critical current density, are highly dependent on the angle between that direction and the rows of atoms in the crystal.

These ceramics are type II superconductors. They exhibit zero resistance, strong diamagnetism, the Meissner effect, magnetic flux quantization, the Josephson effects, an electromagnetic penetration depth, an energy gap for the superconducting electrons, and the characteristic temperature dependences of the specific heat and the thermal conductivity that are described above. Thus, it is clear that the conduction electrons in these materials form the Cooper pairs used to explain superconductivity in the BCS theory. The central conclusions of the BCS theory are demonstrated. Indeed, that theory guided Bednorz and Müller in their search for high-temperature superconductors. It is not known, however, why the transition temperatures of these oxides are so high. It was generally believed that the members of a Cooper pair are bound together because of interactions between the electrons and the lattice vibrations (phonons), but it is unlikely that these interactions are strong enough to explain transition temperatures as high as 90 K. Many theoretical models have been suggested to provide the explanation, but it is not known which, if any, are correct.

Applications of high-temperature superconductors

Films of the new materials can carry currents in the superconducting state that are large enough to be of importance in making many devices. Possible applications of the high-temperature superconductors in thin-film or bulk form include the construction of magnets for medical magnetic-imaging devices, magnetic energy-storage systems, levitated passenger trains for high-speed travel, motors, generators, transformers, transmission lines, computer parts (logic devices, memory elements, switches, and interconnects), oscillators, amplifiers, particle accelerators, and highly sensitive devices for measuring magnetic fields, voltages, or currents. The principal advantages of these superconducting devices would be their low power dissipation, high operating speed, and extreme sensitivity.

Equipment made with the high-temperature superconductors would also be more economical to operate because they can be cooled with inexpensive liquid nitrogen (boiling point, 77 K) rather than with costly liquid helium (boiling point, 4.2 K). The ceramics have problems, however, which must be overcome before useful devices can be made from them. These problems include brittleness, instabilities of the materials in some chemical environments, and a tendency for impurities to segregate at surfaces and grain boundaries, where they interfere with the flow of high currents in the superconducting state. (D.M.Gi.)

Superfluidity

The superfluid properties of liquid helium-4 are heralded by a seemingly infinite lambda- (λ-) shaped heat-capacity peak at a temperature $T_\lambda = 2.2$ K as the liquid is cooled under its own vapour pressure (see CHEMICAL ELEMENTS: *Noble gases*). Below the lambda temperature the liquid has a thermal conductivity some 3,000,000 times that of the liquid just above the lambda temperature, and it can flow or, at least, a part of it can flow without friction and without entropy transport through tiny capillaries and through leaks that just above T_λ are impassable not only to liquid but also to gaseous helium-4. Many other strange phenomena have been investigated.

Theory of helium-4 behaviour

Early contributions to a theoretical understanding of liquid helium-4 behaviour by two physicists, Polish-born Fritz London and Lev D. Landau of the Soviet Union, were originally believed to exclude one another, but, with the later work of Richard Feynman and Chen Ning Yang, United States physicists, and a Norwegian chemist, Lars Onsager, their approaches have been seen to be complementary. Following London's interpretation, stress is put on the fact that helium-4 atoms are bosons; *i.e.,* that they contain an even number of elementary particles (electrons, protons, neutrons) and as such any number of them can occupy the same quantum state of momentum. Fermions, on the other hand, of which the isotope helium-3 is an example, contain an odd number of elementary particles, and no more than a few (only two if they have minimum nuclear or electronic spin values) can occupy the same momentum quantum state. It is assumed that below the lambda temperature a finite fraction of the helium-4 atoms are condensed into the zero momentum state with zero entropy and that these constitute a superfluid component in the liquid. This component is not spatially condensed but distributed uniformly through the liquid in a way not unlike the paired electrons of superconductivity theory. In fact a pair of electrons is a boson.

Liquid helium-3 does not exhibit the superfluid properties of liquid helium-4, thus strengthening the belief that these properties are a result of a boson momentum ordering. Whether or not the helium-3 atoms as fermions will undergo a momentum ordering analogous to that of superconducting electrons is an open question.

Special methods for obtaining low temperatures

ADIABATIC COOLING

When changes are made in a closed or isolated system without gain or loss of heat, the phenomenon is adiabatic. For example, an expanding gas that neither gains nor loses heat is expanding adiabatically. Primary gas thermometry depends on the variation in pressure of a gas with temperature and as such provides a temperature scale independent of the properties of any particular substance. The relation of this scale (T° K) to the Celsius scale (t° C) is $T = t + 273.15$, such that absolute zero is −273.15° C. A convenient way of experimentally determining low temperature is to measure the vapour pressure of a sample of liquefied gas in thermal equilibrium with the sample the temperature of which is desired. This measurement requires use of tables of temperature as a function of the vapour pressure. To prepare these tables a few primary gas thermometric measurements must be made, which are supplemented at intermediate temperatures by simultaneously measuring the vapour pressure and one or more physical properties of an inserted probe that can be related to the temperature by means of empirical equations. These equations must express the probe property as a function of temperature and of a few constants the values of which are known by direct measurement of the probe property and the temperature at a few points. Use of the electrical resistance of certain pure metals or of carbon and the magnetic susceptibility of certain paramagnetic salts is common. Vapour-pressure thermometry with liquid helium-4 can be carried down to about 1 K and with liquid helium-3 to about 0.3 K.

Gas thermometric measurements

By the third law of thermodynamics it is impossible to

reduce the temperature of any sample precisely to zero, but in practice temperatures as low as 10^{-5} K have been attained. Various liquefied gases boiling off under reduced pressure can serve admirably to cool samples down to about 1 K. To liquefy these gases they may be compressed to high pressure at a constant temperature (the heat generated by the compression being transferred via an exchange gas in an intermediate chamber to a liquid that can vaporize at constant temperature) and then isolated thermally (*i.e.,* heat is not allowed to flow in or out of the system) and allowed to expand to low pressure. In the expansion, work is done by the gas, and thus it cools. A series of such steps may be needed until the gas is cooled below its critical temperature and a portion of it liquefies. Alternatively, a process based on the Joule–Thomson phenomenon of adiabatic stationary flow can be made use of. In this process, when a gas flows through a porous plug or expansion valve from a constant high pressure to a constant low pressure, it undergoes a temperature change that depends in sign and magnitude on a balance between the cooling effect of increasing the separation of the molecules and the heating effect of work being done on the gas in pushing it through the valve. At low enough initial temperature and pressure, the effect is a cooling one. An initial expansion technique is quite often used to cool a gas until it is ready for a final Joule–Thomson liquefaction.

Spin ordering

To cool a substance below $T_0 = 1$ K there must exist one or more sets of energy levels with spacings still smaller than kT_0, the entropy of which can be manipulated by varying an external parameter (quantity). In practice there are two such sets of energy levels available: states of net electron-spin–magnetic-dipole orientations associated with paramagnetic atoms and states of nuclear spin–magnetic-dipole orientations associated with some atomic nuclei. It was suggested in 1926–27, independently by the Dutch physicist Peter Debye and the American William Giauque, that, if electron spins could be isothermally (*i.e.,* without change in temperature) ordered at about 1 K in an external magnetic field, thus removing much of their entropy by evolving heat that could be conducted away, then a subsequent adiabatic demagnetization (being practically a constant entropy process in which energy is lost; *i.e.,* work is done by the spin system as the field is turned off) would result in a much lowered temperature for the spin system. This loss would then result in a lowered temperature for the paramagnetic salt, and any sample in contact with it, as these would have a negligible heat capacity compared with that of the spins. Temperatures as low as 10^{-2} to 10^{-3} K have been obtained with this widely used technique.

Adiabatic demagnetization

Temperature measurement below 1 K is carried out by measurement of the magnetic susceptibility of a paramagnetic salt and then relating this property to the absolute-temperature scale. This last procedure must be done separately for each salt, by determining the susceptibility change for known heat inputs and for known entropy changes arising from demagnetization experiments from known liquid-helium temperatures. (For a general survey of low-temperature applications, see REFRIGERATION).

ADIABATIC DILUTION

Another technique to obtain temperatures to about 10^{-2} K, originally suggested by German physicist Heinz London in 1951 but only experimentally realized since 1965, is the adiabatic dilution of liquid helium-3 by liquid helium-4. It is difficult to explain without mathematical formulation, and only a brief reference to the principles can be made here.

Below 0.1 K there is a small, nearly constant limiting solubility of helium-3 in concentrated helium-4 that persists to absolute zero. At the starting temperatures of about 0.5 K reached in ordinary liquid helium-3, helium-4 is entirely superfluid with zero entropy, providing a noninteracting inert background for the helium-3, which in concentrated solutions behaves like a pseudo-one-component liquid and in dilute solutions like a pseudo-one-component gas. The crucial step for cooling, which is a dilution of a concentrated helium-3 solution across the two-phase region to the dilute side, is thus analogous to an ordinary adiabatic evaporation of a liquid. Continuous-refrigeration devices based on this technique have been developed.

(C.E.He./Ed.)

HIGH-PRESSURE PHENOMENA

High-pressure phenomena are those changes that occur in ordinary matter when subjected to pressures that may range from 10 to millions of times greater than normal atmospheric pressure. Under such conditions, matter is compressed, and chemical and physical properties are altered; the material may even become unrecognizable. Laboratory studies conducted at these pressures may also involve temperatures that range from near absolute zero to far beyond the melting point of any element under normal conditions.

Range of pressures

The range of pressure in nature is extreme: at the deepest ocean depths (about 11 kilometres [seven miles]), the pressure is roughly one kilobar (1,000 bars); at the base of the Earth's crust (about 40 kilometres' [25 miles'] depth), the pressure is approximately 10 kilobars; at the base of the Earth's mantle (about 2,900 kilometres [1,800 miles]), the pressure is 1.3 megabars (1,300,000 bars); and at the Earth's centre, it is 3.4 megabars. Pressures at the centre of Saturn and Jupiter are estimated to be 20 and 100 megabars, respectively, and those of some stellar interiors are thought to be 10^{16} bars or more. While it is difficult to understand those laws of physics and chemistry that apply to material properties over the wide range of pressure that occurs in the universe, present technology makes it possible to study materials at all pressures found in the Earth. Studies using explosive techniques have attained pressures greater than those at the centre of Saturn.

The amount of pressure necessary to create similar effects on different materials is highly dependent on the materials involved. Whereas a pressure change of 0.2 bar on a gas, for example, causes the volume of the gas to change by 20 percent, a similar volume change requires 10,000 bars in a liquid and 200,000 bars in a solid. Hence the selection of materials and effects to be studied has been strongly influenced by the historical development of high-pressure apparatus.

Before the 20th century, when research was concerned primarily with the effects of pressure on gases and liquids because of apparatus limitations, considerable work was done at pressures of three kilobars and over a temperature range of 200° C (392° F). Typical of such investigations were those concerned with the volume behaviour of liquids and with the existence of gas–liquid critical points (see above *Phase changes and equilibria*). At the critical point the temperature and pressure are such that the transition from a gas to a liquid takes place with zero change in volume and energy. This transition involves two factors: (1) the critical temperature, which is the highest temperature at which a gas can be liquefied by pressure; and (2) the critical pressure, which is the pressure of a gas at its critical point. The important question of a critical point between liquid and solid matter was also raised during the 19th century, but the pressures available for research at that time were far too small to settle the matter.

Modern approach to research

Modern high-pressure research was made possible primarily by the work of the U.S. physicist Percy Williams Bridgman, who received the 1946 Nobel Prize for his investigation of high-pressure phenomena. Except for those techniques that involve the use of explosives, all current high-pressure apparatus is derived from equipment that Bridgman designed and employed (see below).

Physical and chemical effects of high pressure

In order to describe high-pressure phenomena over extremely wide ranges of variables, it is necessary to use

mathematical statements that relate small increments of the various changes taking place simultaneously during the overall transformation of conditions and reveal graphically what is occurring.

Modern high-pressure research in phase changes, particularly the work of Bridgman, has revealed that a myriad of transformations occur as a result of atomic rearrangements, many of which involve gross changes in the physical properties of a substance.

As previously discussed, two or more phases can exist in equilibrium, a balanced coexistence that is governed by thermodynamic relationships. These relationships are formulated into a concept called Gibbs free energy, which is the difference between the heat content of a substance (called enthalpy) and the product of its temperature and entropy. The equation for this thermodynamic relationship is written in symbols: $G = H - TS$, in which G is the Gibbs free energy of the phases in equilibrium, H is the enthalpy, T is the absolute temperature, and S is the entropy of the phase mixture. As the equation indicates, at any constant pressure and temperature, when the phases are in equilibrium, the mixture has the lowest Gibbs free energy, relative to the energy of the mixture at any other configuration.

By differentiating this equation it may be seen that there is a simple relationship between Gibbs free energy and the volume and entropy of the phase. The partial derivative of the free energy with respect to the pressure (P), the temperature (T) being held constant, is equal to the volume (V). This is written $(\partial G/\partial P)_T = V$. The partial derivative of the free energy with respect to the temperature (T)—the pressure (P) being held constant—is equal to minus the entropy (S). It is written $(\partial G/\partial T)_P = -S$.

If entropy and volume are known as a function of the pressure and temperature for two phases, these equations may be integrated and an equilibrium-phase boundary thus calculated, separating the regions in P–T space (a phase plot of pressure versus temperature) in which the two differing states are stable. Stability does not imply that a phase transformation will take place, however; for example, with care, most liquids may be preserved as liquids for lengthy periods of time below their freezing points, at temperatures where the solid is the normal stable phase. Such supercooling (or any other such sluggish transformation) is known as metastable behaviour, marked by only a slight margin of stability for that phase, because any disturbance will bring about a transformation to the stable phase. A supercooled liquid, for example, becomes solid rapidly if a minute crystal is dropped into it.

Supercooling

EFFECTS ON MELTING

Melting phenomena are best discussed with reference to the above equations. Most melting curves have been studied only up to less than 100 kilobars by static apparatus; melting has been detected, however, at higher pressures in some dynamic shock-wave experiments on copper and on lead. The tangent to the melting curve that results when pressure is plotted against temperature, $\partial P/\partial T$, is given by the ratio of the change in entropy (ΔS) and the change in volume (ΔV) between solid and liquid, or $\partial P/\partial T = \Delta S/\Delta V$. Thermodynamics shows that the liquid, being the high-temperature phase, has the greater entropy. In most cases, because the liquid has the larger volume, both the change in entropy and the change in volume are positive. This is, however, not true for water, bismuth, and a few other substances.

In most cases, pressure raises the melting point, and, most frequently, the liquid is more compressible. Therefore, as pressure is increased, the change in volume decreases, and the rate of temperature increase of the melting curve is slowed. In a few cases, however, such as cesium and potassium nitrates, a maximum temperature is reached at which the liquid becomes denser than the solid. Thereupon, any further application of pressure decreases the melting point. If a transition in crystal structure takes place, thereby transforming one type of solid into another (*i.e.,* if there is a solid–solid transition in a solid), then a triple point occurs on the melting curve at which solid (1), solid (2), and liquid are all in equilibrium at one pressure and temperature. After continuing past this break in the melting curve, the process may be repeated.

Both cesium and potassium nitrates actually have two maximums in their melting curves before a third solid phase intervenes. To illustrate further the occasional complex behaviour of matter, the melting temperature of cesium decreases with pressure until a minimum is reached, after which the melting temperature increases. In this case, even though the solid is less dense than the liquid, it has a lower initial compressibility, hence the volume of solid and liquid approach each other and become equal at the minimum point.

EFFECTS ON POLYMORPHISM

Polymorphism is the phenomenon in which a substance can exist in more than one form. Diamond and graphite, which are radically different forms of carbon, and calcite and aragonite, two forms of calcium carbonate ($CaCO_3$), are examples of polymorphic substances. Although this aspect of solids had been known for over 200 years, it was not until Bridgman extended the range of pressure studies that it was realized how universal this behaviour of matter is. In polymorphic substances, a thermodynamic rule states that for two-phase equilibrium, at constant temperature, the high-pressure phase must be denser. In a crystalline solid, one obvious way of obtaining a denser phase without chemical change is to increase its coordination number (the number of neighbour atoms surrounding a given atom). The common form of iron (alpha iron, or α-iron), for example, is a body-centred cubic crystal at room temperature; and, hence, it is eight coordinated (each atom has eight atoms around it positioned as though at the eight corners of a cube). One might anticipate, therefore, that, at some unspecified elevated pressure, a closer packed form of iron might become stable. Such a form of iron, which is 12 coordinated, has been found at 130 kilobars in shock-wave studies. X-ray crystallography has shown it to be hexagonal close-packed, not cubic close-packed, as the gamma iron already known to occur at elevated temperatures. It is paramagnetic, unlike normal ferromagnetic α-iron.

Coordination number

The statement that an increase in a substance's coordination number results in a higher density, however, is true only if the atoms in the higher coordinated phase do not have drastically larger radii. For some elements—*e.g.,* titanium, zirconium, and ytterbium—the converse is true. In these cases, application of pressure to an initially close-packed 12-coordinated element changes it to a more open-body eight-coordinated cubic, the atomic radii of which are smaller. The tetrahedrally coordinated elements germanium and silicon are semiconductors (see above *Semiconductors and insulators*), the electrical conductivity of which becomes metallic when they adopt the structure of metallic white tin at high pressure, which is 20 percent denser. Other atomic arrangements are known for these elements. Indeed, white tin itself changes to a tetragonal form.

Shock-wave techniques have provided considerable insight into the volume behaviour of matter. They reveal how cations (positive ions), much more than anions (negative ions), influence the pressure at which the coordination structure changes. Sodium chloride (NaCl), for example, in its normal six-coordinated structure has a density of 2.165 grams per millilitre at one bar, 2.248 at 10 kilobars, 2.731 at 100 kilobars, and finally 3.349 at 300 kilobars, at which it transforms to the eight-coordinated modification. In potassium chloride (KCl) the same transition as that in sodium chloride occurs at 18 kilobars, in rubidium chloride (RbCl) at four kilobars, whereas cesium chloride (CsCl) is in this form at one bar. The effect of the chlorine anion on these transformations is trivial by comparison with the cation (*e.g.,* sodium, potassium, or other). Shock-wave techniques have also been used to study the volume behaviour of many elements at a pressure of two megabars. At this pressure the densities of most metals have been increased by 60–100 percent, temperatures of 2,000° C (3,600° F) are attained, and all thermodynamic properties of the materials may be obtained for use in calculating chemical equilibrium. By using the Rankine-

Hugoniot relations (see below *Principles of apparatus design*) the internal energy (E) and the enthalpy (H), with $H = E + PV$, are obtained directly from the shock-wave data. The entropy (S) may be obtained from an equation relating change in energy to pressure, change in volume, temperature, and change in entropy, as follows: $dE = -PdV + TdS$, but not directly; an approximation has to be made, as must also be done for temperature. From these expressions, Gibbs free energy is derivable. Further thermodynamic considerations permit reduction of the data to an isotherm (constant temperature), so that thermodynamic functions covering a large rectangle in P–T space become known. In liquid oxygen, for example, the upper corner in P–T space is 10,000° C (18,000° F) and 1.4 megabars.

EFFECTS ON MINERALS

Although the necessary apparatus had been in existence for some time before World War II, it was not until afterward that there was a major effort to solve high-pressure mineralogical problems in the laboratory. Because the entire range of geological conditions that exist in the Earth's crust (namely, pressures up to 10 kilobars and temperatures up to 1,500° C [2,700° F]) may be duplicated in the laboratory, it is possible to investigate natural reactions in which pressure is a factor. Moreover, by using such fluids as water and carbon dioxide as pressure mediums, physical properties as well as chemical reactions can be studied. Among the geological conditions that can be measured are the effect of pressure on the electrical conductivity of minerals and on various Curie temperatures (a critical temperature above which a ferromagnetic substance becomes paramagnetic). Much laboratory effort has also been expended on measuring the effect of pressure on the propagation of longitudinal and transverse elastic waves through the Earth. It has been found that such measurements, to be meaningful, must be made under pressure; this is because, at a pressure as low as one bar, the initial porosity and lack of compactness of rocks make it impossible to determine values of wave velocity between different samples of nominally the same rock type. In those cases in which the deformation (under confining pressure) of at least a few rock types has been studied in detail in the laboratory, pressure enhances rock flow and helps to explain the tortuous nature of certain folded rocks in nature.

Duplicating nature in geological research

With the development in the mid-1950s of the apparatus illustrated in Figure 57, conditions duplicating those in the Earth's upper mantle could be achieved. There then followed a major change in geophysical concepts, such as the previously unrealized importance of polymorphic changes induced by high pressure. An indication of what was to come was the discovery of a denser (2.91 grams per cubic centimetre versus 2.65 grams per cubic centimetre) form of quartz (coesite), first in the laboratory and subsequently at meteorite craters. Later, a still denser (4.29 grams per cubic centimetre) modification, now called stishovite, was found. Thus it was seen that a common constituent (quartz) of surficial rocks could exist in a form that was 62 percent denser at depths of about 400 kilometres (250 miles). Similar pressure studies have revealed modifications of olivines, pyroxenes, iron oxides, and feldspar in which density increases of 25–50 percent are not uncommon. As a result of such studies, it is safe to say that the mineral components of the Earth at a depth of 1,000 kilometres (620 miles)—corresponding to a pressure of 400 kilobars—will contain completely different crystal structures than their chemical counterparts found in the crust. There may be a few exceptions, however, such as magnesium oxide (MgO) and carbon in the form of diamond.

Laboratory studies also destroyed the belief that the rise time of a shock wave is too rapid to allow a polymorphic transition to occur. Many mineral transitions are found to take place—even diamonds have been made—when explosives are used. Stishovite has also been recovered from explosive studies, and its occurrence at meteorite craters is thought to have been caused by natural explosive events when the meteors struck the Earth.

EFFECTS ON MAGNETIC AND ELECTRICAL PROPERTIES

Solids are commonly classified as insulators, semiconductors, or metals, depending on the magnitude of their electrical resistivity and its temperature coefficient (see above). Although such insulators as magnesium oxide and certain alkali halides have been studied while under explosive shock, comparatively few studies have been made of the effect of pressure on the electrical resistance of insulators because of experimental difficulties. A common method of estimating temperatures in the Earth is to compare the Earth's known electrical conductivity, which varies as a function of depth, with the conductivity at various temperatures of those minerals thought to exist there. Although olivine, which has a wide distribution in the Earth, has been studied extensively in the laboratory, no definitive statement concerning the effect of pressure on electrical conductivity can be made. In general, however, one can say that, if the electrical conductivity in a substance is the result of ionic motion, pressure should inhibit the motion and thus increase the resistivity. Similarly, diffusion rates should decrease. If, on the other hand, the process is electronic—*i.e.*, the result of motion of an electron or a hole—pressure may either increase or decrease the electrical resistance.

Electrical effects

Semiconductors are much more sensitive to pressure than other metals, because pressure can have gross effects on carrier density or mobilities or both through an alteration of the conductive-band structure. Changes of several orders of magnitude in resistivity can occur; for instance, 30 kilobars at room temperature changes the resistivity of the semiconductor germanium by a factor of 100. Even more startling changes in resistivity happen when a phase change occurs, as, for example, when at about 120 kilobars germanium adopts the white-tin structure and becomes a metallic conductor. Even in metals, the high density of charge carriers, the free electrons, can react in a complicated manner with the nuclear potentials, causing great variability in behaviour when pressure is increased. Strontium, for example, has maximum resistivity at about 40 kilobars, whereas calcium has a maximum at over 300 kilobars. Barium's electrical resistance is at a minimum near 10 kilobars and at a maximum near 130 kilobars. Similarly, antimony has a maximum near 25 kilobars, whereas arsenic has a minimum near that pressure but a maximum near 60 kilobars. Phosphorus (the black form), which has the same number of electrons as arsenic in its outer atomic shell, adopts the arsenic structure at an elevated pressure and then is transformed to a simple cubic structure around 110 kilobars. Presumably, both latter forms of phosphorus are metals, although at a pressure of one bar phosphorus is a semiconductor.

Magnetic effects

Few studies have been made of the effect of pressure on the magnetic properties of matter. In the range below 30 kilobars, where hydrostatic mediums may be used, the effect of pressure is slight. In the range above 30 kilobars, where solid mediums must be used, experimental design and interpretation of data are difficult. Nevertheless, the Curie point of iron has been found to be unaffected by pressure up to the point at which paramagnetic γ-iron forms from ferromagnetic α-iron. In a few other elements, pressure causes moderate fluctuations in Curie points.

Applications of high pressure

The most important economic use of "high" pressure, namely in refining and plastic industries, will not be discussed here, because the pressures employed are lower than those with which this discussion is concerned. The primary industrial application of truly high pressure is in the manufacture of diamonds from other forms of carbon. In the belt apparatus (Figure 57A), carbon can be transformed to the diamond modification at about 55 kilobars and 1,500° C (2,700° F) in the presence of a transition metal (one with an incomplete inner shell) such as nickel. The small diamonds made this way are of industrial grade. Although the same technique can be used to manufacture gem-quality diamonds, the time involved makes this process uneconomical at present. A denser form of boron

nitride, with diamond-like hardness, has also been manufactured under pressure.

Another use of high pressure is to alter the mechanical properties of materials, such as increasing the ductility of a metal. In one such application, a new high-speed wire-drawing system has been developed in which external hydrostatic pressure is used to reduce the size of the wire in one die, thereby replacing a whole series of reduction dies that had formerly been needed. Attempts are also being made to compress hydrogen or its compounds in order to create a stable form of hydrogen metal in the solid crystalline state.

The comparative paucity of major applications of ultra-pressure technology is probably not because of any lack of phenomena but rather because of the relative newness of the field and the magnitude of the area to be surveyed. To map out the temperature properties of a hypothetical alloy at normal temperature and pressure, for example, might require 20 samples that differ in composition and, thus, 20 experiments. To make the same temperature study for the 20 samples at intervals between 10 kilobars and 200 kilobars, 400 experiments would be necessary.

Bridgman's works

Use of hydrostatic pressure. In 1909 Percy Williams Bridgman published a description of a leakproof packing that would extend high-pressure research from three kilobars to 30. Although there had been no problem previously in finding theoretical ways to obtain high pressures by hydrostatic means, no one had discovered how to keep a fluid at such pressures from seeping out of the system. To circumvent this problem, Bridgman used the principle of unsupported area in his packing, which is annular; *i.e.,* shaped like a ring. The pressure of fluid is against one side of the disk, which in turn presses on the annulus. Because the annulus has no material in its centre, its area is less than that of the disk. Hence the pressure per unit area on the annulus is greater than on the disk. Thus the fluid, which is confined behind the disk, cannot escape past it. With this packing arrangement, the sealing of gases and liquids was no longer a problem, but then the strength of containers and plungers became an important factor. The problem with plungers was solved by the use of cemented tungsten carbide, which rarely fails for pressures under 50 kilobars.

Containers. As for containers, such as heavy-walled steel cylinders that may not burst until 40 kilobars of internal pressure, they suffer plastic deformation starting at 15–20 kilobars. Therefore, they are inefficient for repeated use. Bridgman devised a scheme for overcoming this weakness, by fashioning a conical container. As pressure is increased in the interior of the container, the conical vessel is forced into a conical sleeve, because this creates a stress on the exterior of the interior vessel that is approximately equal to the pressure on the inside; thus, deformation is avoided. Bridgman modified this principle of external support by immersing an entire small pressure vessel in a fluid at 30 kilobars. This enabled him to perform experiments at pressures as high as 100 kilobars. More recently, by using the same modification principle, pressure generators have been developed, as illustrated in Figure 57. Figure 57A is a diagram of the belt, an apparatus developed in the United States by the General Electric Company for manufacturing the first man-made diamonds. Devices of this general type have been used to achieve pressures of about 200 kilobars and a temperature of about 3,000° C (5,400° F) (but not concurrently). Bridgman anvils, which are shown in Figure 57B, were developed to study the electrical resistance of materials up to 100 kilobars. Their chief use has been by geologists in studying mineral reactions in the laboratory; generally, a furnace is added externally to provide elevated temperatures. Figure 57C shows a modification of Bridgman anvils. By supporting the tapered faces, it has been possible to achieve the greatest static (nonexplosive) pressure to date, about 500 kilobars. These and similar designs are successful for two reasons. One is termed "massive support"; that is, a high pressure is exerted on only a small region of a massive object, as in all three designs of Figure 57. The second reason is that a comparatively soft gasket, such as that filling the space between piston and cylinder in Figure 57A, seals against high pressure by its own internal friction as well as by friction against the walls, but only as long as the piston–cylinder gap is not too great.

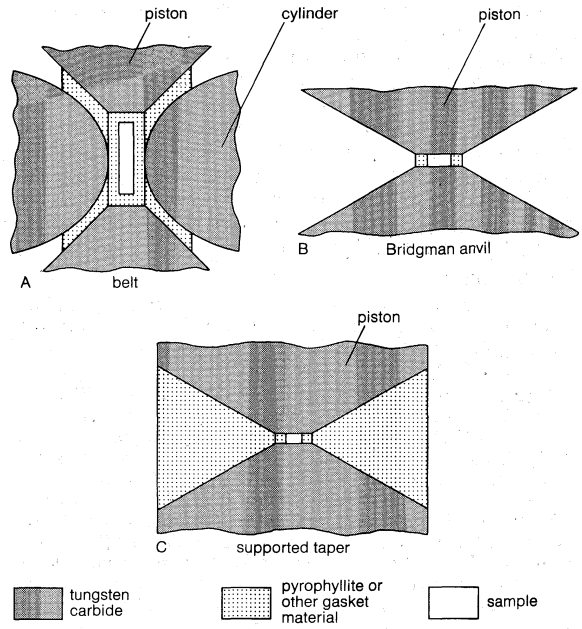


Figure 57: Schematic apparatus designs.
From H. Drickamer, "Scientific Research at Very High Pressures," *Physics Today* (November 1964)

Shock-wave technique. The greatest extension of pressure and temperature studies began after World War II with the development of controlled explosive events. Even higher pressures and temperatures are achieved with nuclear devices. Studies with explosives, termed shock-wave research, are excellent illustrations of how an experiment may be designed to yield valuable information that can be interpreted without mathematical complexity. The basic tool for such studies, displayed in Figure 58, is called a plane-wave generator.

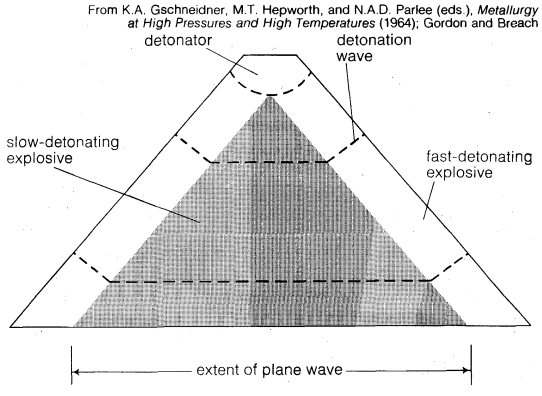


Figure 58: Plane-wave generator.

plane-wave generator. When detonated, the velocities of the two explosive charges in the device are such that, due to the conical angle, the velocity at the edge of the slower central explosive wave is exactly matched by the faster wave of the outer explosive. Thus the inner wave is kept exactly plane, and any waves it induces in another pad of explosive and the sample assembly are also plane, at least for a known distance.

The three governing equations for a shock wave under the above conditions are called the Rankine-Hugoniot relations. These equations, which are based on the conservation of mass, momentum, and energy, show the relationship between the energy, pressure, and specific volumes (reciprocal of the density) as well as the shock velocity and the partial velocity (the velocity at which the shock material is moving after the shock has passed). When the shock velocity and the partial velocity are measured, the

volume behaviour of a substance, which is a function of pressure, can be computed.

The Rankine-Hugoniot equations also apply to projectile impact when it is more convenient to impel a projectile at a high velocity and allow it to impact a sample than it is to use explosives. Whereas explosive research is limited to pressures of about two megabars, high-velocity studies with a gas gun may reach several megabars. Workers in the Soviet Union have reported experimentation at 34 megabars, but they have not specified the technique they employed to reach such pressures. (J.C.J./Ed.)

BIBLIOGRAPHY

General works: E.A. MOELWYN-HUGHES, *States of Matter* (1961); W.F. SHEEHAN, *Physical Chemistry,* 2nd ed. (1970); and G.W. CASTELLAN, *Physical Chemistry,* 2nd ed. (1971); J.H. MULVEY (ed.), *The Nature of Matter* (1981).

Phase changes and equilibria: J.W. GIBBS, "On the Equilibrium of Heterogeneous Substances," *Trans. Conn. Acad. Arts Sci.,* vol. 3 (1875–78), on the deduction of the phase rule; A. FINDLAY, A.N. CAMPBELL, and N.O. SMITH, *The Phase Rule and Its Applications,* 9th ed. (1951); R. HAASE and H. SCHONERT, "Solid-Liquid Equilibrium," in *The International Encyclopaedia of Physical Chemistry and Chemical Physics,* topic 13, vol. 1 (1969); A.W. FRANCIS, *Liquid–Liquid Equilibriums* (1963); J.E. RICCI, *The Phase Rule and Heterogeneous Equilibrium* (1951); L. BREWER, "Viewpoints of Stability of Metallic Structures," in P.S. RUDMAN, J. STRINGER and R.I. JAFFEE (eds.), *Phase Stability in Metals and Alloys* (1967).

(N.O.S./Ed.)

Solid state: LAWRENCE VAN VLACK, *Elements of Materials Science* (1964), an elementary textbook; CHARLES A. WERT and ROBB M. THOMSON, *The Physics of Solids,* 2nd ed. (1970), an intermediate level text; CHARLES KITTEL, *Introduction to Solid State Physics,* 4th ed. (1971), an advanced level text; and JOHN ZIMAN, *Electrons and Phonons* (1960). See also the series FREDERICK SEITZ and DAVID TURNBULL, *Solid State Physics,* 22 vol. (1956–71), written for the expert. GEORGE E. BACON, *The Architecture of Solids* (1981), an introduction to the physical properties, atomic arrangements, and types of bonding of solids.

The classic reference work on bonding is LINUS PAULING, *Nature of the Chemical Bond and the Structure of Molecules and Crystals,* 3rd ed. (1960). A readable treatment of crystal structure is C.S. BARRETT and T.B. MASSALSKI, *Structure of Metals,* 3rd ed. (1966). See also J.F. NYE, *Physical Properties of Crystals* (1957); M.J. BUERGER, *Introduction to Crystal Geometry* (1971); F.C. PHILLIPS, *An Introduction to Crystallography,* 4th ed. (1972); W.L. BRAGG and G.F. CLARINGBULL, *Crystal Structures of Minerals* (1965); R.A. LAUDISE, *The Growth of Single Crystals* (1970); A. VAN HOOK, *Crystallization* (1961); R.H. DOREMUS, B.W. ROBERTS, and D. TURNBULL (eds.), *Growth and Perfection of Crystals* (1958); H.S. PEISER (ed.), *Crystal Growth* (1967); and MICHAEL O'KEEFFE and A. NAVROTSKY (eds.), *Structure and Bonding in Crystals,* 2 vol. (1981–).

Glass and certain other amorphous solids are treated in J.R. HUTCHINS and R.V. HARRINGTON, "Glass" in the *Encyclopaedia of Chemical Technology,* 2nd ed., vol. 10, pp. 533–604 (1966), a review article with an extensive bibliography; G.O. JONES, *Glass,* 2nd ed., rev. by S. Parke (1971), a clear, short introduction to the physics of the glassy state; G.W. MOREY, *The Properties of Glass,* 2nd ed. (1954, reissued 1970), a discussion of the measurements of the physical properties of glasses and of their dependence on chemical composition; and R.H. DOREMUS, *Glass Science* (1973), similar to Morey's classic work, but less comprehensive and more recent.

Coverage of conductors, semiconductors, and insulators is provided in W.A. HARRISON, *Solid State Theory,* rev. ed. (1980); S.M. SZE, *Physics of Semiconductor Devices* (1969); and W. SHOCKLEY, *Electrons and Holes in Semiconductors* (1950).

References on the mechanical properties of solids include A.H. COTTRELL, *The Mechanical Properties of Matter* (1964, reissued 1981); and GEORGE E. DIETER, JR., *Mechanical Metallurgy* (1961). Interaction of radiation with matter is covered in EARLE B. BROWN, *Modern Optics* (1965); and GRANT R. FOWLES, *Introduction to Modern Optics* (1968); RALPH S. BECKER, *Theory and Interpretation of Fluorescence and Phosphorescence* (1969); and P.R. THORNTON, *The Physics of Electroluminescent Devices* (1967). General books on magnetism and magnetic materials are: JAN SMIT (ed.), *Magnetic Properties of Materials* (1971); JOHN E. THOMPSON, *The Magnetic Properties of Materials* (1968); and ROY CAREY and EVAN D. ISAAC, *Magnetic Domains and Techniques for Their Observation* (1966). The classic book on materials of ferromagnetism is RICHARD M. BOZORTH, *Ferromagnetism* (1956). H.G. HOPKINS and M.J. SEWELL (eds.), *Mechanics of Solids* (1982), 18 papers devoted mainly to mathematical treatments of plasticity.

(C.A.W./H.K.B./R.A.L./J.B.Co./Ed.)

Liquid state: P.A. EGELSTAFF, *An Introduction to the Liquid State* (1967), a work that serves both as an introduction to liquid structure and as a more advanced account of the scattering of radiation; D.S. EISENBERG and W. KAUZMANN, *The Structure and Properties of Water* (1969), covers the internal structure of the water molecule as well as its behaviour in gaseous, liquid, and solid states; G.W. GRAY, *Molecular Structure and the Properties of Liquid Crystals* (1962); R.M. MAZO, *Statistical Mechanical Theories of Transport Processes* (1967), a short book that deals primarily with liquids; J.S. ROWLINSON, *Liquids and Liquid Mixtures,* 3rd ed. (1982), covers the classical and statistical thermodynamics of liquids at equilibrium, with some emphasis on the properties at high pressures; H.N.V. TEMPERLEY, J.S. ROWLINSON, and G.S. RUSHBROOKE (eds.), *Physics of Simple Liquids* (1968), an advanced treatise that covers equilibrium and nonequilibrium statistical thermodynamics and computer calculations for molecularly simple liquids. See also D.P. WOODRUFF, *The Solid-Liquid Interface* (1973, reissued 1980).

An introductory but quantitative discussion of liquid solutions is given by A.G. WILLIAMSON in *An Introduction to Non-Electrolyte Solutions* (1967). A comprehensive view is presented by J.H. HILDEBRAND and R.L. SCOTT in *Solubility of Non-Electrolytes,* 3rd ed. (1950); more recent work is reviewed in J.H. HILDEBRAND, J.M. PRAUSNITZ, and R.L. SCOTT, *Regular and Related Solutions* (1970).

(Jo.S.R./J.M.P./Ed.)

Gaseous state: The full theory of the equilibrium properties of gases is given in J.O. HIRSCHFELDER, C.F. CURTISS, and R.B. BIRD, *The Molecular Theory of Gases and Liquids,* corrected printing (1965); the full theory of the transport processes of gases in S. CHAPMAN and T.G. COWLING, *The Mathematical Theory of Non-Uniform Gases,* 3rd ed. (1970); and the full theory of thermal-relaxation processes in gases in J.F. CLARKE and M. MCCHESNEY, *The Dynamics of Real Gases* (1964). A good but rather dated compilation of experimental values of equilibrium and transport properties of gases and liquids is R.C. REID and T.K. SHERWOOD, *The Properties of Gases and Liquids* (1958).

(M.McC.)

Plasma state: LYMAN SPITZER, *The Physics of Fully Ionized Gases,* 2nd rev. ed. (1962); and T.G. COWLING, *Magnetohydrodynamics* (1957) are basic introductory texts. More advanced works are J.L. DELCROIX, *Plasma Physics,* vol. 1 and 2 (1965–68); and WULF B. KUNKEL (ed.), *Plasma Physics in Theory and Application* (1966). Two excellent texts on experimental techniques are W. LOCHTE-HOLTGREVEN (ed.), *Plasma Diagnostics* (1968); and M.A. HEALD and C.B. WHARTON, *Plasma Diagnostics with Microwaves* (1965). Two classic works are HANNES ALFVEN and CARL-GUNNE FALTHAMMAR, *Cosmical Electrodynamics,* 2nd ed. (1963); and SYDNEY CHAPMAN and T.G. COWLING, *The Mathematical Theory of Non-Uniform Gases,* 3rd ed. (1970). An advanced text, with many references, is V.L. GINZBURG, *The Propagation of Electromagnetic Waves in Plasmas* (trans. from the Russian by J.B. SYKES and R.J. TAYLER, 1964).

(B.S.L.)

Colloids: DUNCAN J. SHAW, *Introduction to Colloid and Surface Chemistry,* 3rd ed. (1980), fairly elementary, not containing much on hydrophilic colloids; KAROL J. MYSELS, *Introduction to Colloid Chemistry* (1959), more advanced than Shaw's book but still elementary; HUGO R. KRUYT (ed.), *Colloid Science,* vol. 1, *Irreversible Systems* and vol. 2, *Reversible Systems,* Eng. trans. by L.C. JACKSON (1952, 1949; reprinted 1969), advanced texts, slightly outdated, especially vol. 2, but more modern books at this level are not available; ALBERT E. ALEXANDER and PALEY JOHNSON, *Colloid Science,* 2 vol. (1949; reissued in one vol., 1952), an advanced text, also somewhat outdated but containing more on proteins than Kruyt's work; *High Polymers* (1940–), a continuing classic series on polymers; CHARLES TANFORD, *Physical Chemistry of Macromolecules* (1961), contains much on polyelectrolytes and biopolymers; J.T.G. OVERBEEK, *Colloid and Surface Chemistry* (1971–72), a videotaped self-study course in 4 vol. (55 lessons), with printed study guides, advanced level, and many demonstrations.

(J.T.G.O.)

Low-temperature phenomena: F.E. SIMON *et al., Low Temperature Physics: Four Lectures* (1952, reprinted 1961); and K. MENDELSSOHN, *The Quest for Absolute Zero: The Meaning of Low Temperature Physics,* 2nd ed. (1977), are excellent nontechnical summaries. FRITZ LONDON, *Superfluids,* vol. 1, *Macroscopic Theory of Superconductivity,* 2nd ed. (1961), vol. 2, *Macroscopic Theory of Superfluid Helium* (1954, reprinted 1964), the classic presentation of theory; GUY K. WHITE, *Experimental Techniques in Low-Temperature Physics,* 3rd ed. (1979), a detailed discussion; C.J. GORTER (ed.), *Progress in Low Temperature Physics* (1955–), theoretical and experimental research reports and review articles, all with extensive bibliographies; K. MENDELSSOHN (ed.), *Progress in Cryogenics,* 4 vol. (1959–64), which complements the Gorter series, with more emphasis on applied problems and developments; H.M.

ROSENBERG, *Low Temperature Solid State Physics: Some Selected Topics* (1963, reprinted 1965), fairly elementary; DAVID R. TILLEY and JOHN TILLEY, *Superfluidity and Superconductivity,* 2nd ed. (1986); CHARLES G. KUPER, *An Introduction to the Theory of Superconductivity* (1968); WILLIAM E. KELLER, *Helium-3 and Helium-4* (1969), advanced-level summaries of theory; A.C. ROSE-INNES and E.H. RHODERICK, *Introduction to Superconductivity,* 2nd ed. (1978), basic, with an emphasis on application; MICHAEL TINKHAM, *Introduction to Superconductivity* (1975, reprinted 1980), an intermediate introduction; P.G. DE GENNES, *Superconductivity of Metals and Alloys* (1966), a presentation of the foundations and applications of the theory; R.D. PARKS (ed.), *Superconductivity,* 2 vol. (1969); W.D. GREGORY, W.N. MATTHEWS, JR., and E.A. EDELSACK (eds.), *The Science and Technology of Superconductivity,* 2 vol. (1973); MARSHALL SITTIG, *Cryogenics: Research and Applications* (1963); and two articles from *American Journal of Physics*: D.M. GINSBERG, "Resource Letter Scy-1 on Superconductivity," 32:85–89 (1964), and ". . . Scy-2. . .," 38:949–955 (1970), introductory reviews with descriptive bibliographies. High-temperature superconductors are reported on in M.K. WU *et al.,* "Superconductivity at 93 K in a New Mixed-Phase Y-Ba-Cu-O Compound System at Ambient Pressure," *Physical Review Letters,* 58(9):908–910 (March 2, 1987); ANIL KHURANA, "Superconductivity Seen Above the Boiling Point of Nitrogen," *Physics Today,* 40(4):17–23 (April 1987); K. ALEX MÜLLER and J. GEORG BEDNORZ, "The Discovery of a Class of High-Temperature Superconductors," *Science,* 237(4819):1133–39 (Sept. 4, 1987); and STUART A. WOLF and VLADIMIR Z. KRESIN, *Novel Superconductivity* (1987). See also articles in *Journal of Low Temperature Physics* (semiannual).

(C.E.He./D.M.Gi.)

High-pressure phenomena: P.W. BRIDGMAN, *The Physics of High Pressure* (1949), a classic; R.S. BRADLEY (ed.), *High Pressure Physics and Chemistry,* vol. 1– (1963–); *Advances in High Pressure Research,* vol. 1– (1966–); N.B. BRANDT and N.I. GINZBURG, "Superconductivity at High Pressure," *Scient. Am.,* 224:83–88 (1971); F.P. BUNDY, W.R. HIBBARD, and H.M. STRONG (eds.), *Progress in Very High Pressure Research* (1961); W. PAUL and D.M. WARSCHAUER (eds.), *Solids Under Pressure* (1963); K.A. GSCHNEIDNER, M.T. HEPWORTH, and N.A.D. PARLEE (eds.), *Metallurgy at High Pressures and High Temperatures* (1964); C.T. TOMIZUKA and R.M. EMRICK (eds.), *Physics of Solids at High Pressures* (1965); R.H. WENTORF (ed.), *Modern Very High Pressure Techniques* (1962); and A. ZEITLIN, "High Pressure Technology," *Scient. Am.,* 212:38–46 (1965).

(J.C.J.)

Maxwell

James Clerk Maxwell is regarded by most modern physicists as the scientist of the 19th century who had the greatest influence on 20th-century physics; he is ranked with Sir Isaac Newton and Albert Einstein for the fundamental nature of his contributions. In 1931, at the 100th anniversary of Maxwell's birth, Einstein described the change in the conception of reality in physics that resulted from Maxwell's work as "the most profound and the most fruitful that physics has experienced since the time of Newton." The concept of electromagnetic radiation originated with Maxwell, and his field equations, based on Michael Faraday's observations of the electric and magnetic lines of force, paved the way for Einstein's special theory of relativity, which established the equivalence of mass and energy. Maxwell's ideas also ushered in the other major innovation of 20th-century physics, the quantum theory. His description of electromagnetic radiation led to the development (according to classical theory) of the ultimately unsatisfactory law of heat radiation, which prompted Max Planck's formulation of the quantum hypothesis—*i.e.,* the theory that radiant-heat energy is emitted only in finite amounts, or quanta. The interaction between electromagnetic radiation and matter, integral to Planck's hypothesis, in turn has played a central role in the development of the theory of the structure of atoms and molecules.

By courtesy of King's College, London



Maxwell, engraving by G.J. Stodart.

Early life. Maxwell came from a comfortable middle-class background. The original family name was Clerk, the additional surname being added by his father after he had inherited the Middlebie estate from Maxwell ancestors. James, an only child, was born on June 13, 1831, in Edinburgh, where his father was a lawyer; his parents had married late in life, and his mother was 40 years old at his birth. Shortly afterward the family moved to Glenlair, the country house on the Middlebie estate.

Education and early contributions

His mother died in 1839 from abdominal cancer, the very disease to which Maxwell was to succumb at exactly the same age. A dull and uninspired tutor was engaged who claimed that James was slow at learning, though in fact he displayed a lively curiosity at an early age and had a phenomenal memory. Fortunately he was rescued by his aunt Jane Cay and from 1841 was sent to school at the Edinburgh Academy. Among the other pupils were his biographer Lewis Campbell and his friend Peter Guthrie Tait.

Maxwell's interests ranged far beyond the school syllabus, and he did not pay particular attention to examination performance. His first scientific paper, published when he was only 14 years old, described a generalized series of oval curves that could be traced with pins and thread by analogy with an ellipse. This fascination with geometry and with mechanical models continued throughout his career and was of great help in his subsequent research.

At the age of 16 he entered the University of Edinburgh, where he read voraciously on all subjects and published two more scientific papers. In 1850 he went to the University of Cambridge, where his exceptional powers began to be recognized. His mathematics teacher, William Hopkins, was a well-known "wrangler maker" (a wrangler is one who takes first class honours in the mathematics examinations at Cambridge) whose students included Tait, George Gabriel (later Sir George) Stokes, William Thomson (later Lord Kelvin), Arthur Cayley, and Edward John Routh. Of Maxwell, Hopkins is reported to have said that he was the most extraordinary man he had met with in the whole course of his experience, that it seemed impossible for him to think wrongly on any physical subject, but that in analysis he was far more deficient. (Other contemporaries also testified to Maxwell's preference for geometrical over analytical methods.) This shrewd assessment was later borne out by several important formulas advanced by Maxwell that obtained correct results from faulty mathematical arguments.

In 1854 Maxwell was second wrangler and first Smith's prizeman (the Smith's prize is a prestigious competitive

award for an essay that incorporates original research). He was elected to a fellowship at Trinity, but because his father's health was deteriorating, he wished to return to

Career

Scotland. In 1856 he was appointed to the professorship of natural philosophy at Marischal College, Aberdeen, but before the appointment was announced his father died. This was a great personal loss, for Maxwell had had a close relationship with his father. In June 1858 Maxwell married Katherine Mary Dewar, daughter of the principal of Marischal College. The union was childless and was described by his biographer as a "married life . . . of unexampled devotion."

In 1860 a merger took place between the two constituents of the University of Aberdeen, King's College and Marischal College, and Maxwell was declared redundant. He applied for a vacancy at the University of Edinburgh, but he was turned down in favour of his school friend Tait. He then was appointed to the professorship of natural philosophy at King's College, London.

The next five years were undoubtedly the most fruitful of his career. During this period his two classic papers on the electromagnetic field were published, and his demonstration of colour photography took place. He was elected to the Royal Society in 1861. His theoretical and experimental work on the viscosity of gases also was undertaken during these years and culminated in a lecture to the Royal Society in 1866. He supervised the experimental determination of electrical units for the British Association for the Advancement of Science, and this work in measurement and standardization led to the establishment of the National Physical Laboratory. He also measured the ratio of electromagnetic and electrostatic units of electricity and confirmed that it was in satisfactory agreement with the velocity of light as predicted by his theory.

Later life. In 1865 he resigned his professorship at King's College and retired to the family estate in Glenlair. He continued to visit London every spring and served as external examiner for the Mathematical Tripos (exams) at Cambridge. In the spring and early summer of 1867 he toured Italy. But most of his energy during this period was devoted to writing his famous treatise on electricity and magnetism.

It was Maxwell's research on electromagnetism that established him among the great scientists of history. In the preface to his *Treatise on Electricity and Magnetism* (1873), the best exposition of his theory, Maxwell stated

Research on electromagnetism

that his major task was to convert Faraday's physical ideas into mathematical form. In attempting to illustrate Faraday's law of induction (that a changing magnetic field gives rise to an induced electromagnetic field), Maxwell constructed a mechanical model. He found that the model gave rise to a corresponding "displacement current" in the dielectric medium, which could then be the seat of transverse waves. On calculating the velocity of these waves, he found that they were very close to the velocity of light. Maxwell concluded that he could "scarcely avoid the inference that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena."

Maxwell's theory suggested that electromagnetic waves could be generated in a laboratory, a possibility first demonstrated by Heinrich Hertz in 1887, eight years after Maxwell's death. The resulting radio industry with its many applications thus has its origin in Maxwell's publications.

Maxwell's other contributions to science

In addition to his electromagnetic theory, Maxwell made major contributions to other areas of physics. While still in his 20s, Maxwell demonstrated his mastery of classical physics by writing a prizewinning essay on Saturn's rings, in which he concluded that the rings must consist of masses of matter not mutually coherent—a conclusion that was corroborated more than 100 years later by the first Voyager space probe to reach Saturn.

The Maxwell relations of equality between different partial derivatives of thermodynamic functions are included in every standard textbook on thermodynamics (*q.v.*). Though Maxwell did not originate the modern kinetic theory of gases, he was the first to apply the methods of probability and statistics in describing the properties of an assembly of molecules. Thus he was able to demonstrate that the velocities of molecules in a gas, previously assumed to be equal, must follow a statistical distribution (known subsequently as the Maxwell–Boltzmann distribution law). In later papers Maxwell investigated the transport properties of gases—*i.e.*, the effect of changes in temperature and pressure on viscosity, thermal conductivity, and diffusion.

Maxwell was far from being an abstruse theoretician. He was skillful in the design of experimental apparatus, as was shown early in his career during his investigations of colour vision. He devised a colour top with adjustable sectors of tinted paper to test the three-colour hypothesis of Thomas Young and later invented a colour box that made it possible to conduct experiments with spectral colours rather than pigments. His investigations of the colour theory led him to conclude that a colour photography could be produced by photographing through filters of the three primary colours and then recombining the images. He demonstrated his supposition in a lecture to the Royal Institution of Great Britain in 1861 by projecting through filters a colour photograph of a tartan ribbon that had been taken by this method.

In addition to these well-known contributions, a number of ideas that Maxwell put forward quite casually have since led to developments of great significance. The hypothetical intelligent being known as Maxwell's demon was a factor in the development of information theory. Maxwell's analytic treatment of speed governors is generally regarded as the founding paper on cybernetics, and his "equal areas" construction provided an essential constituent of the theory of fluids developed by Johannes Diederik van der Waals. His work in geometrical optics led to the discovery of the fish-eye lens. From the start of his career to its finish his papers are filled with novelty and interest. He also was a contributor to the ninth edition of *Encyclopædia Britannica.*

Cavendish Laboratory

In 1871 Maxwell was elected to the new Cavendish professorship at Cambridge. He set about designing the Cavendish Laboratory and supervised its construction. Maxwell had few students, but they were of the highest calibre and included William D. Niven, Ambrose (later Sir Ambrose) Fleming, Richard Tetley Glazebrook, John Henry Poynting, and Arthur Schuster.

During the Easter term of 1879 Maxwell took ill on several occasions; he returned to Glenlair in June but his condition did not improve. He died after a short illness on November 5, 1879. Maxwell received no public honours and was buried quietly in a small churchyard in the village of Parton, Scotland.

BIBLIOGRAPHY. A standard biography is LEWIS CAMPBELL and WILLIAM GARNETT, *The Life of James Clerk Maxwell* (1882, reprinted 1969). Modern biographies include C.W. FRANCIS EVERITT, *James Clerk Maxwell: Physicist and Natural Philosopher* (1975), with a useful bibliography; IVAN TOLSTOY, *James Clerk Maxwell* (1981), for the general reader; and MARTIN GOLDMAN, *The Demon in the Aether* (1983). Commemorative publications include J.J. THOMSON (ed.), *James Clerk Maxwell: A Commemorative Volume, 1831–1931* (1931), containing lectures given at Cambridge University; CYRIL DOMB (ed.), *Clerk Maxwell and Modern Science* (1963), lectures concerning his electromagnetic theory; and CYRIL DOMB, "James Clerk-Maxwell: 100 Years Later," *Nature,* 282:235–239 (November 15, 1979). For more details of his contributions to electromagnetism, see R.A.R. TRICKER, *The Contributions of Faraday and Maxwell to Electrical Science* (1966); SIR EDMUND T. WHITTAKER, *A History of the Theories of Aether and Electricity,* rev. and enlarged ed., 2 vol. (1951–53, reprinted 1973). For Maxwell's original papers see WILLIAM D. NIVEN (ed.), *The Scientific Papers of James Clerk Maxwell,* 2 vol. (1890, reissued in 1 vol., 1965). See also Maxwell's own works: *Theory of Heat,* 3rd ed. (1872, reprinted 1970), and *Treatise on Electricity and Magnetism,* 2 vol. (1873; 3rd ed., 1892, reissued 1954).

(Cy.Do.)

Principles, Methods, and Instruments of Measurement and Observation

Measurement is the process used to answer the questions: How many? and How much? In many cases, it is integral to the observation of natural phenomena, particularly those objects and occurrences of scientific interest. Measurements, broadly defined, can be made by the unaided human senses and brain—for example, in estimating distances, dimensions, temperatures, and weights; identifying and matching colours; and estimating chemical composition by smell and taste. In general, however, man's capabilities need to be both extended and refined by instruments.

The modern level of technology would not have been attained without the existence of its instrumentation—namely, the development and use of precise measuring and observational devices. Every advance in technological capabilities from the earliest times was necessarily preceded by the development of instrumentation that made it possible to detect, measure, and gain access to the unknown.

The principles involved in the use of the human senses are employed in the instruments and mechanisms utilized throughout the technological environment. In searching for the instrumental tools to understand the world, investigators have first turned to those that are derived from these same principles, in the process effectively enlarging the natural senses. Instruments enabling pilots and navigators to see through fog and at great distances respond to the invisible radiation of radar and radio waves. Underwater obstacles and objects can be "seen" in terms of the echoes they produce from ultrasonic waves that are received by sonar instruments. And nowhere has the incredible extension of human capabilities by instrumentation been so dramatically revealed as in the exploration of space; precise measurement of the distance between the Earth and Moon is an outstanding example.

Instruments range from simple tools to large and complex systems, from the metre or yardstick to the computerized apparatus used in studying the atom and its constituent particles. To treat so vast a topic in detail is impossible in a brief space. This article thus describes only a few of the thousands of measuring and observational devices in use, focusing on those that have found universal application. For additional information on the instrumentation and methodology of the physical sciences, see ANALYSIS AND MEASUREMENT, PHYSICAL AND CHEMICAL. For coverage of other related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 723, 735, and 738, and the *Index*. (Ed.)

The article is divided into the following sections:

Measurement process and systems 727
- Basic aspects of the measurement process 727
- Functional elements of measuring systems 728
  - Principal elements
  - Performance characteristics
  - Developing instruments of measurement

Major forms of instrumentation 733
- Historical survey 733
  - Early history
  - The 19th and 20th centuries
- Thermometers 734
  - General considerations
  - Development of the thermometer
  - Methods of temperature measurement
- Microscopes 741
  - Historical development
  - The simple microscope
  - The compound microscope
  - Specialized types of microscopes
  - The electron microscope
  - Preparation of specimens for observation
    - Optical microscopy
    - Electron microscopy
- Telescopes 752
  - General considerations
  - Optical telescopes
  - Radio telescopes
  - Other types of telescopes
  - The development of the telescope and auxiliary instrumentation
- Sonar 769
  - Development of sonar
  - Fundamental principles of operation
  - Inherent limitations of sonar
  - Technical advances
- Instruments for monitoring industrial processes 772
  - Monitoring chemical properties
  - Monitoring physical properties
  - Measuring environmental properties
  - Production-line monitoring
- Other significant instruments 777
  - Biomedical instrumentation
  - Devices for atmospheric measurement
  - Pollution detection and monitoring devices
  - Laboratory instrumentation
  - Surveying instruments
  - Instruments for space exploration
  - Research in instrumentation

Systems of weights and measures 786
- Early units and standards 786
  - Ancient Mediterranean systems
  - Ancient Chinese weights and measures
  - Medieval weights and measures
- The English and U.S. customary weights and measures systems 787
  - English units and standards
  - The U.S. customary system
- The metric system of weights and measures 788
  - Development and establishment of the metric system
  - The International System of Units

Bibliography 790

MEASUREMENT PROCESS AND SYSTEMS

Basic aspects of the measurement process

Measurement begins with a definition of a quantity, condition, property, or other characteristic that is to be determined. In what follows, the quantity being measured will be referred to as the measurand to distinguish it from other quantities and variables included in the context. These definitions of measurables may be conceptual or operational. If conceptual, they must be converted to operational definitions in preparation for measurement. That is to say, they must be defined in terms of a sequence of steps or operations that describe a procedure for accomplishing the measurement. The measuring instrument is an embodiment of these operational steps. If the measurand is not fully defined, the measurement process becomes, itself, an essential part of the definition. An example is the measurement of hardness. Several kinds of hardness scales—*e.g.*, Brinell hardness, Knoop hardness, Rockwell

hardness—are each defined by the particular method or device used to measure them. Another example is octane number of gasoline, which is determined by evaluating the fuel when used to operate an internal-combustion engine of specified design under defined conditions. In general, the translation of a conceptual definition into an operational definition is not perfect, and the measurand defined by the measurement process consequently differs in varying degree from the intended or ideal.

Comparison with reference quantity

The information sought by the measurement process is always a comparison of the measurand with a reference quantity *of the same kind.* If a scaled series of reference quantities of all possible sizes is available, the comparison of measurand to the reference reduces to a demonstration that the measurand is equal in magnitude to a particular member of the set of reference values. Alternatively, the measurand may itself be subdivided into equal parts of uniform size shown to be equal to a unit reference magnitude, and the number of such parts is counted. Examples of scaled reference quantities are the subdivision of the metre or yard into centimetres or inches and its multiplication to kilometres or miles.

Frequently the measurand is not accessible or not suitable for direct comparison with a reference, as, for example, the radiation intensity inside a nuclear reactor. The measurand may then be transduced (converted) into an analogous measurement signal, a physical quantity of a more suitable kind that is related to the measurand in such a way that a given value of the signal represents a definite measurand magnitude. The reference quantity must then be similarly transduced to provide a reference signal of the same kind so that the measurand signal and the reference signal may be compared, after suitable multiplication or subdivision. Equality is established by opposing the measurand and reference, or the analogue signals that represent them, so that a transfer of energy occurs in one direction or the other depending on the relative magnitudes of the two. The approach to equality is then evidenced by reduction of the energy transfer to a minimum.

Because energy is required by the comparison process, there is always interaction between the observed measurand and the observing instrument. The value of the measurand observed is, therefore, not the same as that of the undisturbed measurand. If the disturbance is large enough to be significant, the exact value of the measurand must be inferred from knowledge of the disturbing process. The time required for accumulation and transfer of energy limits the capability to measure rapidly changing or dynamic phenomena.

Instruments as systems of joining of functional elements

Most measuring instruments are more or less complex systems in which a number of functional elements are joined to accomplish the desired measurement and to express the results in desired form. In some cases the functional elements are discrete units interconnected by the user to form the complete measuring instrument. In other cases the elements are so integrated that the individual functions are not readily discernible. Thus, light may be measured by connecting together a phototube, a power supply, an amplifier, and a voltmeter, or all of these may be integrated into a single instrument that is called a photometer.

Some of the measurement functions performed by the functional elements are basic to the measurement process. Thus, in any measuring instrument an element must be present that discriminates the measurand from all other conditions and senses its magnitude. In any measurement system this magnitude must be compared with the reference magnitude and the results of that comparison must be communicated to an observer or to a device that performs a control function. Other functions make it possible to present the results in the desired form at a desired time and place. Examples of the latter are signal generation, excitation, modulation, demodulation, magnification, attenuation, amplification, quantizing, counting, coding, digitizing, timing, transmission, telemetering, computation, display, and recording.

Although a specific sequence of functions exists in a given system, there is no standard order in which the functions are performed. Some may even be performed more than once. Many of the functional elements are themselves measuring instruments because they accept a signal as input and produce an output signal related quantitatively to the magnitude of the input.

Functional elements of measuring systems

PRINCIPAL ELEMENTS

Functional elements may be treated from the point of view of an ideally complete instrument system in which all measurement functions are performed without human involvement except for noting the final results. In practice, measurement systems are often only partly instrumented, some of the measurement functions being performed by operators or observers who are, therefore, participants in the measurement system. Thus, a spectrometer for measurement of chemical composition may produce a record of an infrared spectrum that must be interpreted by the observer to produce the desired composition information.

Measurands are often not measured directly but are instead inferred by combining the results of a number of separate measurements. Analogues of most instrument functions can be found in biological systems; *e.g.,* the temperature sensors of most living organisms. Conversely, most measurement functions that are performed by humans can also be accomplished by devices, although the performance of living systems has in some instances, such as in the sense of smell, not yet been equalled in nonliving systems.

Active and passive measurands

Measurement information is transferred throughout the instrument system by physical signals. If the measurand is itself energetic or active—for example, motion, temperature, pressure, or flow—it may provide the power required for the measurement signal. Passive measurands, such as location of an object in space, electrical resistance, and properties of materials, are not sources of signal power. To measure such phenomena, a measurement signal must be excited by causing the measurand to interact with an energetic probe or with a carrier signal.

Probes or sources of excitation signals take a variety of forms. Examples are light sources of various kinds, X-ray tubes, cathode-ray guns, particle accelerators, nuclear reactors, radioactive materials, sources of voltage and current, waveform generators, magnets, mechanical-force generators, vibration exciters, acoustic generators, and heaters.

Interaction of the carrier or probe with the measurand generates a measurement signal that carries information about the magnitude of the measurand. This information may be carried in the form of modulation of the exciting signal or carrier, a modification of a quantitative property of the carrier. If the carrier is an alernating voltage or current, as is quite frequently the case, then amplitude, frequency, and phase-angle modulation are commonly used.

If pulsed electrical signals are used as the carrier, then the height, frequency, position, duration, interval, or time of pulses may be used to carry measurement information. The properties of the carrier or probe that affect the recovery (demodulation) of the information it carries must be accurately controlled at the source.

A reference signal to be compared with the measurand is derived, ultimately, by a process of calibration accomplished by applying the instrument to measurands of known magnitude. The reference signal is often derived indirectly through a sequence of intermediate reference quantities.

Calibration may be effected automatically, but more often an operator is involved in the process. Examples of highly stable intermediate references are the cadmium standard cell and the Zener diode for voltage reference, the quartz-crystal controlled oscillator for frequency and time reference, and radiation emitted by elements such as sodium, cadmium, mercury, and krypton for length reference.

To effect comparison of measurand and reference, these, or the analogue signals that represent them, must be brought to equality, for example, by subdividing or multiplying the reference to provide a series of references of suitable resolution over the ranges of interest. Detection of equality, which closes the basic measurement process,

is frequently performed by the observer—*e.g.*, by noting the coincidence of a pointer and a scale graduation. It can also be performed by an instrument element.

An alternative comparison technique is quantization followed by counting. A signal is quantized by breaking it into parts, or quanta, of equal and known size. Counting the quanta then completes the measurement. Many measurands of interest occur in nature in quantized form, notably chemical species such as atoms and molecules, so that measurement of these can sometimes be accomplished by counting alone.

Amplification. The energy impressed upon or withdrawn from the measurand and from the reference quantity to provide power for the measurement signals must be kept small enough to avoid changing their magnitudes significantly. The permissible power level is frequently too small to permit performance of desired measurement functions, especially when the requirement for high accuracy is coupled with the need for great sensitivity and high speed. Amplification is the instrument function by which a signal carrying measurement information is increased in power without degrading its information content.

Electrical amplification is widely used in instrument systems. Pneumatic amplification, in which the measurement information is represented by pressure or flow of a gas, and hydraulic amplification, in which liquid pressure or flow is the information carrier, are also frequently used. Pneumatic amplification is employed in the technology of fluidics, whereby pneumatic signals are manipulated rapidly and transmitted at speed of sound by means of devices of small dimensions that involve no moving parts other than the gaseous fluid itself. Hydraulic amplification finds application especially in cases in which large amounts of power are to be transmitted in a relatively small space, as for controlling the aerodynamic surfaces of aircraft, for controlling the motion of heavy machine elements, and for manipulating fluids (as in process control). Hydraulic and pneumatic amplification are also used to avoid certain hazards, such as those of electrical currents in explosive atmospheres, and, in nuclear radiation environments, to avoid the effects of radiation damage to which electronic components are sometimes susceptible.

Use of pneumatic amplification

Electrical amplification is important for measurement applications because of the extremely rapid response that can be achieved; the speed with which electrical signals can be transmitted; the relative simplicity of the physical laws involved; the relative freedom of electrical parameters, such as resistance and capacitance, from environmental influences; the availability of electrical conductors that can transmit large amounts of power with little loss within the confines of small dimensions; the ease with which electrical signals can be confined by insulators and by shielding; and the enormous versatility of the devices available for manipulation of electrical signals. Microcircuitry and integrated circuits based on semiconductor electronics have added to the already great versatility of electrical amplification.

Amplifiers function by using the low-power measurement signal to control power drawn from a very large source other than the measurand. Electrical devices used for amplification include the vacuum electron tube, the transistor and related semiconductor devices, and the saturable core reactor, a type of transformer in which a small current in one winding controls a large amount of power in another winding. Electromechanical amplifiers include such devices as relays and magnetic clutches.

The relationship between input and output signals governed by amplifier principles and devices is, in general, only approximately linear and is subject to many environmental influences and to changes that occur in characteristics of components. If greater accuracy is required than can be achieved by isolation and correction or compensation, the principle of negative feedback, or closed-loop operation, is employed.

Negative feedback. The feedback principle is of general utility in overcoming distortion of measurement signals throughout the instrument system. Its operation may be understood by noting that it involves two signal channels—an information channel and a power channel. The information channel operates at the lowest possible power level and employs signal manipulation and transmission principles and devices selected for their freedom from error and instabilities.

The output signal of the power channel, including distortions introduced by processes taking place in the channel (such as amplification and transmission) is compared with the signal from the information channel, which is assumed to be correct, and the difference between the two, the error signal, is returned to a control element in the power channel in such a way that a positive error (power channel signal greater than the information channel signal) causes a reduction in the gain, or amplification factor, of the power channel and a negative error produces the opposite effect. The result is a reduction in the error signal until the difference between the two channels becomes negligible.

Clearly, the automatic adjustment of gain of the power channel to correct the power channel signal must be effected before a significant change in magnitude of the measurand has time to occur. Lags in signal transmission in the two channels will result in incorrect adjustment of gain and unstable operation unless they are small relative to the rate at which the measurand changes.

Digitalization. Another powerful tool for avoiding degradation of measurement information is conversion of the signal to coded or digital form. The analogue signals are subject to instability and to noise interference. As long as the signal remains an analogue, each operation is in essence a new measurement subject to error, so that the measurement information becomes progressively degraded at each step. Once the measurement information is represented by a number, however, such degradation is enormously reduced. Thus, if it is desired to evaluate a measurand with an error as small as one part in 1,000,000, each manipulation of an analogue signal must be performed with an accuracy of at least one part in 1,000,000. If, however, the magnitude of the measurand is represented by a decimal number, the signal simply transmits each individual digit. An accuracy of only one part in 10 permits each digit to be unambiguously identified, which is all that is required.

Furthermore, if binary-number representation is used, an accuracy of only one in two suffices. Digital signals in measurement systems are generally manipulated in binary form, by presence or absence of signal or by positive or negative direction of signal, decimal digits being represented by binary equivalents. The possibility of error in transmission or manipulation can be further reduced by introducing redundancy—that is, repetition of each signal as a check.

In the manipulation of analogue signals, small errors are more probable than large, resulting in a normal (Gaussian) distribution of error. In digital signals the various digits represent different orders of magnitude, but the probability of an error in a given digit is independent of the magnitude it represents. Thus, in digital communication of measurement information, large and small errors are equally likely, but their probability of occurrence can be made as small as desired.

Techniques for rapid manipulation of digital information in electrical form are highly developed so that digital transmission is competitive with analogue transmission even where speed or high-frequency response is important. Digital instruments are rapid enough in their response to permit their use for measurement and control of aircraft operating at supersonic speeds, of space vehicles millions of miles removed from their controllers, and of many manufacturing processes.

Advantages of digital form

Many techniques are available for converting from analogue to digital form with high accuracy. For measurement information carried by an electrical voltage signal, a digital voltmeter may be used for the conversion. When the measurement information is a periodic signal whose frequency is the analogue of the measurand magnitude, an electronic counter that presents in digital form the number of cycles occurring during a fixed counting period provides the measurement information in digital form. When linear or angular displacement is the analogue signal to be converted (for example, the motion of a pointer), digitally

coded scales are commonly used. Because of the desirability of conversion to digital form at the earliest possible stage of the measuring process, much attention has been given to development of so-called digital transducers, in which digitalization is immediately associated in a single device with the first conversion of the measurand magnitude into a measurement signal.

Magnification. Enlarging a measurement signal without increasing its power is called magnification. Mechanical, optical, and electrical techniques are all employed. Examples of mechanical magnifiers are the lever; the screw; pulley and gear systems; and hydraulic and pneumatic force multipliers. Optical magnifiers include the lenses and mirrors used to produce or modify images in microscopes, telescopes, and cameras and the optical lever used for instrument indication. Electrical magnification is provided by transformers.

The purpose of magnification is primarily to perform an interfacing function: to match the output of one element of the measurement system to the input requirements of the next. An early application (and still a major one) is to condition the measurement signal for presentation to the human senses at the interface, or point of contact, between the instrument and the observer. Its frequent function there is to match the magnitude of the readout indication to the resolution capability of the eye. It serves analogously at the interface between instrument elements, as in impedance matching between stages of an electrical measuring circuit.

Analysis. One of the most widespread characteristics of physical systems is resonance, or characteristic periodic variation. For this reason an important measurement is the determination of the resonance frequencies associated with a system. To determine these frequencies, the output of the system may be subjected to harmonic analysis. This may be done by mathematical computation, or the signal may be analyzed by using resonant detection devices.

Analysis of resonance frequencies

The most commonly encountered example of harmonic analysis is that of a radio receiver, in which a complex electromagnetic field that induces currents in an antenna is sorted into separate signals characterized by the resonant frequency of the transmitter from which the signal was broadcast. By scanning across the receiver dial and noting the points at which signals are received, a spectrum analysis of the incoming signal is performed.

Computation. General-purpose computing operations for mathematical manipulation of measurement signals include analogue computation, in which the magnitude of the signal (usually in the form of an electrical voltage) is operated upon, and digital computation, in which the magnitudes of signals are first converted to digital form and then operated upon as numbers. Both of these methods are of growing importance as components of measuring instruments.

The basic element of analogue-computing devices is the operational amplifier (opamp), a high-gain, direct-coupled, wideband amplifier available in compact modular form. Operational amplifiers are combined in various ways with other electrical-circuit elements, such as resistors, capacitors, and inductors, to perform mathematical addition, subtraction, multiplication, division, integration, and differentiation.

Digital computers combine large internal memories for storing numerical information, or any information convertible to numerical form, with arithmetic-manipulation capability. The essential characteristics of modern digital computers are very fast operation and capability for storing and operating from a program of instructions for performing complex computation. This capability of rapidly performing computations, when incorporated into an instrument, permits direct automatic measurements of complex quantities that would otherwise require combination by an observer of the results of several separate operations.

The computer also identifies and monitors signals, normalizes data, linearizes or otherwise transforms instrument response, and corrects measurement data for various influences that can be measured if knowledge of their effects is programmed into the computer; in addition, it performs several different kinds of analysis, summarizes measurement data, and generates reports (see also INFORMATION PROCESSING AND INFORMATION SYSTEMS: *Computers*).

The ability of the computer to store and manipulate instructions as well as data is also used to automate the operation of instruments performing large numbers of complex measurements. Both speed and accuracy of measurement can be improved in this way. Computer generation of complex displays for presentation of interpreted data enlarges the scope of instrument communication with the observer, and it permits the observer to interact with the instrument system, as described further in the discussion of display techniques below.

Telemetry. It is often necessary for results of measurement to be presented, essentially instantaneously, at a location remote from the point of measurement. Space exploration, biomedical research, oceanography, and meteorology are fields in which such requirements are common. Telemetry, or telemetering, the art of remote measurement, makes use of communication channels, especially wireless transmission, to transmit measurement data. In modern telemetry, data are frequently transmitted in digital form.

Display. Communication of measurement information to the observer is a function of all instruments except those that are used for automatic-control-system actuation. Even in automatic control, the results of measurement are usually displayed so that the performance of the system may be monitored. Display serves not only to present results of measurement but also to communicate with any observer who participates in, or must respond to, the measurement process. Display is usually visual, but presentation to the other senses is also common. Auditory display is used for alarms that indicate that some limiting value of measurand has been exceeded; *e.g.*, a fire alarm activated by a thermostat. The auditory sense is also enlisted negatively for detecting the absence of sound when a measurand signal and a reference signal are to be compared or when a feedback error signal is to be brought to zero.

Communicating information to the observer

The commonest visual display is the pointer moving relative to a graduated scale; the graduations of the scale are usually marked to read directly in measurand units. The relationship between measurand and pointer movement may be expressed as either a linear (that is, direct one-on-one relationship) or as a nonlinear relationship, such as logarithmic.

Linear relationship between measurand and pointer movement and uniform spacing of graduations facilitate interpolation between graduations. A logarithmic relationship between measurand and pointer deflection has the characteristic that the percent error of reading is uniform over the scale. Logarithmic scales permit larger ranges of measurand magnitude to be covered usefully by a given length of scale.

An important contributor to pointer-reading error is the parallax, or apparent difference in position, when the pointer and the scale are not in the same plane. Parallax is reduced by providing a mirror sector in the plane of the scale or by using a pointer configured as a thin vane rather than a needle. With the former, the eye of the observer, the pointer, and the mirror image of the pointer are aligned for reading; in using the vane, the eye is positioned for minimum pointer profile to improve consistency of reading.

A special kind of pointer, a cathode ray (beam of electrons) impinging on a phosphor screen, is used in oscilloscope and television-raster types of display. Deflection of the cathode ray is controlled by electric or magnetic fields, and its intensity may also be modulated to provide simultaneous display of the relationship among three variables. Addition of colour permits limited display of a fourth variable.

Use of a beam of electrons as a pointer

The high-speed writing capability of the cathode ray has made the oscilloscope one of the most important components of instrument systems and has contributed significantly to the desirability of using electrical signals to transmit measurement information. The oscilloscope has been developed to a very high level of sophistication for measurement of electrical variables and, hence, for any measurand that can be converted into an electrical signal.

In special types of equipment, the observer can interact with the measurement system through the cathode-ray tube by use of a hand-held probe that is applied to the screen at the light spot and thereby assumes control of the cathode ray and operates upon the information being displayed.

To improve the precision obtainable by analogue display, scale expansion and zero suppression are used, the total measurement range being covered by more than one span of the scale. Unlimited precision of indication can be obtained with a digital display that relieves the observer of the measurement function involved in reading a scale.

Digital display is accomplished by mechanical, electrical, and electro-optical means. Mechanical devices may use rotatable disks or cylinders the peripheries of which carry the numerals. The automobile odometer is an example.

Another mechanical type of digital display uses a series of pointers, one for each digit, that point to numerals in a circular array, as in the common gas meter, water meter, and watt-hour meter. Analogue and digital display are frequently combined in mechanical displays, the former being used for the less significant digits. One form of this uses an analogue pointer that reads on a scale covering a full circle and one or more auxiliary pointers the indications of which count the number of full rotations performed by the analogue pointer.

Electronic and electro-optical devices for digital display frequently use illuminated numerals. Preformed numerals may be selected and illuminated electrically, numerals may be formed by selection and illumination of dots in a matrix, or numbers may actually be written, as required, by cathode-ray beams. Cathode-ray tubes and television-scan displays permit the presentation of complex information in combined analogue and digital form. Graphs, tabulations, and a great variety of symbolic information may be communicated in this way.

When large quantities of measurement information must be communicated to the human observer at high speed and with high accuracy and reliability, the display art becomes especially critical. Such situations arise, for example, in air-traffic control: the observer must assimilate large amounts of measurement information and make rapid decisions based upon it.

Recording. The recording function in instrument systems serves two distinct purposes: first, it provides a record of measurement results for future use by the observer, and, second, it stores intermediate measurement information for later automatic use by the instrument system. While the requirements of the second purpose are somewhat more restrictive than the first, records in machine-readable form can serve both purposes. It is also common to combine in a single device the functions of recording and display.

Variable characteristics of recording techniques

Recording techniques have a number of variable characteristics. Among these are the visibility of the record (for simultaneous display), machine readability, immediate availability of the record, permanence, erasability, susceptibility to accidental erasure, operating-power requirement, standby-power requirement, writing and reading speed, time required for access to a specific item in the record, form of the data (analogue or digital), precision of recording, number of variables accommodated, and, finally, the techniques used to achieve the characteristics desired.

Recording and display are frequently combined by devices that produce visible markings on a surface. One very general way to do this is to substitute for the usual indicating pointer a device that produces a mark on a moving chart. The device may be an inked pen, a sharply pointed stylus used on a waxed surface, a heated stylus on a heat-sensitive surface, or electric current carried by the pointer on an electrosensitive surface.

By using a light beam as an indicating pointer, or the spot of light produced by a cathode ray, a photographic surface may be used to record the measurement information, leading to higher recording speeds. A stream of electrically charged ink particles directed to the proper point on the recording surface by an electrostatic field under control of the measurand has been used similarly to avoid frictional drag on the recording surface.

The recording surface may be a circular disk, frequently of paper, a flat sheet, a drum, or a continuous strip. To record the measurand value as a function of time, the recording surface or chart is moved by a clock motor. To record the measurand as a function of a variable other than time, the chart may be moved under control of the independent variable, or the marking pointer may be moved in two mutually perpendicular directions: one, usually called the Y direction, being used for the measurand and the other, the X direction, indicating the independent variable.

While any readout principle that uses mechanical motion for indication may be adapted to recording if sufficient power is available, general purpose recorders are usually designed for electrical input. Some recorders operate with negative feedback, that is, a portion of the output is fed back to the input out of phase with the incoming signal; this operation serves to decrease distortion and increase stability. Though it slows the operation somewhat, it permits greater precision.

Electrical input

The non-feedback (open-loop) type is characterized by errors on the order of a few percent, but faster operation, with frequency response on the order of 100 hertz (0.01 second for full-scale traverse) for mechanical pointer types and several thousand hertz for the photographically recording types called oscillographs.

The position or motion of an object that can be illuminated or is self-luminous can be directly recorded as the streak produced as the object's image on the photosensitive surface moves. Alternatively, extremely short pulses of light may be used to produce separate images at the pulse-repetition rate of the light source. Electric sparks or flash-tube discharges are used as sources of light for this type of recording.

Visible recording is also accomplished in digital form by devices such as printers and teletypewriters. Very fast printers have been developed for this purpose. When the rate at which measurement information is produced by the instrument system is beyond the capabilities of available recording means, buffers may be used to store the information temporarily.

Although techniques and devices are available to enable visible records to be read automatically into instrument systems, such records are not generally best suited for machine reading. Recording means have, therefore, been developed that have characteristics dictated primarily by the requirements of internal use in the instrument system. Such recording means are frequently called memory systems or devices. Apart from reliability, the most important of these requirements are rapid writing capability and quick access to the recorded data.

Serial-recording techniques

Serial-recording techniques in instrument systems frequently involve magnetic recording on wires, tapes, drums, and disks. By use of multiple channels, limited partial random access is achieved with these recording media.

Access to the record is relatively slow not only because it is a hunting process but also because mechanical motion is involved. Magnetic recording with high-speed random access is provided by magnetic core memory systems and related magnetic memories that involve no moving parts. Several other storage systems, however, have been developed since the mid-1970s. For example, semiconductor memories that employ tiny electronic circuits for individual memory cells are widely used in microcomputers and minicomputers because of their high random-access speed and compactness. They also are supplanting magnetic core memories in a growing number of large-scale (mainframe) computer systems, particularly those requiring extremely fast access and large storage capacity. Additional information on memory units is included in the article INFORMATION PROCESSING AND INFORMATION SYSTEMS: *Computers*.

PERFORMANCE CHARACTERISTICS

Errors in measurement may result from imperfect definition of the measurand, from imperfect understanding of the physical laws applied to converting the measurand into a measurement signal, from failure to establish completely the conditions prescribed for performing the

measurement, and from imperfections in the performance of the instrument or the operator. Some of the characteristics that describe the accuracy of instruments in a general way that is independent of the nature of the measurand or of the specific operating principles used in the instrument are described below.

Interaction of instrument and measurand. As already mentioned, all instruments either withdraw energy from the measurand or transfer energy to the measurand. In some cases the power derived from or transmitted to the measurand is negligibly small. In other cases it contributes significantly to measurement error. The measurement-signal power reacts upon the measurand so that the magnitude of the latter differs from that of the undisturbed measurand. It is the modified value that actuates the measuring instrument.

Noise. The elements of which the instrument is comprised generate internal disturbances that are indistinguishable from the signals that represent the magnitudes of measurand and reference. These disturbances are called noise by analogy with unwanted sounds. By extension the term is also applied to any unwanted signal, regardless of source. Noise interferes with the ability to evaluate the measurand. The ratio of the desired signal magnitude to the unwanted noise magnitude often sets an upper limit on the accuracy of measurement. The signal-to-noise ratio is characteristic of a particular instrument system for noise originating within the instrument. A given level of noise is most damaging when the level of the signal is low, as in the early stages before amplification.

Signal-to-noise ratio

The first stage of amplification is therefore critical for establishing a good signal-to-noise ratio. Amplification of the measurement signal as early as possible in the sequence of functions is beneficial. Thermal agitation is an important source of noise, so that operation of the first functional elements at low temperature is often used to achieve low-noise operation.

Interference. The environment in which the instrument operates contributes influences that interfere with the measurement. Once these interfering influences enter the instrument system, they become, like any other noise, indistinguishable from the measurement signal. Unlike noise, however, they differ from the measurand before they enter the system and, hence, can be dealt with by isolating the instrument from them or by providing compensation against them. Neither the isolation nor the compensation is, in general, perfect.

Dynamic response. The physical system of the measuring instrument does not respond instantaneously to the value of the measurand. In mechanical systems inertia, elasticity or imperfect rigidity, and friction or dissipation of energy all introduce lags in response. In electrical systems analogous effects are produced by inductance, capacitance, and resistance. And in thermal systems lags are introduced by thermal capacity and by thermal resistance.

The result of these lags is that the output of the instrument is not a true representation of the instantaneous value of the measurand if the latter is changing. Instead, the indicated value may be an average that is characteristic of the instrument system. This characteristic is related to the natural frequency of the system (the frequency at which the output of the instrument would oscillate if left free after a momentary input disturbance) and to its degree of damping or internal dissipation of energy.

Drift. Unless the measuring elements in the instrument are completely stable, which is never entirely the case, the indication of the instrument will drift from the correct value in the interval between calibrations. A major cause of short-term drift is self-heating, causing changes in temperature of various parts of the instrument. It is common practice to allow a minimum warm-up time before making measurements, and instruments are often maintained in continuous operation to minimize drift for important measurements.

Changes caused by drift

Instrument drift may cause two changes: (1) change of zero indication and (2) change of sensitivity, the ratio of output to input. For accurate measurement, frequency of calibration must be appropriate to the drift characteristics of the instrument. Continuous calibration based on local reference standards that are built into the instrument system may be used in critical measurement situations.

Hysteresis. Measuring instruments may not respond equally to a measurand value that is approached from a lower value and to the same value approached from a higher value. Lagging of response behind input is called hysteresis by analogy with the similar phenomenon in magnetization of materials. More complex dependence of measurement results upon the past history of the measuring instrument may also be present. To reduce such effects, the measurement may be made in both ascending and descending directions, or the instrument may be cycled through a range of values preparatory to making the final measurement.

A special case of hysteresis is the dead band. The output of an instrument may not respond to a change of magnitude of the measurand until that change has reached some finite magnitude characteristic of the instrument. Thus, a dead band exists that sets a lower limit to the magnitude of change that can be detected. A cause of dead band in mechanical instruments is backlash or lost motion in gear systems and other linkages.

The dead band

Linearity. It is desirable that the relationship of the output to the input signal can be described by a simple mathematical statement. The actual characteristic of an instrument always departs more or less from the ideal, and the precise relationship, ascertained by the process of calibration, must usually be expressed either by a tabulation of outputs and corresponding inputs or, graphically, by a curve relating the two. A linear relationship is frequently desired because this greatly simplifies the calibration process.

Resolution. The closeness to each other of two values of measurand that can be distinguished by the measuring instrument is described as the resolution capability of the instrument.

Precision. The ability of an instrument to determine with a high degree of discrimination the relationship between two values of measurand that fall within its nominal measurement range, independently of a reference standard external to the instrument system itself, is its precision. To be capable of high precision, an instrument must be capable of a high degree of resolution, but the converse does not follow. A partial measure of precision is the dispersion or scatter of measurement results obtained when the same measurand is repeatedly measured with all conditions of measurement maintained within the range specified for proper operation of the instrument.

Precision of a measurement can be improved by averaging a number of separate readings. Precision usually increases as the square root of the number of repetitions. Similar improvement is obtained by allowing the instrument to average its measurement over a period of time during which the measurand remains constant.

Accuracy. The capability of an instrument to yield the correct (or true) value of the measurand is its accuracy. Of special importance is the accuracy with which the indications of an instrument system agree with the true value defined by international agreement. This kind of accuracy provides the most broadly significant measurement data in the sense that the widest possible range is provided for communication about the data resulting from the measurement process. The International System of Units, abbreviated SI, is a modernized version of the metric system established to provide a logical and interconnected framework for all measurement (see the section *Systems of weights and measures*). Traceability of measurements to SI is provided by a hierarchy of intercomparisons of standards performed through the agency of the International Bureau of Weights and Measures, located at Sèvres, near Paris. National laboratories of various countries provide access to the International Bureau for the purpose of achieving international compatibility of measurements.

The internationally defined true values, the SI units, are disseminated in a variety of ways starting with the establishment of national standards that are the physical embodiments of the units of measurements. Methods of dissemination include calibration services, development and distribution of standard reference materials that em-

body measurands of known value, determination and publication of standard reference data, or best values, relating to materials and phenomena that are widely accessible for measurement, round-robin tests in which standards laboratories cooperate to compare results of measurements performed on identical measurands, and broadcast of time and frequency signals.

Basic SI units

The basic units of SI are the metre, the unit of length; the kilogram, the unit of mass; the second, unit of time; the kelvin, unit of temperature; the ampere, unit of electric current; the candela, unit of luminous intensity; and the mole, unit of quantity of matter. Plane angle and solid angle are expressed in terms of the radian and steradian, respectively. From these units all other units of physical measurement are conceptually derived. Except in the case of the unit of mass, the base units are defined so that they can be realized independently of the calibration hierarchy described above. High accuracy, however, is in most circumstances more readily achieved through the hierarchy.

Safeguards against failure. A measuring instrument may fail completely, ceasing to operate, or its performance may gradually deteriorate until its accuracy falls below the level of acceptability without giving any indication of failure. A common safeguard in measuring instruments is visibility, as in the liquid-in-glass thermometer, in which the condition of the instrument is easily ascertained by inspection. Instrument systems in which reliable operation is critically important may have built-in self-diagnosis capabilities.

DEVELOPING INSTRUMENTS OF MEASUREMENT

Five stages can be distinguished in the evolution of instrument systems: definition, research and development, production design, production and, application phase. These phases are not clearly separable; even in the last, it is frequently desirable to return to earlier phases to improve operation of the instrument.

The first phase includes a definition of the measurand and a set of requirements that define performance desired and conditions of operation under which this performance is to be attained. Unless the measurand is directly observable and accessible to the human senses, it must be defined in terms of other measurables. The conversion of the measurand is called transduction, and the devices that effect the conversions are transducers. The first conversion is unique in that the measurand and the circumstances in which it occurs are not subject to choice by the instrument engineer but must be accepted as presented. The first, or primary, transducer must discriminate the measurand from among all accompanying conditions and convert it into a signal accurately related to the measurand and of such a form that it can be, in turn, measured without loss of accuracy.

Conversion of the measurand

Next is formulation of a schematic functional design, a sequence of transductions that will provide the measurement information in the time, place, and format desired under the prescribed conditions of operation. After this, each of the functional elements is designed or selected to have the desired characteristics and to connect appropriately with other contiguous elements and with the human observer or operator who participates in the measurement process.

Detailed design is then effected and implemented by constructing a prototype of the instrument. When the desired performance has been successfully demonstrated, the instrument enters the so-called production-development stage, which entails full-specification design for production, selection or development of fabrication techniques, and design of the system of tests and inspections to be used for quality control. Production of the instrument is followed by application, including installation, operation, maintenance and repair, and assurance of accuracy of measurement results by standardization and calibration.

(J.Sn./Ed.)

MAJOR FORMS OF INSTRUMENTATION

The field of instrumentation is broad; it extends into almost every area of science and technology. There are instruments based upon almost every principle and practice in the fields of physics, chemistry, and electronics. Described in this section are a few of the more notable varieties and the basic principles of physics and chemistry involved in their operation.

Historical survey

EARLY HISTORY

Though modern science and industry are inconceivable without precision instruments, only since the Industrial Revolution and the rise of experimental science have such devices been indispensable. Yet there has always been a need for measuring the dimensions of buildings, the area of fields, the capacity of vessels, the weight of coins, and the passage of time; and from antiquity simple methods and devices have existed. The greatest need for precise measurement in early days occurred in astronomy and navigation, and it was in these fields that the most significant and accurate instruments were invented.

The armillary sphere

The oldest known astronomical device was the armillary sphere, used to represent the great circles of the heavens. A later modification, originating with the ancient Greeks, was the astrolabe, which could be used to determine the time, the length of the day or night, or the rising and setting of a star. The invention of the compass represented a striking advance in instrumentation. This device, which for the first time gave a directional indication without using astronomical indicators, may have originated in western Europe, or it may have been imported there from China. The earliest models consisted of a small bar of magnetized iron floating in a bowl of water on a reed or section of wood.

The invention of the telescope in the early 1600s gave mankind an instrument that extended the most important of the senses, that of sight. Galileo was first to make effective use of the device. Using a simple telescope with two lenses that magnified three times, Galileo observed lunar craters, four of the satellites of Jupiter, the phases of Venus, and many of the individual stars present in the Milky Way. Use of the telescope led to a tremendous growth in astronomical knowledge and measurement ability (see below *Telescopes*).

THE 19TH AND 20TH CENTURIES

The development of instrumentation in the 19th and 20th centuries derived impetus from four major historical movements. The first was the Industrial Revolution of the 18th and 19th centuries, with its accompanying scientific advances. The second was the 20th-century growth of modern industry. The third impetus was provided by the introduction of computers in the early 1950s, and the fourth came with the advent of the Space Age.

The Industrial Revolution. The many advances during this period can be illustrated by a discussion of instrumentation in three critical areas: dimensional measurement, electrical measurement, and analytical determination.

The manufacture of both the industrial equipment and the scientific instruments of this period necessitated the adoption of new standards of precision. Devices were urgently needed to measure the dimensions of machine tools and their products. The most commonly used instrument for this purpose, first developed in 1638 for astronomical measurements, was the screw micrometer, in which the turning of a screw is translated into a precisely measurable linear movement. As its use became widespread, refinements brought standard models capable of measuring to 0.0025 millimetre (0.0001 inch), and special models measured to 0.000025 millimetre (0.000001 inch). The precise work made possible through use of these instruments

The screw micrometer

paved the way for mass-production methods. No longer was it necessary to build each machine separately; instead, individual parts could be manufactured in large numbers and the final machine assembled from the necessary components drawn from a stock of interchangeable parts.

During the 19th century the principles of electrical phenomena were discovered and applied in the development of the electrical industries. It became necessary to devise instruments to detect and measure electrical phenomena, and these instruments laid the groundwork for the variety of modern instrumentation that depends upon electronic circuitry.

The instrumentation primarily in use during this time measured current, voltage, or resistance. The earliest measurement of current and voltage followed the Danish physicist Hans Christian Ørsted's discovery in 1820 that a current-carrying wire would cause deflection in a nearby pivoted magnet. A number of instruments called galvanometers were then developed that indicated the amount of current or voltage by using the electromagnetic force in a coil of wire carrying a current to move a pointer across a scale. The controlling force against which the pointer operated was either gravity, a spring, or a magnetic field.

Galvanometers

Precise electrical measurement was finally established in 1837 by the invention of the tangent galvanometer, the first such device in which the angle of deflection was directly proportional to the current. The growing demands of the electrical industries led to the invention in 1879 of a portable device to measure current and voltage, the spring ammeter. In this device the attraction of an iron core within a tubular coil compressed a flat strip spiral, resulting in a twist at one end, to which an indicating pointer was attached.

The most important method of resistance measurement was the Wheatstone bridge, first introduced in 1843. This method, which consisted of comparing an unknown resistor with a standard resistance, was used routinely throughout the century for resistance measurement and is still a common basic circuit used in thousands of instruments.

The third important area of instrumentation advance from the time of the Industrial Revolution to the end of the 19th century was in the field of analysis. Analytical instrumentation was primarily used in astronomy, medicine, and chemistry. Improvements in the microscope, which had been invented in the early 17th century, led by 1820 to some spectacular advances in biology and medicine (see below *Microscopes*).

The microscope gained use in the detection of adulterants in food as well as in crystallography (the scientific study of crystals) and metallurgy. Sugar solutions, such as syrup, were analyzed commercially by the polarimeter, an instrument developed in 1840, which passed polarized light through the solution. The spectroscope, developed in the mid-19th century, uses a prism to separate light into its different wavelengths. The presence of certain wavelengths reveals the chemical elements present in the substance emitting the light. The new instrument's great sensitivity permitted the detection of trace elements previously unobservable. The spectroscope achieved great importance in astronomy by making possible identification of the elements present in the stars.

Modern technology. The rapidly evolving technology of the 20th century furnished an impetus for the development of modern instrumentation, as well as the knowledge required to bring it into being. Though the greatest needs for instrumentation have been in the production and process industries, applications developed in virtually every field. The growth of technology has created a number of new instrumental needs. First, technological growth creates a need for the measurement of more variables: an industry that once needed to monitor only temperature and pressure, as it acquires greater complexity of operation, finds it necessary to monitor flow rate, humidity, acidity, and chemical composition. Second, technological growth creates situations in which old instrumentation is too slow or too insensitive. Finally, situations arise in which measuring instruments must operate or actuate control devices or serve to record various measurements automatically. Many new instruments and techniques were developed to meet these needs.

Instrumentation in the 20th century has been characterized by the increasing use of electronic circuitry. The following is a generalized, simplified description of the operation of electronic instrumentation. First, a transducer (a device for changing energy from one form to another) transforms the variable to be measured into an electrical signal. This signal may be displayed in a number of ways; it also may be recorded or used to regulate devices that control the measured variable.

Examples of transducers are the thermocouple and the photocell: heat, in the thermocouple, and light intensity, in the photocell, are converted to an electrical voltage. By no means do all transducers operate by a simple and direct conversion of the variable into an electrical signal; many utilize various indirect methods. An example of indirect conversion is the thermistor, which measures flow rate by measuring the amount of electric power needed to maintain its temperature against the cooling effects of a flow of liquid. The faster the liquid flows, the more quickly the thermistor cools, hence the more power is needed to maintain a steady temperature.

The invention of the vacuum tube greatly increased the capability of electronic instrumentation. An even greater step forward resulted from development of solid-state devices, such as transistors, which have almost completely replaced the electron tube in instrumentation, and integrated circuits, which—though smaller than a fingernail—can replace hundreds of separate electronic components.

Effect of the transistor and integrated circuit

Some of the most profound advances in instrumentation during the growth of technology have been made in analyzing the composition of various substances. This widespread and important use of instrumentation gained most momentum in the years following World War II; until that time analysis was carried out for the most part by traditional, noninstrumental methods. Spectacular gains in accuracy, versatility, and speed over the older methods were brought about in the 1940s by the introduction and refinement of such instruments as spectrophotometers, mass spectrometers, and gas chromatographs (see below).

The commercial introduction of computers in the early 1950s had a revolutionary effect upon instrumentation. Computers easily process, display, or store the output of measuring and observational devices, and they perform an endless variety of sophisticated functions with data obtained by those instruments. Such functions include detailed analysis, checking for unusual patterns, virtually instantaneous correlation of the data with information from other measuring instruments, and use of the data for immediate actuation of control systems.

The development of the computer revolutionized instrumentation in many ways. The speed of its operation made possible its use in combination with many types of instruments both in the laboratory and in monitoring processes and production lines. The computer's ability to perform complex analyses of the measurement data made possible measurements that formerly had been impossible. The exploration of outer space brought a great demand for miniaturization and computerization of instruments for remote control.

(F.W.K.)

Thermometers

GENERAL CONSIDERATIONS

A thermometer is any of a class of instruments that measure temperature. There are many definitions of temperature. In lay terms, temperature is a measure of the hotness or coldness of a system, or a measure of the ability of a system to transfer heat to another system. More scientifically, it may be defined as the property of a system that determines whether it is in thermal equilibrium with other systems. If two isolated systems at different temperatures are placed in intimate contact (*e.g.*, a piece of hot metal in a pail of cold water), the hotter system will transfer heat to the colder until both are at the same temperature, intermediate between their original temperatures; that is, until thermal equilibrium is reached. Temperature is not

a measure of the total quantity of any form of energy in a system. For a more rigorous treatment of temperature, see the article THERMODYNAMICS.

Temperature is one of the seven basic physical quantities in terms of which all other physical quantities are defined. It differs from the others (such as length and mass) by being "intensive," whereas the others are all "extensive." Length, for example, is additive: two separate one-metre lengths placed end to end establish a two-metre length. Temperature, being intensive, is not additive, however: one litre of water at 20° C (68° F) added to another litre of water at 20° C gives two litres of water still at 20° C. This phenomenon has fundamental implications for the establishment of a practical scale of temperatures. The scale or its unit cannot be established at a single temperature and extrapolated to others; it must be carefully defined over the whole range of temperatures.

Importance of thermometry

Temperature measurement and control play an integral role in a wide variety of industrial, scientific, and domestic activities. Those industries where the need is especially obvious include agriculture, aircraft and automobile industries, heating and air conditioning, refrigeration, electronics industries, mining and metallurgy, manufacturing. A familiar and important application of precise thermometry in science is in medicine. Domestically, nearly every household has at least one thermometer, be it on the furnace or the kitchen range.

DEVELOPMENT OF THE THERMOMETER

Early history. Although the ancient Greeks conceived of degrees of hot and cold and performed various simple experiments which could, in retrospect, have formed the basis of a thermometer, not until the late 16th century did the first thermoscope appear; most authorities attribute its invention to the Italian scientist Galileo (Galilei), probably in 1592. Probably not provided with a scale, the thermoscope measured only changes in temperature and these only qualitatively (see Figure 1, top left). Galileo's thermoscope consisted simply of a large glass bulb with a long, narrow, openmouthed neck inverted over a vessel of coloured water or spirit of wine (alcohol). After some air had been driven out of the bulb, the liquid rose a short distance into the neck; subsequent changes in the temperature of the bulb produced expansion or contraction of the air in it, which in turn produced changes in the level of the liquid in the neck. The idea of providing such a thermoscope with a scale, thus producing a thermometer, probably originated with Sanctorius Sanctorius, a colleague of Galileo, in 1611. He noted the level at which the liquid stood when the bulb was surrounded by melting snow and again when it was heated by a candle and divided the intervening distance into 110 equal parts.

By 1632 Jean Rey, a French physician, apparently unaware of the earlier Italian work, had invented the first thermometer with a liquid instead of air as the indicating fluid (see Figure 1, top right). He simply filled a flask having a long, slender neck with water until the liquid level lay in the neck. Changes in the temperature of the bulb produced corresponding changes in the liquid level.

In succeeding years the sensitivity of these thermometers to barometric pressure was observed. This difficulty was first overcome in about 1644 by the grand duke Ferdinand II of Tuscany, who conceived the idea of hermetically sealing the end of a liquid-filled thermometer of the Rey type (see Figure 1, bottom left). Ferdinand contributed further to the advance of thermometry by promoting the establishment in 1657 of the Accademia del Cimento in Florence. During the 10 years of its existence this society conducted extensive thermometric experiments while its workmen became highly skilled in the manufacture of thermometers of reproducible dimensions. These Florentine thermometers were widely used for over a century, some surviving to the present. The Accademia was also among the first to try mercury in its thermometers, although spirit of wine was preferred because of its larger expansion coefficient; *i.e.*, because it expanded more in response to increases in temperature.

The first mercury thermometer

These advances in liquid-in-glass thermometry did not preclude continued study of air thermometers. In Paris in 1672 Hubin (given name unknown) invented the first air thermometer to be independent of atmospheric pressure by adding a sealed bulb to the open end of a two-liquid U-tube barometer (see Figure 1, bottom right). A sealed air thermometer was similarly constructed from a three-liquid barometer in 1695 by Guillaume Amontons, a French physicist who studied seriously the principles of its operation. These were important developments, leading in the 19th century to fundamental studies in gas thermometry.

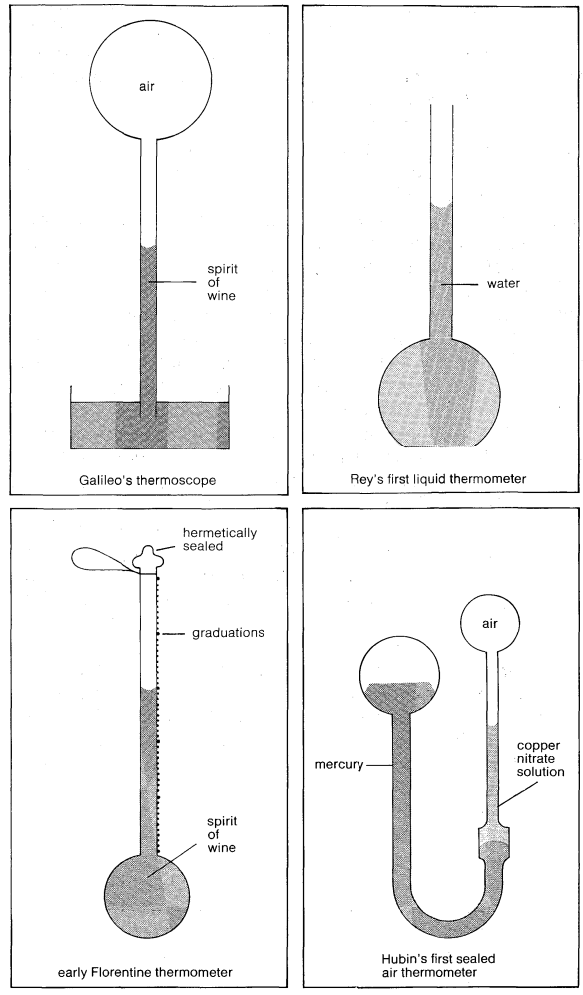


Figure 1: Early thermometers.

Development of temperature scales. For the next century, thermometry was devoted largely to the attainment of reproducible temperature scales to enable measurements made with different thermometers, in different locations, or at different times to be directly comparable. The list of contributors to these advancements includes many of the most famous names of science, such as the Englishmen Robert Boyle and Sir Isaac Newton and the Dutchman Christiaan Huygens. By the early 18th century as many as 35 temperature scales had been devised, most of them based upon one or other of two principles: (1) calibration of the thermometer at two temperatures to establish two fiduciary marks, division of the interval between them into equal parts, and linear extrapolation beyond the fiduciary marks, or (2) calibration of the thermometer at one temperature with subsequent scale divisions based upon the calculated expansion of the fluid in it. The first of these is simpler to apply, is more precise, was then the widest used, and in a more general form is the basis of practical temperature scales to this day. The second provides to some extent a fundamental rather than an empirical scale and may be considered the forerunner of the absolute temperature scale.

The two principles of thermometer scales

Fahrenheit scale. Of these diverse scales only a few survived, one of the best known having its foundations in the work of the Danish astronomer Ole Rømer. For

several years before his death in 1710, Rømer turned his attention to thermometry. He calibrated his standard thermometers according to principle (1) above, assigning the value 60 (degrees) to the boiling point of water and 7½ (later 8) to the melting point of ice, these serving as his two fixed points. For his meteorological observations he seldom used the upper part of his scale, and therefore he changed his upper reference temperature to that of blood heat, labelling it 22½. Rømer's scale is important only insofar as it influenced Daniel Gabriel Fahrenheit, a Dutch instrument maker of German descent. Fahrenheit became justly famous in the period 1700–30 for the quality of his mercury thermometers and for the agreement among them when they were calibrated on his scale. Preferring finer scale divisions than Rømer used, he chose 32 for the melting point of ice and 96 for blood heat. On this scale he determined the boiling point of water to be 212 degrees, although he did not use it as one of his fixed points. The Fahrenheit scale commonly used in the English-speaking world today derives from the original one but uses 32 and 212 rather than 32 and 96 for its two fixed points. Thus, the unit of temperature (Fahrenheit degree, ° F) on this scale is 1/180 of the difference between the boiling and freezing points of water. It is a tribute to the skill of Fahrenheit that the presently accepted value for the body temperature of a healthy person (98.6° F) is so consistent with his original scale.

Rømer's scale

Réaumur scale. At about the same time (1730), a scale based on principle (2) above was developed by the French scientist René-Antoine Ferchault de Réaumur. Though it is not absolutely clear how he proceeded, it seems he chose for his thermometric fluid spirit of wine diluted with some particular amount of water, assigned 0° Re to its level at the freezing point of water, and graduated the thermometer so 1° was 1/1000 of the total volume occupied by the fluid at 0° Re. The alcohol was so diluted as to boil at about 80° Re, but because of confused reporting it was believed that this was the boiling point of water. Thus, what has been handed down as the Réaumur scale uses 0° Re as the freezing and 80° Re as the boiling point of water. But for this historical accident Réaumur might have been credited with establishing the first centigrade scale (*i.e.,* one with the fundamental interval divided into 100 parts), because, on his original scale, the boiling point of water would have been near 100° Re. The Réaumur scale found widest acceptance in Germany.

Celsius scale. The first truly centigrade scale is generally attributed to the Swedish astronomer Anders Celsius in 1742. Earlier workers, including Fahrenheit, had already been aware that the temperature of boiling water is dependent upon atmospheric pressure, but Celsius was the first to allow for this explicitly in his calibrations. For his scale he assigned 0° to the level at which the mercury stood when the thermometer bulb was immersed in water boiling at a particular barometric reading and 100° to the level when the bulb was immersed in melting snow. Shortly thereafter the inverted form of Celsius' scale (*i.e.,* 0° for the freezing point and 100° for the boiling point of water) made its appearance in Sweden and in France. It later spread throughout most of the world and was known for two centuries as the centigrade scale until, in 1948, the name of Celsius was officially attached to it.

Absolute temperature scale. All of the temperature scales described above are essentially empirical. The basis for a fundamental scale, however, had already been laid by experiments on the properties of gases. In 1661 Robert Boyle observed that the product of the pressure (P) and volume (V) of a given quantity of air at constant temperature was constant over a moderately wide range of pressures. What is now known as Boyle's law may therefore be written

Boyle's law

$$PV = \text{constant}.$$

Further experiments many years later showed that Boyle's law is only approximately true except in the limit of very low pressures, in which case it is exactly true for all gases.

In about 1800, Jacques-Alexandre-César Charles and Joseph-Louis Gay-Lussac concluded from independent experiments that the volume occupied by any gas increases linearly with increase in temperature when the gas is maintained at constant pressure. Known as Charles's law, this result may be expressed in the form

$$V = V_0 (1 + at),$$

where V is the volume at temperature $t°$ C, V_0 is the volume at 0° C (32° F), and the constant symbolized by the Greek letter alpha, a, is the coefficient of thermal expansion. According to Charles and Gay-Lussac, a has the same value for every gas; just as before, the law is only approximate but becomes exact in the limit of low pressures, in which case a has the value $0.003661 = 1/273.15$. If this value is inserted in the above equation, it follows that, when $t = -273.15°$ C, the gas has no volume at all. For this reason the temperature −273.15° C is called the absolute zero; no lower temperature is possible. This statement of zero volume should not be interpreted literally; what it approximately means is that at −273.15° C the molecules comprising the gas are so tightly packed that no space is left between them. In practice, all substances at this temperature are in the liquid or solid state, for which the law no longer applies anyway.

If now a new temperature scale is defined so that temperatures T on it are given by $T = t + 273.15$, Charles's law may be rewritten in the form

$$\frac{V}{T} = \text{constant}.$$

This, when combined with Boyle's law, results in the general gas law (see MATTER: *Gaseous state*)

$$\frac{PV}{T} = \text{constant}.$$

It is evident from this equation that, if either the pressure or the volume of a given quantity of gas is held constant, measurement of the other may be used to determine temperature. This condition is the basis of the gas thermometer. A gas for which the general gas law holds is called ideal, or perfect; the scale on which the temperature T was defined is called the absolute temperature scale, or the ideal-gas temperature scale.

Ideal gas

Kelvin scale. Because no real gas behaves exactly as an ideal gas, most of the early experiments utilizing these principles produced slightly different scales, not only for different gases but also for the constant pressure and constant volume modes of operation. The need remained for a truly fundamental temperature scale independent of the properties of any particular substance. This need was met by the British physicist William Thomson (later Lord Kelvin), who proposed in 1848 to define a thermodynamic temperature scale in terms of the efficiency of an ideal heat engine described by the French physicist Sadi Carnot in 1824 (see THERMODYNAMICS). A heat engine may be described as a machine that converts heat into mechanical work, such as a steam or an internal-combustion engine. To produce continuous work, the machine performs a given cycle of operations over and over again. Any heat engine operates on the principle that during one part of its cycle the working substance absorbs heat from a reservoir at a high temperature, and during another part, after using some of this heat to do the work, it rejects a lesser quantity of heat to a reservoir at a lower temperature. In the case of the steam engine, for example, the working substance is water vapour, the high-temperature reservoir is the boiler, and the low-temperature reservoir is the condenser. No heat engine can be perfectly efficient; *i.e.,* can convert all of the heat extracted from the high-temperature reservoir into useful work. Carnot's contribution was to show that his ideal engine worked at the maximum possible efficiency and, further, that this efficiency was independent of the nature of the working substance and dependent only upon the temperatures of the two reservoirs.

Thomson (Lord Kelvin) showed that for the Carnot engine, if Q_1 units (joules) of heat are extracted from the hot

reservoir and Q_2 joules of heat are returned to the cold reservoir, a temperature scale could be defined such that

$$\frac{T_1}{T_2} = \frac{Q_1}{Q_2},$$

where T_1 and T_2 are the temperatures assigned to the hot and cold reservoirs, respectively, on his thermodynamic temperature scale. This is often also called the Kelvin scale, in his honour. Thomson went on to establish that the temperatures so defined are identical with those defined by the ideal-gas laws; *i.e.,* the absolute and thermodynamic temperature scales are the same.

Establishment of the Kelvin unit

To establish the size of the unit on the Kelvin scale, Thomson pointed out that in principle one temperature could be defined arbitrarily, and, by the above equation, all others would be measured in terms of it. This was impractical in his day. As an alternative and to keep the scale compatible with those already in use, he accepted the Celsius principle of setting the difference between the freezing and boiling points of water equal to 100 kelvins (the name now used for the unit on the Kelvin scale). Thus, the kelvin (symbol K without the degree sign °) has the same magnitude as the degree Celsius, and the two scales are related by

$$T\,(\mathrm{K}) = t\,(^\circ\,\mathrm{C}) + 273.15.$$

(The degree sign is still occasionally used with the symbol K when referring to temperatures or temperature differences on the Kelvin scale.)

This definition of the kelvin suffers the disadvantage that the position of the two fundamental reference points relative to the absolute zero (*i.e.,* the number 273.15) must be determined by experiment. Suppose this value can be found to within 0.01 K; then the freezing point of water on the Kelvin scale (273.15 K) is uncertain by about one part in 27,000. A very low temperature, such as the boiling point of liquid helium (4.2 K), however, is uncertain by about one part in 400. For scientific work such a large uncertainty in converting from one scale to the other is undesirable. For this reason Kelvin's first proposal for defining the kelvin is preferred.

Rankine scale. It is of course possible to set up an absolute temperature scale with a degree the same size as the Fahrenheit degree. If this is done, the absolute zero is found to occur at −459.67° F. The absolute Fahrenheit scale is defined by setting $T\,(^\circ\,\mathrm{R}) = t\,(^\circ\,\mathrm{F}) + 459.67$; it bears the same relation to the Kelvin scale as the Fahrenheit does to the Celsius. This scale is commonly called the Rankine scale, in honour of W.J.M. Rankine, a famous 19th-century Scottish engineer.

Advances in thermometer technology. Coincident with these fundamental developments, advances in thermometer technology continued. Any property of a substance that changes with temperature may in principle form the basis of a thermometer; properties other than the expansion of liquids and gases are capable of higher sensitivity and wider range. The English scientist Sir Humphry Davy reported in 1821 that the electrical resistance of a metal is dependent upon temperature. The German scientist Sir William Siemens proposed in 1871 to exploit this phenomenon to build an electric resistance thermometer. His choice of platinum for the material of the sensor was superb; as perfected by the British physicist Hugh Longbourne Callendar and others, the platinum resistance thermometer is the most stable and accurate thermometer available. It is described in detail later; at this point it is important for the role it plays in the present International Practical Temperature Scale, below.

Another electric thermometer, called a thermocouple, is based upon the discovery by Thomas Johann Seebeck in Germany in 1822 that an electric current is generated in a simple closed circuit consisting of two dissimilar metals (*e.g.,* copper and iron) when the two junctions are held at different temperatures. Thermocouples composed of platinum and an alloy of 10 percent rhodium in platinum have been developed into precise instruments, especially for reading high temperatures; they, too, form a part of the International Practical Temperature Scale.

That the temperature of a hot object might be measured by suitably sampling the heat it radiates was recognized by the 19th century. Various theoretical attempts to establish a relation between the emitted radiation and the temperature of the object culminated in 1900 in the famous law of the German physicist Max Planck. For the very special case of a blackbody (*i.e.,* a body that totally absorbs all of the radiation incident upon it), Planck developed the equation

Planck's law

$$L_\lambda(T) = c_1/\lambda^5(\exp\,[c_2/\lambda T] - 1),$$

where $L_\lambda(T)$ is the spectral radiance of the source (*i.e.,* the energy radiated per unit time per unit of projected area per unit solid angle at a fixed wavelength symbolized by the Greek letter lambda, λ), T is the thermodynamic temperature, and c_1 and c_2 are the first and second radiation constants, respectively. Not only has this law been found to describe the spectral distribution of the energy emitted by a blackbody accurately, but it also marked the beginning of the quantum theory of radiation. For his discovery Planck received the Nobel Prize for Physics in 1918. It is important that the same temperature appears in Planck's equation as in Kelvin's thermodynamic scale, so that radiation thermometry, in addition to gas thermometry, provides a practical means of measuring thermodynamic temperatures. Planck's law as applied to a particular radiation thermometer, the optical pyrometer, is also an integral part of the International Practical Temperature Scale.

International temperature scales. Although the thermodynamic scale as propounded by Kelvin was recognized as fundamental, there were practical difficulties in using it. Gas thermometers were cumbersome and unsuited to many applications. Furthermore, in many cases, knowledge of the thermodynamic temperature was not as important as the precise reproduction of temperature or temperature intervals on some practical scale. A multiplicity of such scales was obviously undesirable. Thus, in 1927 all countries adhering to the International Bureau on Weights and Measures adopted the International Temperature Scale. This scale was defined basically by extending the principle (1) listed earlier. Because of subsequent improvements in measurements and techniques, the International Temperature Scale was replaced in 1948 by the International Practical Temperature Scale of 1948 (IPTS-48), in 1960 by a revised version, and in 1968 by the currently used International Practical Temperature Scale of 1968 (IPTS-68).

The IPTS-68 recognizes the thermodynamic temperature as the basic temperature and defines its unit (the kelvin) to be 1/273.16 of the triple point (equilibrium among the solid, liquid, and vapour phases) of water (see Table 1). The IPTS-68 itself is so constructed as to closely approximate thermodynamic temperatures; its unit is also called the kelvin, or the degree Celsius, which has the same size. Temperature (T_{68}) on the IPTS-68 is defined by assigning numerical values to 11 reproducible fixed points and interpolating between them in prescribed ways. The IPTS-68 is not defined below −259.34° C. The 11 fixed points of the IPTS-68 and their assigned temperatures are listed in Table 1.

An important but subtle point regarding these scales is worth noting. By definition, the unit for thermodynamic temperatures is the same in all temperature regions; this is not necessarily so for the IPTS-68. Even if the temperatures assigned to the fixed points are their true thermodynamic values, it has not yet been established that the assigned interpolating formulas will exactly reproduce thermodynamic temperatures except above the gold point. Thus, the unit on the practical scale may not be constant; a degree in one temperature region may be of different size from that in another. Hence, the accurate determination of thermodynamic temperatures in all temperature regions is the major task of modern thermometry.

METHODS OF TEMPERATURE MEASUREMENT

Temperature measurements are required over such a wide range and in such a wide variety of situations that a very

Table 1: Defining Fixed Points of the International Practical Temperature Scale of 1968*

fixed point	T_{68} (K)	t_{68} (°C)
Equilibrium among the solid, liquid, and vapour phases of equilibrium hydrogen (triple point of equilibrium hydrogen)	13.81	−259.34
Equilibrium between the liquid and vapour phases of equilibrium hydrogen at a pressure of 33,330.6 newtons per square metre (25/76 standard atmosphere)	17.042	−256.108
Equilibrium between the liquid and vapour phases of equilibrium hydrogen (boiling point of equilibrium hydrogen)	20.28	−252.87
Equilibrium between the liquid and vapour phases of neon (boiling point of neon)	27.102	−246.048
Equilibrium among the solid, liquid, and vapour phases of oxygen (triple point of oxygen)	54.361	−218.789
Equilibrium between the liquid and vapour phases of oxygen (boiling point of oxygen)	90.188	−182.962
Equilibrium among the solid, liquid, and vapour phases of water (triple point of water)	273.16	0.01
Equilibrium between the liquid and vapour phases of water (boiling point of water)	373.15	100.00
Equilibrium between the solid and liquid phases of zinc (freezing point of zinc)	692.73	419.58
Equilibrium between the solid and liquid phases of silver (freezing point of silver)	1,235.08	961.93
Equilibrium between the solid and liquid phases of gold (freezing point of gold)	1,337.58	1,064.43

*All of the assigned boiling and freezing point temperatures are for a pressure of one standard atmosphere (101,325 newtons per square metre). The freezing point of tin (t_{68} = 231.9681° C) may be used as an alternative to the boiling point of water.

large number of types of thermometers have been developed. The principles of the more important of these are outlined below. Many considerations influence the choice of thermometer for a particular task, among them cost, durability, size, speed of response, accuracy, sensitivity, and temperature range.

Gas. Gases are used in several ways as thermometric fluids. In the gas thermometer they provide the best means of measuring thermodynamic temperatures over the wide range from about −263° to 1,000° C (−441° to 1,800° F). Only for very low temperatures, however, is the gas thermometer a practical measuring device.

Gas thermometer

The three forms of the gas thermometer. The gas thermometer may be constructed in one of three forms: constant-volume, in which the pressure of a given quantity of gas confined to a constant volume is measured as a function of temperature; constant-pressure, in which the volume of a given quantity of gas maintained at constant pressure is measured as a function of temperature; and constant-temperature, in which both the pressure and volume of a given quantity of gas at temperature T are measured and then measured again when a known fraction of this gas has been transferred to a reference volume. In all three cases one of the temperatures is a known reference (frequently the triple point of water), and the others are determined relative to it. Most gas thermometers in use are of the first type. In a typical installation a quartz or metal bulb containing the gas is held at a constant temperature and connected to a pressure-measuring device through a complex network of capillary tubing. An appropriate standard IPTS-68 thermometer placed so as to be at the same temperature as the bulb allows a direct comparison of the practical with the thermodynamic scale. The gas is usually helium, especially at low temperatures, but hydrogen, nitrogen, argon, and others have been used at high temperatures.

A number of small but important effects must be considered if accurate temperatures are to be obtained. The volume of the bulb itself, for example, and hence of the gas in it changes with temperature; the amount of change can be calculated if the thermal-expansion coefficient (fractional increase in volume for a 1° C rise in temperature) of the material of the bulb is known; metals are preferred over glasses in this regard. Again, a small portion of gas is not in the bulb but in the connecting capillary, which usually is exposed to a temperature gradient. This volume and its temperature distribution must be measured. Many possibilities for errors arise. Perhaps the most important are adsorption of gas on and desorption of gas from the walls of the bulb. These result in the nature and quantity of the gas not being constant. The uncertainties in temperatures measured with very good gas thermometers vary from a few thousandths of a degree below 100° C to a few tenths of a degree at 1,000° C.

Factors leading to error

Vapour-pressure thermometry. Gases are also used in vapour-pressure thermometry. From thermodynamic principles the equilibrium pressure of the vapour over a liquid surface may be related to thermodynamic temperature by an equation involving the latent heat of vaporization and the heat capacity of the substance. If the latter two quantities are known, then measurement of the vapour pressure will determine thermodynamic temperature; if not, the pressure–temperature relation may be established by gas thermometry and subsequently used for practical vapour-pressure thermometry. These thermometers are used only at cryogenic (very low) temperatures with high-purity gases such as helium, hydrogen, neon, and nitrogen as the working fluids. From 0.2 to 5.2 K, helium-vapour-pressure scales are recommended by the IPTS-68. Because the vapour pressure is independent of volume, the thermometers are easier to use than gas thermometers, but their ranges are discontinuous—from the triple to the boiling point of each fluid.

Filled-system thermometers. Both gases and vapours are used in industrial filled-system thermometers. In the usual construction the fluid (gas alone or liquid plus vapour) is contained in a bulb that connects through a capillary to an indicating device. Changes in the temperature of the bulb are transmitted by changes in pressure of the gas or vapour to the indicator. Gas-filled thermometers commonly contain nitrogen; the liquid in a vapour-pressure thermometer is chosen to suit the temperature range desired.

Liquid volume. The popularity of liquid-in-glass thermometers arises from their simplicity, low cost, long lifetime, wide temperature range, and general applicability. The basic instrument is not significantly different from those of the 17th century. It consists of a glass bulb connected to a closed, graduated capillary; a liquid (almost invariably mercury) fills the bulb and part of the capillary, while the remainder of the capillary is commonly filled with dry nitrogen. The liquid level rises in the capillary when the bulb temperature increases because the mercury expands more than the glass containing it. The freezing point of mercury (−39° C [−38° F]) determines the lower limit of the mercury-in-glass thermometer; the softening point of the glass determines the upper limit. Hard borosilicate-glass thermometers may reach almost 600° C and quartz higher, provided the nitrogen pressure is sufficiently high to prevent boiling or significant evaporation of the mercury. Coloured organic liquids are used below −39° C: alcohol will reach −80° C and pentane −200° C.

For accurate work these thermometers must be used in their intended manner. They may be designed to have all of the bulb and liquid column at the temperature to be measured (total immersion), all of the bulb and a specified portion of the column (partial immersion), or the entire thermometer (complete immersion). If some portion of the liquid column is not at its prescribed temperature, the observed reading may be corrected for the length of emergent stem. A thermometer with a separated mercury column will read incorrectly. The column may usually be reunited by cooling the bulb until all of the mercury has retracted into it. These thermometers also exhibit small amounts of hysteresis (they will read differently when cooled down to a specific temperature than when warmed up to that same temperature) and dimensional instability. The very best of them are capable of reading temperatures to within a few thousandths of a degree.

Liquid-in-glass thermometers are manufactured in many special forms. The clinical thermometer has a constriction

Clinical thermometers

in the bore to prevent the mercury from retreating from its maximum reading when the thermometer is removed from the body. The thermometer must then be shaken to return the mercury to the bulb. The mercury column in an oceanographic reversing thermometer is arranged to separate when inverted; the amount of separated mercury indicates the temperature at the time of inversion. Another type indicates both the minimum and maximum temperatures reached in a given time interval.

Liquids are also used in liquid-filled thermometers (similar to gas-filled ones) in which the temperature indications are produced by changes in the volume of the liquid.

Electrical resistance. *Platinum resistance thermometer.* The basic principle of the platinum resistance thermometer has already been described. Over a very wide temperature range, from 0° to 630° C, the thermometer's electrical resistance is related to temperature on the IPTS-68 by a rather complex equation. Below 0° C the relationship is even more complicated. The great value of the platinum resistance thermometer lies in the fact that, once calibrated, its readings are reproducible to a high degree of accuracy. The platinum in standard thermometers must be free from mechanical strain and have a minimum purity of about 99.995 percent. The most successful design has platinum wire of about 0.08-millimetre (0.003-inch) diameter wound as a bifilar (two-strand) helix on a mica support, the whole being about four millimetres (0.2 inch) in diameter and 20 millimetres (0.8 inch) long with a resistance at 0° C of about 25 ohms. This sensor is mounted in the bottom of a glass or quartz sheath about 40 centimetres (16 inches) long, with fine gold lead wires extending from it to the head of the sheath. Its resistance is measured when a small current (usually less than 0.002 ampere) is passed through it. This construction is suitable from about −180° to 500° C (glass sheath) or 650° C (quartz sheath). For lower temperatures the sensor is encapsulated in a short, close-fitting metal sheath; for higher, designs incorporating ceramic supports and capable of reaching 1,100° C were developed. For accurate measurements the platinum thermometer must be annealed at about 450° C for several hours at frequent intervals. It is fragile and must never be subjected to rough handling. In laboratory research these thermometers may attain an accuracy approaching 0.0001° C. More rugged designs, capable of lesser accuracy, are used industrially. Because platinum thermometers are expensive, other metal resistance thermometers, notably copper and nickel, have also been developed. Copper is limited to temperatures below 120° C, nickel to below 300° C. Neither is as stable or accurate as platinum.

Cryogenic electrical resistance thermometers. Other types of electrical resistance thermometers are suitable for cryogenic temperature measurements. One is the small (less than two-watt) carbon resistor commonly used in electronic circuitry. Its resistance increases rapidly with decreasing temperature, so that at very low temperatures it forms a sensitive, moderately stable thermometer. Different resistors are chosen for different small portions of the range from below 1 K to above 50 K.

Semiconductor thermometers

Semiconductor thermometers, in particular germanium thermometers, are the most important cryogenic electrical resistance thermometers. High-purity germanium has too high and constant an electrical resistance to be useful in itself for thermometry. When it is doped with minute quantities of an impurity such as arsenic, however, its resistance is very much lower and highly temperature-dependent. Like carbon and in contrast to metals, the resistance increases with decreasing temperature. Germanium thermometers are useful from about 0.2 to 50 K but especially from 0.2 to 20 K, where the platinum-thermometer sensitivity is very low. In this range reproducibility with 0.001 K can be obtained. A typical germanium thermometer is a small crystal enclosed in a helium-filled cylindrical platinum shell about three millimetres (0.1 inch) in diameter and 12 millimetres (0.5 inch) long with fine gold lead wires protruding.

Another class of semiconductor resistance thermometer is thermistors, which are usually small beads or cylinders about one millimetre in diameter composed of a mixture of complex metal oxides. Different thermistors suitable for various temperatures from 4 K to 300° C are manufactured; it is preferable to operate any one thermistor over a narrow range. At room temperature they are ten times more sensitive than platinum thermometers but much less stable; at cryogenic temperatures they are inferior to carbon or germanium. Germanium and platinum thermometers are expensive, however, whereas carbon and thermistors are cheap; where their lower accuracies suffice, they are widely used.

Thermocouples. Thermocouples are among the most important thermometers. The basic thermocouple consists simply of two wires of different materials joined together at one end (the hot junction), connected to separate copper lead wires at the other (the cold junction), with the leads connected to a voltage-measuring device. The wires are enclosed in a nonconducting sheath (cloth, rubber, ceramic) to insulate them electrically from one another. If necessary, a considerable length of thermocouple extending from and including the hot junction is placed in a ceramic or metal tube for protection from the environment. A temperature difference between the hot and cold junctions develops a small voltage; thus, when the cold junction is maintained at a known reference temperature (frequently 0° C), the voltage is a measure of the temperature of the hot junction. Typical sensitivities are 10 to 60 microvolts per degree Celsius. Thermocouples may be used from very low (1 K) to very high (3,000° C) temperatures—a wider range than any other type.

Innumerable combinations of wires have served as thermocouples. The most important one for laboratory use is platinum, 10 percent rhodium-platinum, with a range from −50° to 1,600° C. Because platinum is costly, however, the largest number of industrial thermocouples employ such base metals as iron, copper, and various alloys.

Rugged, long-lasting, and suitable for continuous recording or control of temperature, thermocouples are perhaps the most widely used industrial thermometers. Their lifetime is dependent upon many things, chief of which are embrittlement, from oxidation or reduction, and contamination.

The sonic pyrometer

A thermocouple is also the sensor in a sonic pyrometer, which is designed to measure the temperature of slowly moving hot gases. A portion of the gas, drawn over a radiation-shielded thermocouple probe at sonic speed to give good convective heat transfer to the probe, is subsequently returned to the gas stream. Variations are applied in pneumatic, aspiration, and subsonic pyrometers.

Radiation. Radiation thermometers have the distinct advantage that they need not touch the material the temperature of which they are measuring; their greatest application is in the field of high temperatures. The most important such thermometer is the disappearing filament optical pyrometer, in which an objective lens produces an image of the source in the plane of the tungsten filament of a small battery-powered lamp. The filament current is varied until the filament and the source image appear equally bright when viewed through an eyepiece lens and a red filter. The source temperature is obtained from a previous calibration of the pyrometer lamp (brightness versus lamp current). The red filter and the eye sensitivity together restrict the observed radiation to a very narrow wavelength band. For temperatures above about 1,500° C the source brightness is first attenuated by a neutral glass filter before the match is made. In recent forms of this instrument a photoelectric detector replaces the eye, an extremely narrow band interference filter (which may transmit an infrared rather than visible wavelengths) replaces the red filter, and the brightnesses are matched automatically by an electronic mechanism as the source and lamp are alternately focussed on the detector.

The optical pyrometer measures an apparent, or brightness, temperature and not the true one, unless the source is a blackbody. If the emissivity of the source (*i.e.*, the ratio of the energy the source radiates to that which a blackbody at the same temperature would radiate) is known, the true temperature may be obtained from the measured brightness temperature. Two-wavelength or multiwavelength pyrometers attempt to eliminate this emissivity correction by measuring the ratio of the source brightnesses at two

or more widely separated wavelengths. They have been successful only for special applications.

In contrast to these monochromatic instruments, a total-radiation pyrometer accepts approximately the whole spectrum of energy radiated by the source, which is usually simply focussed on the pyrometer's detector. Though not as accurate as an optical pyrometer, it can measure much lower temperatures (below 200° C) where the radiation lies wholly in the infrared. Radiation pyrometers are widely used in industry for continuous recording and control of temperatures. Infrared pyrometers manufactured with photodetectors respond to a broad wavelength band of infrared radiation but not to the whole spectrum. Their operation is similar to that of a total-radiation pyrometer. Both of these types also require emissivity corrections.

Spectroscopic thermometry

Another class of radiation phenomena forms the basis of spectroscopic thermometry. Extremely high temperature gaseous sources emit radiation in discrete spectral lines or in groups of closely spaced lines rather than (or in addition to) over a continuous spectral range. The source temperature may be estimated by studying the intensities, or widths, of these lines, by measuring the ratio of the intensities of a line and the continuum at the same wavelength, or by observing the absorption by the gas of blackbody radiation transmitted through it. These techniques are applied, for example, in determining the temperatures of flames, plasmas (ionized gases), and stars. In comparison with more conventional thermometry, the accuracy is very low.

Other methods. *Bimetal.* The bimetal strip, one of the most rugged and maintenance-free of all thermometers, consists simply of two pieces of different metals bonded together and rigidly fixed at one end. When this compound strip is heated, one metal expands more than the other, thus causing a rotary motion of its free end. A pointer affixed thereto moves across a graduated scale to indicate temperatures. Because the sensitivity of this thermometer is determined in part by the length of the strip, it is frequently wound in a spiral to obtain reasonable length in a compact unit. One of the metals is generally Invar, an iron–nickel alloy with a very small thermal-expansion coefficient. The other may be brass or an iron or nickel alloy. The bimetallic strip is similar in range to that of liquid-in-glass, though less accurate; it is not as fragile, and it is simpler to read. It is commonly used in thermostats and similar control devices.

Acoustic. The velocity at which a sound wave travels through a medium is dependent upon its temperature; instruments that measure temperature from a determination of the velocity of sound are called acoustic thermometers. Those that operate at frequencies above the range of human hearing are designated ultrasonic interferometers. Acoustic thermometry is an important supplement to gas thermometry, because it allows independent determinations of thermodynamic temperatures, especially from 2 to 40 K.

In most acoustic thermometers a sound-wave generator such as a fixed-frequency audio oscillator or a resonating quartz-crystal ultrasonic oscillator sets up standing waves in a variable-length cylindrical cavity having precisely parallel end faces and containing helium at a known pressure. The change in the number of nodes in the cavity when its length is changed by a precisely measured amount determines the wavelength (symbolized by the Greek letter lambda, λ) of the wave; this, together with the known oscillator frequency (f), gives the wave velocity from $W = \lambda f$. By repeating the measurement at different pressures and extrapolating to zero pressure, the wavelength at zero pressure is obtained, allowing the temperature to be calculated from a theoretical relation between the wave velocity and the temperature of an ideal gas. In this way temperatures may be measured to accuracies of ±0.005 K.

A practical type of ultrasonic thermometer having high resolution from about −20° to 300° C uses the linear temperature dependency of the frequency of a resonating quartz crystal. It is calibrated by comparison with a platinum resistance thermometer.

Magnetic thermometers

Magnetic. Gas, acoustic, and vapour-pressure thermometers become less sensitive with decreasing temperature, so that their useful lower limit is about 1 K; magnetic thermometers, on the other hand, become more sensitive with decreasing temperature and are therefore important for establishing thermodynamic temperatures from 1 to about 0.002 K. These thermometers employ the temperature dependency of the magnetic susceptibility (χ)—*i.e.*, a parameter that indicates the extent to which a material is susceptible to magnetization—of paramagnetic salts, which is expressed by

$$x = \frac{C}{T - \theta,}$$

where C and θ (theta) are constants. Cerium magnesium nitrate obeys this relation down to about 0.006 K. Below that point the equation defines a magnetic temperature, which must be related to thermodynamic temperature by some kind of auxiliary measurements. The susceptibility is usually determined by placing a spherical single crystal or a small cylindrical powdered specimen of the salt inside a coil, the mutual inductance of which is measured with an alternating-current bridge. The constants C and θ are determined above 1 K by comparison with another type of thermometer and the equation extrapolated below.

Magnetic thermometers with other salts (*e.g.*, gadolinium sulfate) make convenient interpolating instruments from 1 to 30 K. They are calibrated at two or three temperatures over the range, and intervening temperatures are interpolated using the above or some other simple relation between χ and T. It is interesting that the extension of this principle to nuclear magnetism (as distinct from atomic paramagnetism) will measure the lowest temperatures yet attained, below 10^{-5} K.

Thermal noise. In any electrical conductor the electrons move about randomly, the extent of their motion increasing with temperature. Because they are charged particles, their motion produces a small fluctuating voltage; it commonly manifests itself, for example, as noise or static in electronic circuits. From thermodynamic principles it has been shown that this thermal-noise voltage is systematically related to temperature. A thermal-noise thermometer takes advantage of this relationship to measure thermodynamic temperatures. Generally, two platinum resistors connected to an electronic amplifier are used, the first at a known reference temperature and the second at an unknown temperature. The noise generated in the two resistors is compared, and one is attenuated until they are equal. Or, alternatively, one resistance may be varied until the noise voltages are equal. In either case an inherent difficulty is to separate the noise generated in the amplifier itself from that generated in the resistors. Variations of these techniques have been proposed; *e.g.*, rather than attenuate the noise from the resistor at the higher temperature, one may inject noise into the one at the lower temperature, with the advantage of reducing the dependence upon amplifier noise. Noise thermometers may be operated at temperatures from a few kelvins to above 1,000° C.

The phenomenon of pyroelectricity

Pyroelectric. Crystals of certain solids, such as tourmaline and triglycine sulfate, exhibit spontaneous polarization when their temperature is changed (*i.e.*, they become electrically charged with opposite faces having charges of opposite polarity). The magnitude and polarity of the induced voltage is dependent upon the magnitude and direction of the temperature change. This phenomenon, called pyroelectricity, has been suggested as the basis of a differential thermometer but had not yet been fully exploited in the early 1970s. Extreme sensitivity is clearly possible, for in some materials a 1° C temperature change generates several hundred volts. Temperature changes smaller than 10^{-5}° C have been measured in this way, and 10^{-7}° C appears possible. Pyroelectric thermometers will not measure temperature itself, because the voltage gradually disappears when the crystal is maintained at constant temperature.

Gas viscosity. A gas-viscosity thermometer depends upon the resistance offered by a capillary tube to the flow of gas through it. In one such device the flow of argon through a variable-length capillary at a reference temperature is matched to the flow through a ceramic capillary at a high temperature as sensed by a differential manometer (pressure-measuring instrument). At balance, the pressure drop along both capillaries is the same, and under these conditions a simple equation relates the resis-

tance of the reference capillary to the temperature of the ceramic capillary. The temperature is then obtained from an independent calibration of the instrument.

Temperature indicators. Certain materials give one single indication of a temperature. Crayons or paints that melt or change colour within a very small temperature range may be applied to any surface and will indicate when the surface reaches a specified temperature. Pellets that melt and tapes that turn black are similarly applied. More than 100 such temperatures between 40° and 1,800° C may be measured in this way by suitable choice of indicator. The accuracy is typically within 1 percent. Pyrometric cones, small clay cones, their axes inclined slightly from the vertical, soften and bend at particular temperatures and are much used in the ceramic industries. Cholesteric liquid crystals, which indicate temperature by changing their colour, differ from the other indicators in that the colour change is reversible, and the crystals may traverse a range of colours to indicate more than one temperature. Sensitivity is high; a 0.1° C temperature change is detectable.

Thermography. Thermography, or mapping of external temperature distributions over extended areas, can be accomplished with the indicators described above or by photography, especially with infrared-sensitive emulsions. Scanning and recording with infrared pyrometers is common in meteorological applications. Another method employs phosphors that emit visible light when irradiated with ultraviolet. The intensity of the emitted radiation is temperature-dependent. The phosphor may be painted on a surface, irradiated, and the temperature distribution observed visually or photographed.

Mapping internal temperature distributions

Shadowgraph, schlieren, and interferometer techniques when applied to transparent mediums, such as high-temperature gases, can be employed to map internal temperature distributions. These techniques exploit the temperature dependence of the index of refraction, that property determining the direction and velocity with which a light ray traverses a medium. In the shadowgraph and schlieren techniques an extended uniform beam of light is shone through the medium; temperature gradients within cause small displacements of individual rays within the beam. These effects are reproduced as variations in the intensity of the emergent beam and may be displayed by suitable projection on a screen or photographic plate, where they appear as shadows. In the interferometer an initial monochromatic beam is split so that only one portion traverses the medium, is subsequently reunited, and is then projected on a screen. Interference fringes, indicating contours of constant temperature, result from the velocity changes undergone by the individual rays in one beam relative to those in the other. This method is capable of quantitative evaluation. (R.E.Be./Ed.)

Microscopes

The microscope is an instrument for producing enlarged images of objects that are too small to be seen unaided. Such images may be viewed directly or with a viewing screen or photographic apparatus. Objects for microscopy may be either transparent or opaque. The compound light microscope (see below *Compound microscope*) can be used for viewing objects as small as bacteria, and the electron microscope (see below *The electron microscope*) extends the range to small viruses and large molecules.

HISTORICAL DEVELOPMENT

Simple microscope. The simple one-lens microscope may date from the middle of the 15th century, when low-power magnifiers were used for the examination of insects. From these early beginnings, glass lenses of increased power were developed, culminating in ground lenses of remarkable excellence produced by a Dutch microscopist, Antonie van Leeuwenhoek (1632–1723). In the 18th century, simple microscopes were produced in a cheap, portable form known as the screw-barrel type, which used a screw thread to focus the lens upon the object. Further developments of the simple microscope included perfection of mounts and microscope stands intended for specialized studies in botanical and other fields. After the introduction, in the 1830s, of perfected achromatic lenses, which rendered images free from colour distortions, simple microscopes lost favour except for very low power work, such as in dissection.

Early application

Compound microscope. The basic idea of the compound microscope, in which two or more lenses are arranged in such a way as to form an enlarged image of an object, probably occurred independently to several people at about the same time, in the period 1590–1609, in The Netherlands; Hans Jansen, his son Zacharias, and Hans Lippershey have all been cited at times as its inventor. Shortly after that, microscopes made by opticians in Italy and in England were circulating widely.

Though no real advances were recorded in the development of the optical parts of the microscope during the 18th century, some mechanical variations were introduced. Brass superseded wood, cardboard, and leather as a material for construction of the stand, and better mechanical means were introduced for focussing.

The image produced by the lenses of early compound microscopes suffered considerably from chromatic aberration; *i.e.,* objects in the field of view appeared to be surrounded by fringes of colour. Experiments with telescope lenses had shown that images free from colour could be produced by combining lenses of different dispersions. Dispersion is the variation of refractive index with wave-

By courtesy of (left) the Wellcome Trustees, London, (right) the Medical Museum of the Armed Forces, Institute of Pathology, Washington, D.C.

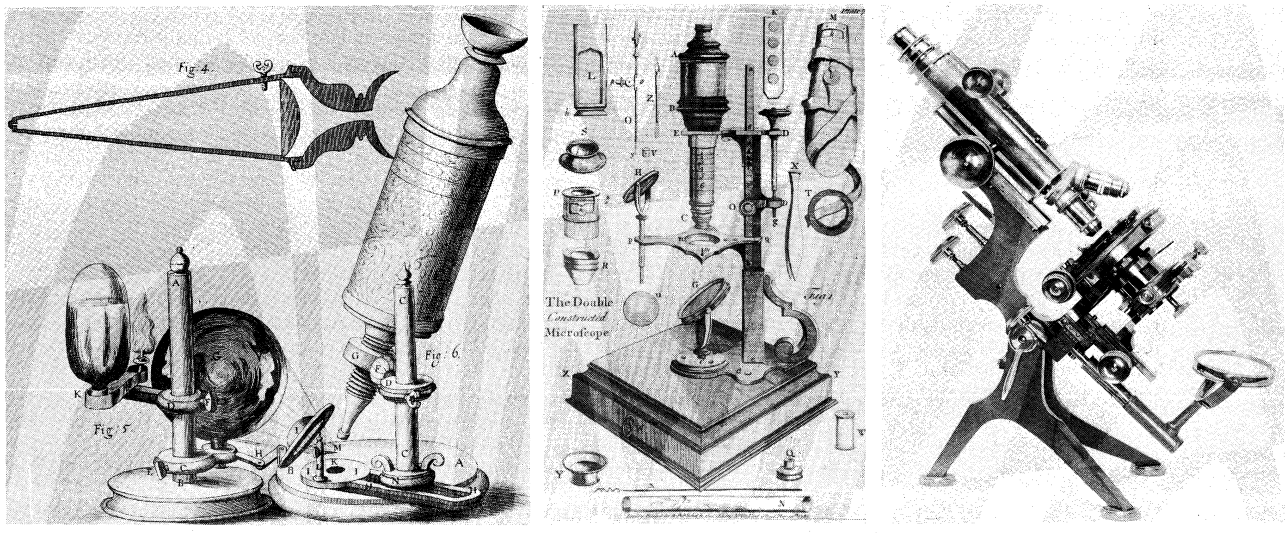


Figure 2: (Left) Hooke's microscope from *Micrographia* by Robert Hooke, 1665. (Centre) Cuff microscope from the second edition of *Essays on the Microscope* by George Adams the Younger, 1787. (Right) Watson microscope made to the specifications of Henri van Heurck by W. Watson & Sons, London, 1904.

length, that for blue light being greater than for longer wavelengths in the red region of the spectrum. The telescopes used a combination of a convex (outward-curved) lens made of optical glass of low dispersion (crown glass) together with a concave (inward-curved) lens of flint glass with a high dispersion (see below *Telescopes*).

Achromatic microscope lens. As glass technology advanced, application of the principles used for telescopes to microscopes became possible; the first achromatic microscope lenses were constructed in The Netherlands at about the end of the 18th century. Many more achromatic lenses of low power (and often of dubious quality) were produced in the period between 1810 and 1820. High-power achromatic lenses proved much more difficult to construct, because the accumulation of spherical aberrations or distortions caused by curvature in the lens glass often negated the effects of achromatism.

At that time lenses were still being designed empirically, and introduction of the use of test objects, such as the markings on the wing scales of moths and butterflies, enabled some progress to be made with the higher powered lenses. It was with the work of the English microscopist Joseph Jackson Lister, published in 1830, that construction and design of achromatic microscope lenses was placed on a theoretical basis. Lister showed how two separate achromatic lenses could be combined to act as a single lens so that the image rendered would be completely free from chromatic aberration and from distortions caused by curvature of the lens. This discovery opened the way to the construction of high-power microscope lenses in the next 50 years.

Abbe's contributions

For both his contributions theory of the instrument and his practical designs, a German physicist, Ernst Abbe, stands out as the foremost figure in the development of the microscope in the latter part of the 19th century. Abbe placed the theory of microscopic image formation on a sound basis by emphasizing the importance of the light diffracted by the object in resolving (making distinct) its fine detail. He introduced the concept of numerical aperture as an expression of the resolving power and light-gathering capacity of a microscope lens, thus permitting its direct comparison with other lenses. Abbe also developed oil-immersion lenses, in which a drop of oil was placed between the object and the front element of the lens. Very soon this type of lens, with its much improved resolving power and high-quality images, was an essential part of the microscopist's equipment. Later, in conjunction with the German firm of Carl Zeiss, Abbe introduced a whole new series of excellent, highly corrected lenses, which rendered the highest quality images yet obtainable. These distortion-free (apochromatic) lenses allowed full practical realization of the theoretical resolving power of the optical microscope by the end of the 19th century.

Developments so far in the present century have included introduction of larger stands with built-in illumination systems and higher standards of accuracy of construction. Especially notable is phase-contrast microscopy (see below *Phase-contrast microscope*), introduced in 1934, which made it possible to examine living cells and tissue specimens without staining them. Other recent trends are the development of flat-field objective lenses for photomicrography and of methods of using computers for the analysis of microscopical images.

Interest has continually centred on the increase of resolving power of the microscope; the work of Abbe showed that this was possible only by increase in the numerical aperture of the optical system (which was not feasible beyond a certain limit) or by the use of a shorter wavelength for the illuminating light. This latter possibility led to the use of ultraviolet radiation as an image-forming agent, but there were technical difficulties with this type of microscope. Little progress in the use of shorter wavelengths was made until the development of the electron microscope.

THE SIMPLE MICROSCOPE

Basic design

A simple microscope, a single positive lens or lens combination, is often known as a magnifier, or loupe. Supplied in a mount furnished with a handle, such a lens is commonly called a reading glass. Magnifiers are often furnished in a hinged mounting between two cover plates, so they may be conveniently carried in the pocket. Occasionally, they are mounted on an arm above a simple specimen-support stage so as to serve as a dissecting microscope. Magnifiers supplied in a short tubular mount, designed to be held in place in front of the eye by contractions of the facial muscles around the orbit, are used for fine instrument work and are called watchmaker's eyeglasses. The apparent size of an object viewed directly by the eye depends upon its distance from the eye. As an object is brought nearer to the eye, the angle which it subtends at the eye increases, and so the retinal-image size (and hence the apparent object size) also increases.

This process continues until the object has been brought so close that the eye no longer forms a clear image. This point is generally termed the nearest distance of distinct vision (in a young adult, about 25 centimetres, or 10 inches), and further movement of the object toward the eye will cause the image to be blurred. If, however, a convex lens is placed between the object and the eye, the sharpness of the image may be restored. If the object is at or within the distance between the lens and the point at which parallel rays passing through the lens converge, or focus (focal length), then light rays entering the lens from the object are bent, or refracted, toward the optical axis (to the centre). If these ray directions were prolonged backward, they would meet at a point in space behind the object. It is here that the brain imagines an enlarged image of the object to be located. Termed a virtual image, this image is magnified and erect.

Field of view. Maximum apparent magnification is reached when the virtual image of the object lies at the nearest distance of distinct vision. The area of the object that is seen through a magnifier (the field of view of the magnifier) is related directly to the diameter of the lens and inversely to the distance between the lens and the eye. Thus, to achieve the largest possible field of view with any given lens, the lens should be held as close as possible to the eye. The position of the object should then be adjusted to obtain the sharpest image. In practice, simple magnifiers suffer considerably from aberrations that affect the peripheral parts of the field of view; curvature of the field, distortion, and lateral chromatic aberration (different colours produce different image sizes) are principally involved.

The simplest form of low-power magnifier is a single biconvex (convex on both sides) or plano-convex (convex on one side and flat on the other) lens. Such lenses have a wide field of view; the eyeball rotates in its socket in order to scan the whole field. Spherical aberration, coma (variation in magnification over the area of a lens), and lateral chromatic aberration are not especially troublesome at the low magnifications involved, but there is appreciable curvature of the field and pronounced pincushion distortion (variation in image magnification radially outward from central point of image). In high-power magnifiers (up to 20 magnitudes, or 20 ×), the field of view is limited, and the eyeball does not rotate to scan it. The chromatic and spherical aberrations of this type of magnifier must be very carefully corrected. Distortion is often troublesome, because the distance between the eye and the lens is no longer small in comparison with the focal length of the magnifier (as it is in lower power lenses).

Types of magnifiers. Four principal types of high power magnifiers are recognized.

Stanhope lens. The Stanhope lens (Figure 3A) consists of a solid block of glass with one polished plane (flat) surface lying in the focal plane of the spherical convex upper surface of the lens. To be viewed, the object must be placed in contact with the plane surface; such lenses suffer from a large degree of distortion, and there is considerable lateral chromatic aberration.

Coddington lens. The Coddington lens (Figure 3B) has two convex surfaces but is characterized by a circular groove cut around the periphery of the lens in such a way as to act as a diaphragm. This improves the image quality, and these lenses are superior to the Stanhope type.

Wollaston doublet. The Wollaston doublet (Figure 3C) uses two plano-convex lenses, with their plane faces

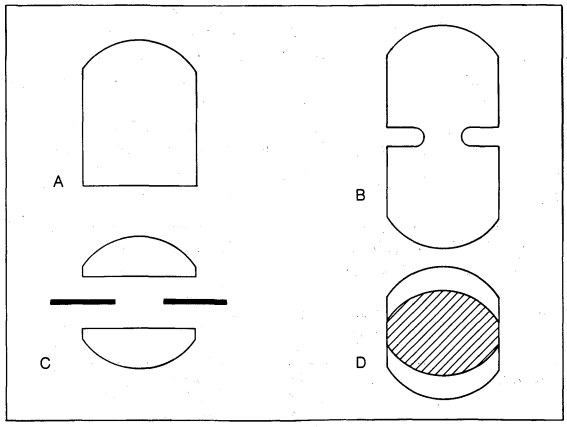


Figure 3: *Various magnifiers.*
(A) Stanhope lens. (B) Coddington lens. (C) Wollaston doublet. (D) Steinheil achromatic triplet.

pointed inward. The lenses are separated by a distance calculated so as to allow a good correction of the lateral chromatic aberration and the spherical aberration and are fitted with a central diaphragm. This type of lens, when used as a magnifier, suffers from an inconveniently short working distance and a restricted field of view.

Steinheil and Hastings magnifiers. Achromatic doublets or triplets are now almost universally used for practical magnifiers; the best known are due to Steinheil and to Hastings. Lenses of this type generally consist of a biconvex (convex on both sides) lens of crown glass cemented between and enclosed by two lenses of flint glass (Figure 3D). These cemented triplets are well corrected for spherical aberration and coma (for rays up to 25° from the optical axis), and their distortion is low. At the same time, they produce a very bright image and have a conveniently long working distance.

THE COMPOUND MICROSCOPE

A compound microscope has an objective lens (the lens closest to the object being viewed) and an eyepiece so mounted that they may conveniently be moved relative to the specimen to focus the image. An illumination system, usually added, consists of a condenser-lens system to focus the light onto the object and either a mirror with an external light source or, as is more common today, a built-in light source (commonly a low-voltage tungsten bulb).

Illumination system

In the compound microscope, magnification depends on the objective lens and the eyepiece. The objective lens produces a real image at some finite distance behind its back focal plane, a distance which is known as the optical tube length.

Compound microscopes may be provided with a single body tube for viewing with one eye; such monocular instruments are used for teaching and for work that is not extended or too exacting on the eyes. The more elaborate microscopes used for research purposes or for long periods of study are generally binocular (two eyepieces fitted with some means of distributing the light between them). A typical monocular microscope would normally be provided with three objective lenses with focal lengths of 16, 4, and 2 millimetres, giving initial magnifications of about 10, 40, and 100 magnitudes, respectively. If two eyepieces were provided with a magnifying power of six and 10, for example, the range of total magnification would extend from 60 to 1,000 magnitudes.

Monocular microscopes

Body tube. The body tube, with its associated focussing mechanism, is carried on the arm of the microscope, which is in turn supported by a horseshoe or tripod base. Attached to the arm, at a right angle to the optical axis, is the stage on which the object is placed. Below the stage is the substage fitting, which carries the condenser lens (used to focus light on the object) and either the mirror (if fitted) or the built-in light. The simplest microscopes have no substage. Instead, they have a plain object stage and a body tube fitted with only a coarse focussing movement. A more elaborate instrument has, in addition, a substage with its own centring and focussing movements. The body tube has both coarse and fine focus and a revolving turret at the bottom fitted to carry several objective lenses so that any one may quickly be brought into the optical axis. The largest and most elaborate research stand is of much sturdier construction and has an illuminating system built into the base of the instrument. A mechanical stage carries the slide and allows it to be moved easily and precisely. A binocular tube is provided for viewing with, possibly, a third, monocular tube permanently fitted to carry a camera. In the past, microscopes had their focus controls mounted on the upper part of the arm. The current trend is to place these, together with the stage-movement knobs, at a low level so they may be operated without moving the hand and arm from the table.

For many years the horseshoe form of foot was popular, though the tripod foot was preferred in England. With the introduction of the built-in light source, both of these have given way to construction in which the arm and foot are solidly joined together, the foot being in the form of a box.

The body tube is mounted on a slide or slides to allow it to be moved for focussing. The coarse focus is actuated by a rack-and-pinion mechanism, while the fine focus, working on a slide with a much more restricted range of movement, is worked either by a screw of very fine pitch that has its motion converted and translated to the slide by a lever or by some cam or bell-crank arrangement. The fine focus must have a very free action but one that operates without any appreciable backlash. In some modern stands, in order to achieve maximum rigidity, the coarse- and fine-focus mechanisms act on the stage instead of on the body tube. Such an arrangement is easily adapted to the provision of low-level controls and to the fitting of concentric coarse- and fine-focus control knobs.

Specimen stage. The specimen stage holds the object under examination and must be fixed accurately at right angles to the optical axis of the microscope. In many microscopes, a simple stage with a central hole and spring clips to hold the slide suffice, but research instruments are often provided with either a detachable or a built-in mechanical stage. This device holds the slide and allows it to be moved in two directions by screw threads. Such stages allow the slide to be moved very precisely and systematically over its whole area, and, as they are often fitted with graduated vernier scales, they enable the position of objects to be located and recorded precisely. For work with polarized light and photomicrography, it is an advantage to have a mechanical stage that also allows rotation of the specimen around the optical axis of the microscope.

The objective lenses of the microscope are fitted to a disk of metal known as the nosepiece, which is accurately centred to the microscope axis. Rotation of the nosepiece allows any of the lenses to be brought quickly into use, and the accurate alignment, or centration, ensures that the object will still remain approximately in the centre of the field of view. If more accurate centration is required, it is usual for each objective lens to be mounted on its own individual carrier, which is provided with centring screws.

The substage carries the condenser and the mirror, if fitted. It slides in a rackwork mechanism to allow adjustment of its position. In some of the cheaper instruments, however, the substage is fixed to the underside of the stage and consists of a helical mount in which the condenser slides. An iris diaphragm, a device which opens and closes to control the size of the illuminating cone of light, is often carried on the substage, rather than on the condenser itself, although this latter arrangement (which is found in the more expensive instruments) is preferable. One essential feature of the substage of any instrument intended for serious work is the provision of screws that allow the condenser to be centred with respect to the optical axis of the microscope.

The substage

The built-in light source of large modern microscopes is often mounted in the base of the instrument and comprises a small low-voltage bulb provided with collector lenses and arrangements for focussing and centring the light source. The light is reflected into the condenser either by a front-surface mirror or by a prism. The binocular body employs an arrangement of prisms to

present the light from the microscope objective to each eye.

Objective lenses. The value of a microscope lies in its ability not only to magnify objects but also to make their fine details visible. These two features depend for the most part on the quality of the lens that forms the primary image. Such an image must be sharp; the aberrations that reduce the sharpness must therefore be reduced to a minimum. Microscope objective lenses originally were single biconvex lenses that suffered severely from aberrations; nowadays all objective lenses, even those of the lowest power, are corrected as fully as possible. Chief corrections are for chromatic aberration (which produces colour fringes around the edges of the image of an object) and for spherical aberration (which is caused by the curvature of the lens). Spherical aberration causes a hazy, indistinct image, lacking in contrast. Chromatic aberration may be corrected by combining a concave negative element made from glass of a different dispersive power from that of the original convex lens. At the same time, the curvatures of such a combination may be calculated to correct part or all of the spherical aberration as well. Though such corrections are relatively easy with a lens of low magnification, satisfactory correction requires increasingly complicated lens design as power and numerical aperture (discussed below) increase. This usually involves addition of more elements. The design and construction of lens systems of high power and aperture are quite difficult, and these still remain the most costly parts of any microscope.

In addition to the primary aberrations just mentioned, coma, astigmatism, distortion, and curvature of field also have to be corrected by the designer of the lens. Coma causes magnification to differ according to the zone of the lens through which the rays pass, so that portions of the image axis are badly distorted. Astigmatism, if present, results in different appearances in different focal planes. A pinhole would appear to be a radial line at one focus and a tangential line at another. This aberration may cause misinterpretation of the shape of objects if severe at the centre of the field of view; a slight amount of astigmatism is tolerable, however, at the periphery of the field of view. Curvature of field causes the image of a plane object to appear on a curved surface. This means the centre of the field of view may be in sharp focus while peripheral areas are blurred, or vice versa. This defect increases in severity with increase in the power of the objective lens. Though not an intolerable drawback in visual work, curvature of field is undesirable when photographing through a microscope. It may be largely corrected by the use of a special eyepiece, but a more satisfactory solution involves the use of specially designed flat-field objective lenses. Empirical studies in the 19th century led microscopists to realize that the ability of their microscopes to resolve fine details was in some way connected with the angle of the cone of light rays accepted by the objective lens. This was termed the angular aperture. Ernst Abbe, in 1873, showed that this was not entirely accurate, and he was able to demonstrate that the detail resolved was proportional not to the angular aperture but to the sine of half this angle (α). Abbe gave the name numerical aperture (N.A.) to this mathematical relationship and showed that in the case of a dry lens—*i.e.*, one in which the cone of light rays from the object passes through air—the numerical aperture = sine α. In an immersion system, in which oil or sometimes water or glycerin is placed between the front lens of the objective and the specimen, the numerical aperture is increased and the value of sine α has then to be multiplied by the value of the refractive index of the immersion medium.

Numerical aperture

The numerical aperture of a lens thus is a direct measure not only of its resolving power but also of its light-gathering capacity. Light-gathering capacity, however, is not of great significance in microscope objective lenses, and high numerical apertures are used for the consequent gain in the resolving power that accrues. A further advantage of the use of the concept of numerical aperture is that the fine-detail-resolution performances of two lenses may be directly compared by the use of their respective numerical apertures.

Microscope objective lenses may be classified according to either their focal length or their type and degree of corrections. The most common focal lengths are 75, 50, and 25 millimetres for low-power lenses, 16 and eight millimetres for medium power, and four and two millimetres for high powers. Lenses may also be classed as achromats, semi-apochromats, or apochromats, in ascending order of accuracy of corrections; flat-field objective lenses may be produced in any of these grades. Some of the features of the common lens types are summarized in Table 2.

Table 2: Characteristics of Various Types of Microscope Objectives

type	focal length (mm)	numerical aperture	initial magnification	working distance (mm)
Achromatic	50	0.11	2.7 ×	50
dry	25	0.18	5.5 ×	16
	16	0.25	10 ×	5.0
	4	0.65	40 ×	0.5
Achromatic	4	0.85	40 ×	0.35
oil	2	1.25	100 ×	0.1
Apochromatic	16	0.30	10 ×	4.85
dry	8	0.65	20 ×	0.75
	4	0.95	47 ×	0.18
Apochromatic	2	1.40	90 ×	0.06
oil				
Semi-apochromatic	16	0.30	10 ×	4.0
dry	8	0.60	25 ×	0.54
	4	0.75	40 ×	0.33
Semi-apochromatic	3	1.25	63 ×	0.65
oil	2	1.30	100 ×	0.12

Achromatic objective lenses. The commonest types of microscope objective lens, achromatic lenses, are corrected so that rays of two widely differing wavelengths (usually in the red and the blue) are brought to a common focus. This correction is usually carried out for an intermediate zone of the lens; if it were done for the marginal zone at the edge of the lens, then the centre would be only poorly corrected, and vice versa. A suitable compromise, therefore, is achieved by correcting for the intermediate zone, which then results in minimal aberration in both the centre and marginal zones. As only two colours are fully corrected, there is at times some remaining colour, known as the residual colour or the secondary spectrum. This appears as a fringe around the edge of objects with sharp contrasts, but it is usually not prominent.

Residual colour

Achromats are also corrected for spherical aberration and coma. This type of objective lens is much used on microscopes intended for teaching and for routine visual observations.

Apochromatic objective lenses. The invention of new types of optical glass in the 1880s, together with the use of the natural mineral fluorite in lens making, enabled Abbe to develop a new type of microscope objective lens in which the colour correction was greatly improved but at the cost of a considerable increase in lens complexity. An apochromatic lens usually possesses a greater numerical aperture than an achromat of equivalent focal length. Care is needed in the use of apochromats because the high power lenses have very short working distances. Expensive, they are ideal for critical work at the limit of resolution of the optical microscope and for photomicrography in colour.

Improved colour correction

All apochromatic lenses (and semi-apochromatic lenses, which are more common) suffer from a kind of error known as lateral chromatic difference of magnification. This means that the image will be larger in blue light than in red. Thus, dark objects with sharp contrasts at the periphery of the field of view appear fringed with red or orange on their outer margin and with blue on their inner sides. This may be eliminated by the use of special compensating eyepieces that have been computed to introduce a compensatory error of similar magnitude but of opposite direction. This cancels out the lateral chromatic differences of magnification and thus gives the best possible image.

Semi-apochromatic objective lenses. Sometimes called fluorite lenses, these are a compromise between the achromats, on the one hand, and the apochromats, on the other.

This compromise allows a reduction in complexity and, hence, in cost. Fluorite and the new optical glasses are used in their construction, but there are fewer elements than in an apochromatic lens. The colour and spherical corrections are better than those of an achromat, the secondary spectrum being much less obvious and, since the number of points of meeting between air and glass (air–glass interfaces) in such a lens is fewer than in an apochromat, the image contrast is often surprisingly high. Again, the numerical aperture of a semi-apochromat is usually 15–30 percent higher than that of an achromat of equal focal length. The lateral chromatic difference of magnification noted above is present in this type of lens, so it is used with a compensating eyepiece to obtain the best possible results. Recently, makers have been deliberately introducing lateral chromatic differences of magnification into their achromatic lenses (which do not normally show this particular fault) in order to eliminate the need for two separate types of eyepiece.

Spherical aberration and the coverslip. With some high-aperture dry objective lenses the protective piece of glass, or coverslip, overlying the mount may introduce a surprising amount of spherical aberration. Though the effect is negligible with low-power objective lenses and may be disregarded, high-aperture apochromatic lenses will give a satisfactory performance only if this factor is taken into consideration. Lenses are normally computed for a definite thickness of coverslip (often 0.17 millimetre), but extra thickness of mounting medium and variation in the thickness of the slip itself may cause wide divergence from this figure. Errors introduced in this way may be compensated for by altering the tube length of the microscope, shortening it if the cover is too thick and vice versa. Alternatively, most high-aperture apochromats are provided with a rotatable collar on the lens mount that varies the separation between the elements of the lens and so adjusts the correction of the lens for spherical aberration. The performance of an oil-immersion lens is, of course, not affected by such variations in coverslip thickness, since there is no air between it and the front element of the lens.

Metallographic lenses

Lenses intended for metallographic work and for studies on opaque objects are normally computed so as to work without a coverslip. In the majority of cases, these lenses are also designed to work with vertical illuminators and must be provided with special antireflection coatings at points of contact between air and glass.

Flat-field objective lenses. All objective lenses of the achromatic, semiapochromatic, and apochromatic types suffer to some extent from curvature of the field, which becomes more marked as the power increases. Such field curvature is troublesome in photomicrography (photography through a microscope), and new lenses have been developed in which this fault has been largely removed. But this has made it necessary to increase the complexity of the lens. In one manufacturer's series of flat-field lenses, the curvature of field produced by the objective lens is cancelled out by a lens in the microscope tube that interposes an equal but opposite curvature of the field. The system is so designed that the same negative lens is common to all the objective lenses of that series, and, hence, it may be fixed in the microscope tube itself. Both apochromats and achromats are available in flat-field versions.

Objective-lens mounts. All microscope objective lenses are fitted into metal mounts that are attached to the nosepiece. To protect the optical elements from damage resulting from the lens inadvertently being brought into contact with the microscope slide, most modern lenses are in resilient spring mounts that compress on light contact with the object slide. Most objective lenses of any one maker are fitted into mounts in such a way that very little adjustment of focus is needed on changing from one power to another. Such lenses are said to be parfocal. Mounts are normally engraved with the magnification of the lens, its numerical aperture, and, usually, the focal length, the tube length at which the lens operates, and the thickness of coverslip for which the lens is computed. A coloured band around the barrel assists in rapid identification.

Eyepieces. The eyepiece forms a further enlarged final image from the real primary image furnished by the objective. The eyepiece may also carry pointers or transparent measuring scales (graticules) and correct for residual aberrations in the image produced by the objective lens. This is important in the compensating eyepieces used with apochromatic and semiapochromatic lenses.

Graticules for expediting measurements

Huygenian eyepiece. Two types of eyepieces are in common use. The Huygenian eyepiece (Figure 4A), sometimes called a negative eyepiece, has a plano-convex field lens nearer the objective lens with a second, smaller plano-convex eye lens above it. The field lens is so called because

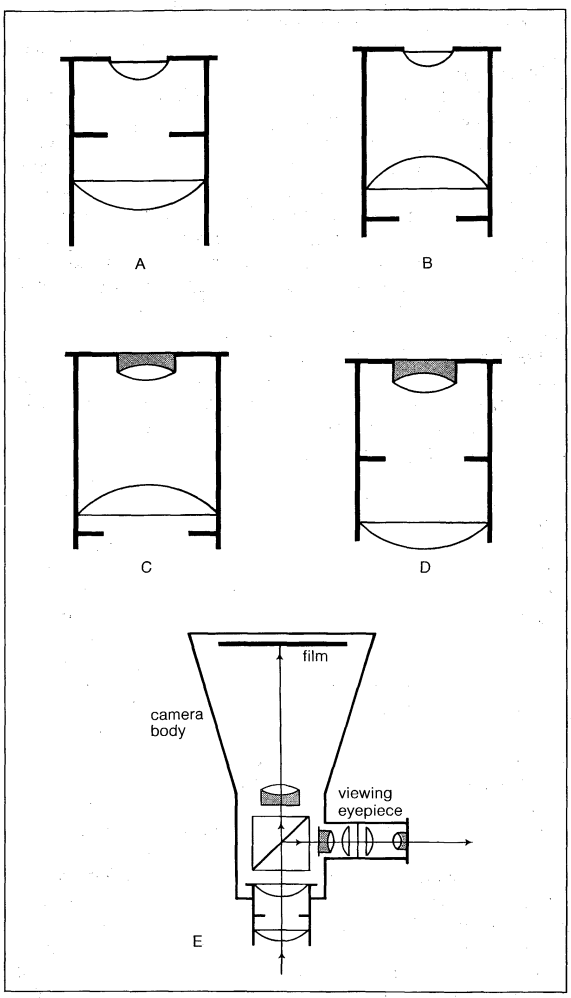


Figure 4: *Various eyepieces.*
(A) Huygenian. (B) Ramsden. (C) Achromatized Ramsden. (D) Compensating. (E) Photomicrographic.

it bends the pencils of rays from the objective towards the optical axis of the microscope without altering their individual convergence or divergence, thus increasing the angular field of view. After leaving the eyepiece, all the rays pass through a small circular area above the eye lens; this is known as the Ramsden circle, or eyepoint. Its diameter should approximate the diameter of the pupil of the observer's eye in order to obtain the largest field of view of maximum overall brightness. In the Huygenian eyepiece, the focal length of the field lens is about twice that of the eye lens, and the two lenses are separated by a distance roughly twice the focal length of the eye lens. It is usual to place a diaphragm between the lenses, which limits the field of view.

Ramsden eyepiece. The Ramsden, or positive, eyepiece (Figure 4B) has two plano-convex lenses of approximately equal focal length, arranged with the plane side of the field lens towards the objective lens and the plane side of the eye lens towards the observer. These are separated by a distance of approximately two-thirds their focal length. The focal plane of the eyepiece lies below the field lens, and the diaphragm is placed at this point. The Ramsden eyepiece is especially suited to carry measuring graticules

and pointers. Both Huygenian and Ramsden eyepieces have uncorrected lenses, but the spherical and chromatic errors of each lens are arranged to cancel out.

Kellner eyepiece. The Kellner (achromatized Ramsden) eyepiece (Figure 4C) resembles the Ramsden very closely but has an achromatized doublet as an eye lens.

Compensating eyepiece. The compensating eyepiece (Figure 4D) used with apochromatic lenses also has an achromatized doublet for its eye lens; the corrections are matched to a specific objective lens. Formerly, eyepieces were available with adjustable compensation. These could be used with any apochromatic objective lens simply by adjusting the separation between the field and eye lenses of the eyepiece.

Photomicrographic eyepiece. For photomicrography, the eyepiece shown in Figure 4E is now much used, in conjunction with a miniature camera. A device called a beam-splitting cube produces two identical images, thus allowing the user to focus and observe the object up to and at the moment of exposure. Demonstration eyepieces allow a student and teacher to examine the same microscope field simultaneously by means of a similar type of beam-splitting arrangement. The addition of a pointer that is visible to both viewers is common in this type of eyepiece. A comparison eyepiece presents the images from two separate microscopes side by side, so that a detailed comparison of them may be made. Such a device is of great value in many industrial applications and in forensic microscopy (*i.e.*, the study of material intended for use as evidence in legal proceedings).

Condensers, illumination, and resolution. Although very low power objective lenses may be used on the microscope with the illumination provided from the light cone of a concave mirror, all higher power lenses require a special system of lenses below the stage to provide a cone of light rays of sufficient numerical aperture to match that of the objective lens in use. A lens system of this type is called a condenser. It is usually in a mount that can be moved along the optical axis in order to focus the condenser and that has spring-loaded screws to centre the condenser to the optical axis. Other functions of the condenser are to provide sufficient light on the object to allow the use of high magnifications and, in some cases, to provide special types of illumination. Examples of these types are oblique illumination and hollow cone illumination, sometimes used to obtain contrast in unstained or transparent preparations.

For low powers, a simple convex lens (sometimes called a spectacle lens condenser) may be used. This is preferable to the use of a concave mirror because the condenser is provided with a diaphragm that allows some control over the working aperture of the system. The commonest type of condenser is a two-lens system known as the Abbe condenser. It is inexpensive, simple, and much used for visual observations with routine and student microscopes, though it suffers very badly from chromatic and spherical aberration, which often make it impossible to utilize the full aperture of illuminating cone provided by the condenser, which in turn makes it impossible to use the full aperture of the objective. This results in loss of resolution at high powers. An improved three-lens Abbe condenser has slightly better spherical corrections. For certain kinds of careful work, an achromatic condenser is essential. Such condensers are corrected for both spherical and chromatic aberration.

Abbe condenser

Illuminated field. The size of the illuminated field is a function of the focal length of the condenser. Thus, when low-power objective lenses are used, it is advantageous to employ a long-focal-length condenser to obtain a fully illuminated field of view. The same end may sometimes be achieved, however, by removing the top lens of a standard condenser, which then works at a reduced aperture but with an increased focal length. A pancratic (zoom) system, which permits change of focus from a large illuminated field at low power to a small area at high power, may be employed. Alternatively, an auxiliary lens may be fitted in the light path to achieve the same end.

If the condenser is arranged to provide a hollow cone of light rays which are of such an obliquity that they cannot enter the objective lens directly, then only the light diffracted by the object enters the lens and participates in image formation. The effect of this is that the specimen appears bright upon a dark background. Dark-ground microscopy gives images of high contrast but requires an intense light source and an accurately centred optical system.

Objects may be examined with either transmitted or incident light. Most of the work in biology is with transmitted light; the opposite is true for materials such as metals and stone. Either transmitted- or incident-light methods may be adapted to give a reversed contrast effect (the object appearing bright against a dark background).

Light source. In all modern microscopes, the light source is an electric lamp of some form. The condenser of the microscope is used to focus an image of the surface of the lamp into the plane of the object. As the light source is in focus at the same time as the object detail, this method of illumination is called source focus; the term critical illumination has also been used. It is essential that the light source be fairly large and homogeneous to obtain an evenly illuminated field of view. A valuable refinement involves the addition of an iris diaphragm in front of the lamp to provide some control over the area of the objective that is illuminated. To minimize glare, this area should not be much in excess of the visual field. In most research microscopes, the illumination is arranged for the instrument in such a way that an image of the light source is formed in the back focal plane of the microscope condenser by a corrected condenser on the lamp house. The lamp is provided with a variable field stop (similar to the aperture diaphragm on a camera), and the microscope condenser is focussed so that this stop is seen in the plane of the object. When this has been achieved, the evenly illuminated lamp condenser serves as the apparent light source; hence, an evenly illuminated field of view results from the use of a small, intense, but nonhomogeneous light source. This method of illumination is called Köhler illumination, after its inventor. For optimum results, the image of the lamp filament should be large enough to occupy a large portion of the area of the aperture of the back focal plane of the condenser. In practice, this means that the filament should be composed of several parallel coils—the so-called solid-source filament—and that the lens of the lamp-house condenser should be of a suitable focal length. The built-in systems of illumination that are now common comprise a small low-voltage bulb, collector lens train, and field stop.

Source focus

Illumination of an opaque specimen from above the stage is often accomplished by some form of vertical illuminator. In its simplest form, this is a thin glass slip inserted into the objective lens holder at 45° to the optical axis. Light is admitted through a small aperture in the side of the lens mount and is reflected down into the object through the lens, which thus acts as its own condenser. The light reflected from the object then passes back through the objective lens and glass slip to form the image. Many of the incident illuminators now available are designed for dark-ground work. In such cases, the incident light is reflected downward by an annular (ring-shaped) mirror around the objective and is then concentrated onto the object by a ring condenser mounted in the housing. An image of the specimen is formed by the objective lens, which is inserted into the centre of the illuminating housing. The reflected light enters the microscope tube by passing through the central hole in the mirror. Such dark-ground systems often give greatly enhanced contrast, as compared with standard vertical illumination.

Resolving power. Because of diffraction effects, the image of a point source of light can never be perfect, even assuming that all the imperfections in the lens are corrected. The image of a luminous point thus appears as a bright circular disk surrounded by a series of concentric light and dark rings. This pattern is called the Airy disk, after Sir George Biddell Airy, a 19th-century English astronomer who first studied it mathematically.

Any image may be considered as being composed of innumerable overlapping diffraction patterns, each corresponding to a given point in the object. Two point sources

that are close together will produce Airy disks that appear side by side. If the separation between the objects is reduced, then the diffraction patterns begin to overlap; if the overlap is considerable, then the two points are not seen as separate—*i.e.,* they are not resolved.

The best resolution should be realized when the full aperture of the objective lens is used and the wavelength of the illuminating light is decreased. This has been shown to hold in practice up to a point, beyond which diminution of contrast causes an apparent loss of resolution, in that the visibility of object detail is impaired. This loss of contrast is, however, distinct from the resolution proper. Modern expressions of the theory of microscopic resolution take into account the degree of coherence of the illuminating light.

If an optical instrument is resolving two closely situated points in the object, this fact will not be apparent to the observer unless the magnification of the instrument is such that the final image points subtend an angle of at least one minute of arc at the observer's eye. There is thus a minimum magnification below which the detail will not be perceived. In practice, this is usually taken as a value 500 times the numerical aperture of the lens. Conversely, excessive magnification is useless, since no more information will be presented to the eye, and the diffuse nature of the separate Airy disks making up the image will become apparent, giving an impression of poor image quality. A magnification, therefore, of 1,000 times the numerical aperture is regarded as the upper limit, values in excess of which will constitute so-called empty magnification.

SPECIALIZED TYPES OF MICROSCOPES

Stereoscopic binocular microscopes. In contrast to the binocular image of the ordinary microscope, these produce a truly stereoscopic image; they are invaluable for the examination at relatively low magnifications of many objects, for the assembly of very small electronic circuits and solid-state devices, and for dissecting and detailed surgical operations. An early form of this type of instrument was the Greenough binocular, which consisted of two complete microscope systems with their axes inclined at an angle of about 16° so that they converged in the plane of the object. Most stereoscopic binocular microscopes are provided with prisms to give an erect image for more convenient examination and manipulation of the object. Though the objective lenses may be duplicated, the tendency today is to use a single large objective lens with a device called an aperture-splitting system arranged above it that presents the images to the two eyepieces.

Aperture-splitting system

Ultramicroscope. The ultramicroscope was developed in 1903 as part of a project to study colloidal particles. These tiny particles, suspended in a liquid and too small to be seen in an ordinary microscope, are illuminated from the side with a powerful beam of light at right angles to the microscope's optical axis. Because the particles scatter light, their positions and movements may be seen as flashes against a dark background. Particles as small as 5 to 10 nanometres (about one-tenth the wavelength of visible light) can be observed.

Metallurgical microscope. This type of instrument is designed for studying the structure of opaque materials, particularly metals. It may be inverted so that a flat surface of the polished and etched specimen can be placed face down on the stage. It has a vertical illuminator and is usually equipped for photomicrography.

Phase-contrast microscope. Contrast in the image rendered by a microscope is usually due to variations in the light-absorbing ability (optical density) of the specimen. Staining the specimen enhances this effect for certain wavelengths. Living cells and other transparent specimens, however, are very difficult to study, and inducing contrast by staining is frequently impossible. The phase-contrast microscope provides an invaluable tool for the examination of such materials, since contrast is introduced by optical means alone and without any manipulation of the specimen. It does, however, suffer from some optical defects, the most notable being that accurate measurements of the image are almost impossible to make.

Interference microscope. With the interference microscope, contrast is introduced into transparent objects just as in the phase-contrast instrument but without optical defects inherent in the phase-contrast system; in addition, measurements are relatively easy to make with this type of microscope.

Ultraviolet and fluorescence microscope. Invention of the electron microscope was preceded by continuous effort to increase the resolution of the microscope. One approach was to use shorter and shorter wavelengths, moving into the ultraviolet region of the spectrum. The first ultraviolet microscopes, using ultraviolet light to illuminate the specimen and special quartz lenses, showed a twofold increase in maximum resolving power, as compared with an instrument of the same numerical aperture using green light. Actual operation was tedious, however, largely because of difficulties in focussing the image and the necessity for recording all of the results by means of photography. Recently, the development of television cameras sensitive into the ultraviolet region of the spectrum has much simplified the recording and display of the image produced by the ultraviolet microscope.

With the development of the high-pressure mercury arc burner, a source rich in shortwave radiation became easily available. Similarly, the development of glass filters that transmit light between accurately known limits of wavelength in the ultraviolet portion of the spectrum have combined to make practical use of ultraviolet light for excitation of fluorescence (conversion of ultraviolet light to visible light) under the microscope. The light is focussed onto the specimen, using either a normal or a dark-ground condenser. Certain compounds absorb this energy and reemit it as fluorescence of a longer wavelength in the visible region of the spectrum. This primary fluorescence is often of characteristic colour and may be of value for the identification of chemical compounds. The objective lens and eyepieces of a fluorescence microscope are similar to those normally used, except that it is desirable to choose lenses in which the cement between components does not itself show appreciable fluorescence. In order that any ultraviolet light that has not been absorbed by the specimen should not reach and damage the eye of the observer, a barrier filter is usually fitted between the objective lens and the eyepiece. The image is formed by the light emitted from the specimen itself, which thus appears luminous against a dark background. Ultraviolet microscopes find increasing use today in biology and medicine.

Fluorescence

Reflecting microscope systems. Interest in ultraviolet microscopy stimulated the redevelopment of reflecting systems, which had been tried at various times since the 18th century. The use of mirrors rather than lenses for ultraviolet is advantageous because of the complete freedom of mirrors from chromatic aberration and because, once focussed in visible light, the same focus applies to ultraviolet and infrared wavelengths. Coupled with these advantages is the fact that the working distance of a mirror system is much longer than that of a conventional lens system of comparable numerical aperture. Drawbacks to the use of mirror systems are that at least one of the reflecting surfaces must be aspherical (*i.e.,* with a nonspherical surface, which makes them difficult to manufacture) and that the front surface mirrors of such a mirror system do not have a very long life because of atmospheric attack and tarnishing. The reflectors used in this instrument consist of a large, primary, aspherical, concave mirror and a smaller, secondary, spherical convex mirror arranged axially with it. Only a few microscopes of this type have actually been constructed.

Reflecting systems have primarily been employed as additions to conventional optics. One ingenious system is added to a lens in order to obtain a long working distance. The mirrors are arranged to project an image into the front focal plane of a lens. Such devices are valuable for examining objects heated in a small vacuum furnace on the stage of the microscope or for carrying out the dissection of living organisms under the microscope.

Infrared microscope. By the addition to the microscope of an image converter (a device that converts an image formed by invisible light, such as infrared or ultraviolet, into a visible image), the longer wavelengths that constitute

the infrared region of the electromagnetic spectrum may be made visible. This device attached to a microscope has been used to investigate the infrared transmission or reflectivity of objects and to render visible details in objects which may be invisible in ordinary light. Though applications are numerous, one particularly valuable field is the microscopic investigation of forgeries. Another expanding use is detection of surface defects in silicon crystals used for transistors and other semiconductor devices.

Polarizing microscope. The polarizing microscope is used by geologists and crystallographers as well as by biologists. Ordinary light consists of waves vibrating in all directions. Plane-polarized light waves vibrate in only one plane. A polarizing microscope uses plane-polarized light to illuminate the specimen, which appears bright against a dark ground. In the course of passing through the microscope, the polarized light splits into two beams with different velocities. Their relative difference when they leave the specimen is called the retardation. When measured in nanometres (1 nanometre equals 10^{-9} metre) and divided by the thickness of the object measured in the same units, it gives a quantity termed the birefringence, a numerical value of help in the identification of minerals under the microscope.

Birefringence

Television microscope. The application of a television camera to a microscope is useful for recording and displaying the image to large audiences and for teaching purposes. More recently the scanning inherent in the formation of a television image has been used in measuring portions of the image and in automatic processing of the data obtained. An example of this application is found in the Quantimet image-analyzing computer, in which a microscope image is projected into the photocathode of the tube of a television camera; the output from the line scan passes after amplification not only to a monitor screen for visual inspection but also to a detector that responds to changes in the output voltage as the scanning spot passes over features of varying optical density. Specialized computer circuitry then enables the measurement of the area of specific features and permits the number of items in the field to be counted. Various other parameters may also be measured. Not only does such automation save operator time, but it also increases the degree of accuracy of the measurements.

Laser microprobe. A laser is a device for producing a very intense beam of light that can be sharply focussed. In the laser microprobe, a tiny beam of light from a pulsed laser (which produces laser light in pulses rather than continuously) is focussed onto the specimen from a vertical illuminator system that uses the objective lens of the microscope to focus the beam and allow observation of the area to be analyzed. The intense heat generated by the pulse of laser light vaporizes part of the sample. In one commercial form of this instrument, this vapour short-circuits an auxiliary gap between two electrodes positioned just above the specimen, and it is thus raised by the ensuing electrical spark to a temperature at which its spectral (light) emission may be measured by means of a device for analyzing light (spectrograph). This instrument has been used to analyze traces of metals that occur in living tissues with a minimal amount of damage to the tissue and has applications in the metallurgical, ceramic, and geological fields.

Holographic techniques. The coherence and intensity of laser light can be used to record the microscope image as a hologram, or three-dimensional representation in the form of an interference pattern. The hologram freezes the specimen at one instant in time, and it may then be reconstructed and viewed at leisure.

X-ray microscopy. Theoretically, the use of shortwave X rays should provide a great increase in the resolving power of the microscope; the practical difficulties, however, are substantial, and predicted improvements in resolution have not been achieved. X rays are not refracted by solid glass lenses, so that conventional methods will not serve for the formation of an X-ray image. A few experimental X-ray mirror microscopes have been built; aberrations are high, however, and resolution thus suffers.

In practice, X-ray microscopy is used not to achieve high resolutions but to analyze the X-ray absorption characteristics of the specimen. For this purpose, contact microradiography may be used. In this process, the object is placed in close contact with a fine-grained photographic emulsion and exposed to a beam of X rays. After the plate has been developed, the X-ray image may be examined with a normal optical microscope.

An alternative technique is point-projection X-ray microscopy, in which an electron gun and condenser system mounted in an evacuated column direct an intense beam of electrons onto a metal cathode, which forms the end window of the tube. A beam of X rays thus generated emerges and radiates from the foil. If a specimen is placed close to the target foil, its shadow will be cast onto a fluorescent screen or photographic plate placed some distance away. With a focal camera with a length of 10 centimetres and with the object about 0.1 millimetre from the X-ray spot, an initial magnification of 1,000 magnitudes will be achieved. Such an instrument suffers in that exposure times are long because there is difficulty in getting sufficient current into a very small electron spot, and the X-ray output is thus very low.

Point-projection X-ray microscopy

The X-ray projection microscope has the advantage that its depth of field is very large; stereoscopic photographs may be taken to show the organization of the structures in thick specimens. The image contrast is largely due to differences in the absorption of X rays, and it is often possible to identify substances present by measuring the X-ray absorption at different wavelengths.

THE ELECTRON MICROSCOPE

Fundamental research by many physicists in the first quarter of this century suggested that cathode rays (*i.e.,* electrons) might be used in some way to increase microscope resolution. Louis de Broglie, a French physicist, in 1924 opened the way with the suggestion that electron beams might be regarded as a form of wave motion. De Broglie derived the formula for their wavelength, which showed, for example, that, for electrons accelerated by 60,000 volts, the effective wavelength would be 0.05 angstrom—*i.e.,* one hundred-thousandth that of green light. If such waves could be used in a microscope, then a valuable increase in resolution would result. In 1926 it was demonstrated that magnetic or electrostatic fields could serve as lenses for electrons or other charged particles, and this discovery initiated the study of electron optics. In 1931 a two-lens electron microscope was devised. It produced enlarged images of the electron source. Later, in 1933, a true electron microscope was built. A year or so later, the resolution of the optical microscope was surpassed, but the specimen was severely damaged by the heating effects of the electron beam, a problem that was not solved for some years. In 1935 the construction of the first commercially built electron microscope was begun in England, and it was soon followed by production microscopes in Germany and the United States. Progress in the construction of electron microscopes was rapid, with a great impetus to the achievement of high resolution being the introduction, in 1946, of a device called a stigmator, which compensates for astigmatism of the objective lens.

Electron microscopes are now available that are easy to operate and reliable and that give point-to-point resolutions of less than three angstroms. Such high resolutions, coupled with suitable advances in the methods of specimen preparation, permit the direct visualization of large molecules. Recently, the transmission electron microscope has been supplemented by the introduction of a scanning electron microscope (Figure 5), in which a beam of electrons is scanned over the surface of an object and used to build up an image of the details of the surface structure. Another development is the electron-probe microanalyzer, which allows an elemental analysis of the composition of materials to be made by using the incident electron beam to excite the emission of characteristic X radiation by the various elements composing the specimen. These X rays are detected and analyzed by X-ray spectrometers built into the instrument. Such probe microanalyzers are able to produce an electron scanning image so that structure and composition may be easily correlated.

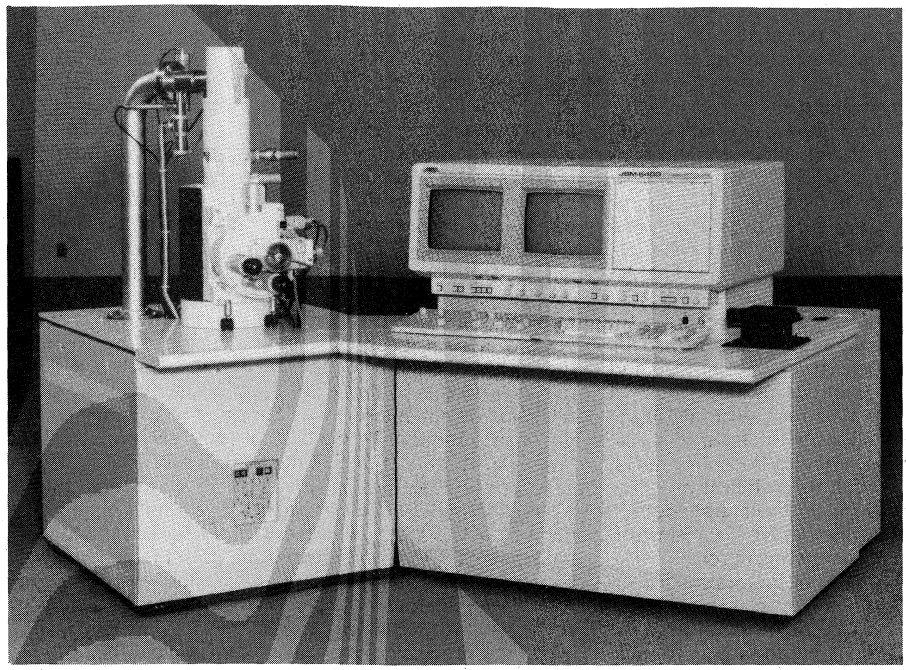


Figure 5: Digital scanning electron microscope.
By courtesy of JEOL U.S.A., Inc.

Differences of types. There are many differences between optical and electron microscopes. The electron microscope requires that the electron beam be in a vacuum, because electrons cannot travel any appreciable distance in air at atmospheric pressure. The column of the electron microscope is evacuated by pumps, and the specimen, photographic plates, and any other necessary apparatus are introduced into the vacuum by means of airlocks or seals. Unlike the optical microscope, in which the lenses are of fixed focus and the distance between specimen and objective lens is varied, the electron microscope has variable-focus lenses, and the distance between specimen and objective lens and separation of the lenses remain constant. The magnification is determined mainly by the value of the current (for magnetic lenses) through the intermediate and projector lens coils. The image is focussed by changing the current through the objective lens coil. A high degree of electrical stability is necessary because the current in a magnetic lens controls its focal length and magnification. Another difference is that the optical microscope is usually operated (for visual observations, at least) so that the image is a virtual one, and in the electron microscope the final image is invariably real and is visualized on a fluorescent screen or recorded for study on a photographic plate.

In the optical microscope the image is formed by absorption of light in the specimen; in the electron microscope the image results from a scattering of electrons by atoms in the specimen. A heavy atom is more effective in scattering than one of low atomic number, and the presence of heavy atoms will increase the image contrast. The electron microscopist may incorporate more heavy atoms into his specimen for this purpose.

Principal forms

Lenses. Electron lenses may be either electrostatic or electromagnetic. The simplest form of electrostatic lens consists of two coaxial cylinders with voltage potential difference applied between them. Such a lens is not of practical use in the electron microscope. A more convenient arrangement is the unipotential, or Einzel, lens. In it, three coaxial cylinders or apertures in plates form the lens. The outer two are grounded, while the middle one is at a high negative potential.

Electromagnetic lenses, which are used for the condenser system and for the projector of the electron microscope, differ from the objective lens only in details. The manufacturing and performance tolerances of condenser-system and projector lenses are much tighter than they are for those used as objective and intermediate lenses.

Efforts to improve the resolution of the electron microscope have tended toward production of a single-field condenser-objective lens of low aberrations. In such a lens, the first part acts as a condenser and the second as the objective; the specimen is inserted into the centre of the lens, where the axial magnetic field (the field along the axis of the instrument) is at a maximum.

Electrostatic lens aberrations

Electrostatic lenses show spherical aberration, distortion, coma, astigmatism, and curvature of field. Electromagnetic lenses, because of the rotation of the electron beam during its passage through the lens, also suffer from other less significant aberrations. In addition, chromatic aberration due to variation in the wavelength of the electrons may be of considerable importance in both types of lens. Such changes of electron velocity may be due either to variations in the high-voltage supply to the electron gun or to energy losses from collisions of electrons with atoms in the specimen. The first of these may be controlled by careful stabilization of the high-voltage supply, but variations in the second cannot be reduced with the electron sources presently in use. For the very thin specimens and the electron energies commonly used, such variation due to energy loss in the specimen may be neglected. The resolving power of the microscope is severely limited by the spherical aberration of the electron lenses. It is not possible to correct this aberration in the same manner as in the optical microscope, because magnetic electron lenses are always convergent lenses. Careful lens design may help, but at present the only practical answer to this problem is reduction of the numerical aperture of the electron lens.

Astigmatism in the electron microscope is largely due to asymmetries in the axial component of the magnetic field of the lens and is the result of imperfect construction of the lens or of the interaction of the electron beam with residual gas molecules in the column. This leads to deposits on the specimen that introduce asymmetries into the magnetic field when they charge up under the influence of the beam. Some astigmatic images may cause great errors in interpretation.

Principal components

Transmission electron microscope. The transmission electron microscope has three essential systems: (1) an electron gun, which produces the electron beam, and the condenser system, which focusses the beam onto the object, (2) the image-producing system, consisting of the objective lens, movable specimen stage, and intermediate and projector lenses, which focus the electrons passing through the specimen to form a real, highly magnified image, and (3) the image-recording system, which con-

verts the electron image into some form perceptible to the human eye (Figure 6). It usually consists of a fluorescent screen for viewing and focussing the image and a camera for permanent records.

Additional requirements are a vacuum system, consisting of at least two pumps and their associated valves, and electric-power supplies with essential stabilizers and control equipment.

The electron gun and condenser system. The source of electrons, the cathode, is a heated, V-shaped tungsten filament with its apex, from which the bulk of the emission occurs, toward the condenser lenses. The filament is surrounded by a shield, sometimes called a Wehnelt cylinder, with a central aperture arranged on the axis of the column; the apex of the filament is arranged to lie at or just above or below this aperture. The cathode and shield are at a high negative potential and are insulated from the rest of the instrument. The final electrode of the electron gun is the anode, which takes the form of a disk with an axial hole. Electrons leave the cathode and shield, accelerate toward the anode, and, if the stabilization of the high voltage is adequate, pass through the central aperture at a uniform speed. The control and design of the electron gun are critical in ensuring satisfactory operation.

The intensity and angular aperture of the beam are controlled by the condenser lens system between the gun and the specimen. A single lens may be used to converge the beam onto the object, but, more commonly, a double condenser is used. In this the first lens is strong and produces a reduced image of the source, which is then imaged by the second lens onto the object. Such an arrangement is economical of space between the electron gun and the object stage and is more flexible, because the reduction in size of the image of the source (and hence the final size of illuminated area on the specimen) may be varied widely by controlling the first lens. The use of a small spot size minimizes disturbances in the specimen due to heating and irradiation.

The image-producing system. The specimen, mounted on a copper grid, is carried in a small holder in a movable specimen stage. The operator can control it by a system of micrometer screws and bell cranks working through vacuum seals, so that a large area of the specimen can be examined systematically. The objective lens is usually of short focal length (one to five millimetres) and produces a real intermediate image with a magnification of 20–200 magnitudes. An area of this image is rendered in turn with further magnification by the projector lens or lenses. A single projector lens may provide a range of magnification of 5:1, and by the use of interchangeable pole pieces in the projector a wider range of magnifications may be obtained. Modern instruments employ two projector lenses (one called the intermediate lens) to permit a greater range of magnification and to provide a greater overall magnification without a commensurate increase in the physical length of the column of the microscope. Such a system may be operated so that each stage in turn provides magnification, or the intermediate lens may be used to reduce the image produced by the objective and the projector set to magnify part of it to give the final image.

For practical reasons of image stability and brightness, the microscope is often operated to give a final magnification of 1,000–50,000 magnitudes on the screen. If a higher final magnification is required, it may be obtained by photographic enlargement. The quality of the final image in the electron microscope depends largely upon the accuracy of the various mechanical and electrical adjustments with which the various lenses are aligned to one another and to the illuminating system. The lenses require power supplies of a high degree of stability, and stabilization is provided electronically for a satisfactory standard of resolution. Image quality

The gun, specimen chamber, and column and camera are evacuated so that the electron beam can pass freely. At the place where the specimen and photographic plates are inserted, air locks provide that only a minimal amount of air is introduced into the column; this reduces pumping time before the microscope is again operational.

Image recording. The electron image must be made visible to the eye by allowing the electrons to fall on a fluorescent screen fitted at the base of the microscope col-

By courtesy of JEOL U.S.A., Inc.

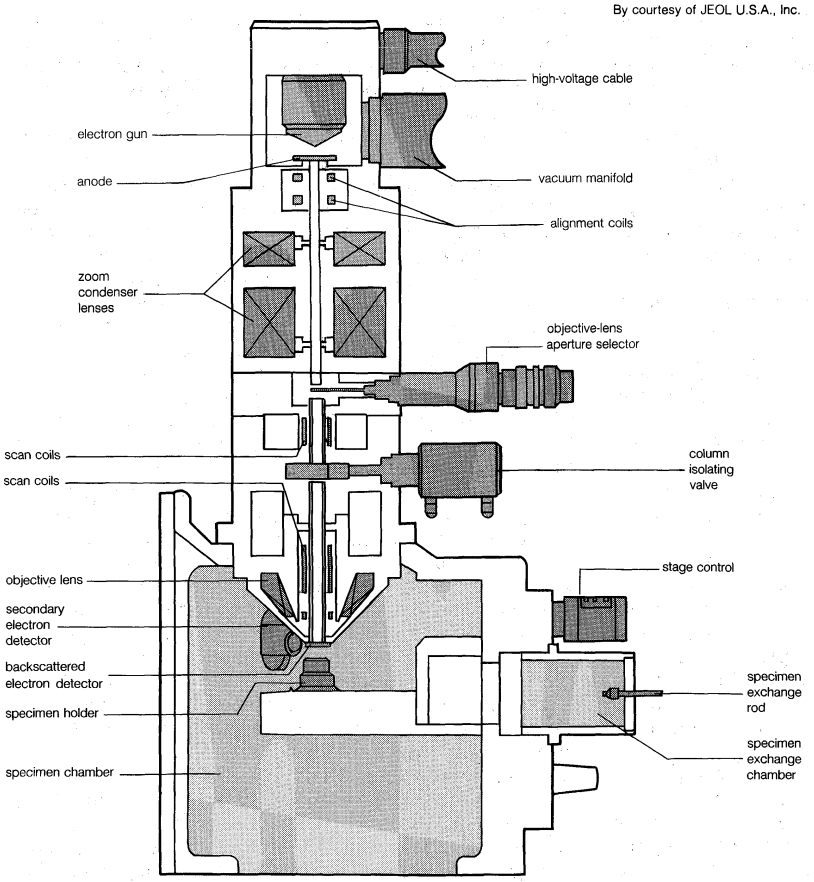


Figure 6: Components of a scanning electron microscope.

umn. Such a screen is satisfactory for quick observations and for focussing and aligning the instrument. A low-power optical microscope fitted outside the column allows the image on the screen to be inspected at a magnification of about 10 magnitudes, but the image suffers from poor contrast, lack of brightness, and the inherent grain of the screen. By allowing the electrons to fall on the emulsion of a photographic plate, the latent image can then be developed and fixed exactly as in ordinary photography, and prints at any degree of further enlargement may be obtained from the negative. Such further enlargement allows all the detail present in the image to be resolved by the eye and also allows some control over the contrast.

Image-intensification systems have been developed using some form of television tube, with the final image presented on a cathode-ray-tube monitor. Image intensifiers permit electronic amplification. Beam currents that would not register any image on a fluorescent screen may be used to permit examination of materials susceptible to radiation damage, such as certain types of plastics.

Special types of electron microscope. *High-voltage instruments.* Several different types of instrument have been constructed to operate at accelerating voltages in excess of the 100 kilovolts normally used in the conventional transmission instrument. Such high-voltage microscopes are capable of accelerating voltages of 1.5 million volts or more.

These instruments have several advantages over the conventional microscope: (1) The theoretical resolving power of the instrument increases as the accelerating voltage increases. (2) Thick specimens are easily penetrated by the beam. (3) Chromatic aberration caused by energy losses and scattering in the specimen is decreased. (4) Specimen heating and irradiation damage are decreased. (5) The resolution of electron diffraction patterns is improved.

The gain in resolution is theoretically about two and a half times that of a 100-kilovolt instrument. It is primarily for the increased penetrating power and consequent reduction in chromatic aberration that the high-voltage instruments are used. Recent advances have resulted in the development of high-voltage electron microscopes with near-atomic resolution. These instruments have enabled researchers to obtain detailed information about the atomic structure of materials (see Figure 7).

The operator must have adequate protection from the X rays that are produced when an electron beam impinges on objects in its path. Progress has been made toward solutions to these problems, and the high-voltage microscope is proving invaluable in the study of metals and alloys. It is possible that biological materials may also be examined if they are enclosed in microchambers inserted into the path of the beam. Bacteria have been examined in this way.

The scanning electron microscope. This type of electron microscope, designed for studying the surfaces of objects directly, utilizes a beam of focussed electrons as an electron probe that is scanned in a regular manner over the specimen. The action of the electron beam stimulates emission of secondary electrons from the surface of the specimen. These are collected and passed to a scintillator crystal (a crystal that produces tiny flashes of light when bombarded with electrons). The light flashes in the crystal operate a photomultiplier tube, which amplifies the electrical signal. The resultant amplified current is used, in turn, to modulate the brightness of a display cathode-ray (television) tube that has its line scan driven in synchronism with the probe beam in the microscope column. There is a point-by-point correspondence between the brightness of each point on the display tube and the number of secondary electrons emitted from any point on the surface of the specimen. In this way, a picture of the surface topography of the specimen is built up. The magnification of the image may be controlled by changing the ratio between the relative sizes of the area covered by the scanning probe and the picture on the display tube.

No elaborate specimen-preparation techniques are required for examination in the scanning electron microscope. Large and bulky specimens may be accommodated. It is desirable that the specimen be rendered electrically conducting, otherwise a sharp picture will not be obtained. Conductivity is usually achieved by evaporating a 10-to-20-nanometre-thick layer of metal onto the specimen in a vacuum (such a thickness does not materially affect the resolution of the surface details); or, under certain circumstances, an aerosol antistatic spray will achieve the same purpose.

The scanning electron microscope has the valuable feature that the image contrast may be changed by differences in potential across the surface under study. This facility is used widely in the semiconductor field and in studying microcircuit devices while they are operating. Scanning microscopes, rather than conventional transmission instruments, have been employed for transmission electron microscopy because they permit the use of very low magnification (of about 20 magnitudes). At the same time, large fields of view are obtained and very thick sections may be studied; this allows easy three-dimensional reconstructions to be made.

The electron-probe microanalyzer. A limitation of the conventional electron microscope is that it provides no chemical information. Electron-probe microanalyzers have

By courtesy of Fernando A. Ponce, Xerox Palo Alto Research Center

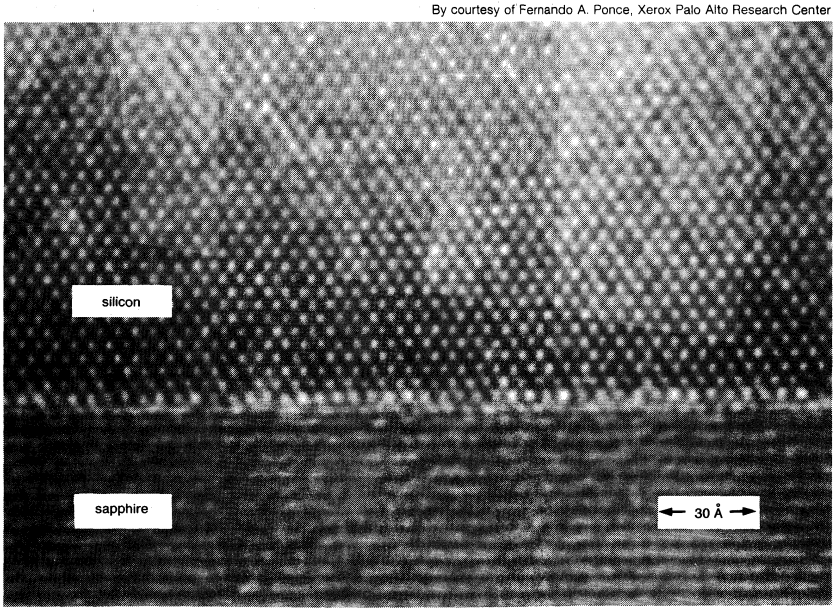


Figure 7: An electron micrograph of an interface between silicon and sapphire at near-atomic resolution. Advances in high-voltage transmission microscopy have provided scientists with detailed information about the atomic structure of materials.

been developed (since 1947) to carry out nondestructive elemental analysis at resolutions approaching those of the transmission electron microscope. This is done by measuring the intensity of the characteristic X rays emitted by a specimen when a focussed electron beam impinges on it. Early instruments used a static electron probe, but now scanning is the usual mode of operation, and a scanning X-ray image is built up on a cathode-ray tube. Such an instrument is called an X-ray microprobe analyzer.

X-ray microprobe analyzer

The X-ray microprobe analyzer is so valuable that X-ray spectrometer attachments are now provided for existing transmission electron microscopes. Combined instruments are specially designed with both transmission-electron-microscopy and X-ray microprobe analysis facilities.

The electron-probe microanalyzer has applications in mineralogical and metallurgical work and is also used in biological research.

Field-emission microscope. In this instrument a fine wire with a rounded tip is mounted in a cathode-ray tube. Electrons are drawn from the tip by a high electrical field and travel toward the screen on which the image is formed. Only heat-resistant metals, such as tungsten, platinum, and molybdenum, can be examined in this way, because the high current density at the metal tip has a great heating effect. The magnification is proportional to the ratio of the radius of curvature of the screen on which the image forms to the radius of the metal tip.

A variant of the field-emission microscope, called the field-ion microscope, is used in metallurgy for direct observation of defects in the crystalline lattice and for assessment of the effects of radiation damage and chemical reactions on the metal surface.

The proton-scattering microscope. This instrument uses a beam of protons from an ion source or gun. The beam is focussed and accelerated onto a crystal of a metal such as copper or tungsten, which serves as the target. The pattern of scattered protons is recorded on a fluorescent screen; the blocking patterns provide information on the orientation of the crystals.

PREPARATION OF SPECIMENS FOR OBSERVATION

Optical microscopy. The surface of bulk materials can be studied at relatively low powers with incident light by simply mounting the specimen for ease of handling. Often, more information may be gained if the surface is flattened by grinding and polishing. In the case of metals, it may be necessary to etch the specimen in acids to allow the grain boundaries to be easily differentiated. With smaller amounts of material, such as powders and dusts, the sample is usually mounted on a slide in a resin medium of known refractive index and covered with a coverslip. For examination by incident light, the specimen material may be simply strewn on a slide.

Biological material, both plant and animal, generally requires more preparation. Organisms must be reduced to pieces that are small enough for study under the microscope. If unicellular organisms or small invertebrates are to be studied alive, they are mounted in a suitable fluid medium. Permanent slide preparations may be made of material that is first killed and then treated with chemical reagents that react with the protoplasm and prevent any further change in it. Contrast is obtained by soaking the specimens in dye solutions. Finally, the water in the objects may be removed by treatment with a series of alcohols of gradually increasing strength before the preparation is cleared in an organic fluid and mounted in a synthetic resin. This sets hard, fixing the coverslip firmly onto the preparation. When large masses of tissue are studied, the individual cells are separated from one another by the process of maceration (softening by soaking), or, more commonly, the object is embedded in a supporting medium, such as paraffin wax, which is then cut into very thin slices. The resulting sections are attached to glass slides, the embedding material dissolved away, and the section stained in various ways to obtain the required contrast. Dehydration, clearing, and mounting in resin complete the preparation of the slide. In special cases, the tissue is made rigid enough for sectioning by freezing it solid.

Maceration

Rocks and other hard specimens may be cut into thin sections for microscopic examination by using a diamond saw or impregnated wheel. The section is then reduced to its final thickness by grinding it down with an abrasive powder.

Electron microscopy. For transmission electron microscopy, the specimen must be in very thin slices, so that the electron beam may penetrate it. Biological material is fixed and dehydrated essentially as for optical microscopy but is embedded in methacrylate or epoxy resin, which sets much harder than the wax and nitrocellulose used in optical microtechnique and allows thinner sections to be cut. The machines used for slicing (ultramicrotomes) usually operate with a knife broken from plate glass and use thermal expansion of the support to move the block forward in the small increments required for each successive section. The thin sections, as they come off the knife, are floated onto the surface of a liquid (usually distilled water) and then picked up on a copper grid; after drying, this grid, bearing the section, may be inserted directly into the microscope.

Contrast in material of biological origin is usually introduced at or just after the fixation stage by the use of osmium tetroxide, which is believed to change chemically after its introduction into the material to a substance (perhaps a lower oxide of osmium or the metal itself) that then provides the required electron density. Further increase in the contrast may be obtained by so-called electron staining of the finished sections when they are on the grids. This is done by floating the grids onto solutions of lead or uranyl salts for short periods of time; after washing and drying, the sections may be examined in the microscope.

Electron staining

For the examination of surfaces in the transmission microscope, it is necessary to take a replica of the surface in plastic, coat the replica with a film of evaporated carbon for strength, and finally evaporate a thin layer of metal onto the surface of the replica at a low angle in order to introduce contrast. Particulate material may be supported on a thin film of plastic carried on a copper grid. Contrast may be added to the biological material by shadowing it in a vacuum with heavy metals at a low angle. This shows up the surface detail just as the setting sun shining along a road shows up the irregularities in the road surface in clear relief. Alternatively, a negative staining technique may be used. In this, the specimens are surrounded by a matrix of the salt of some metal that is opaque to the electron beam. When examined in the microscope, a negative image results with the shape of the particle outlined in sharp relief. This technique is used in studying viruses and in molecular biology.

Metals may be prepared for the electron microscope by obtaining them as foils and then thinning them down further electrolytically to the point at which they will transmit the electron beam. Plastics and rubber may now be sectioned for examination in the electron microscope by freezing and then slicing.

Specimens for study in the scanning electron microscope simply need to be mounted on a specimen-stub holder and rendered electrically conductive either by an antistatic spray or by evaporating onto them a very thin layer of metal. (S.Br./Ed.)

Telescopes

The telescope is undoubtedly the most important investigative tool in astronomy. It provides a means of collecting and analyzing radiation from celestial objects, even those in the far reaches of the universe.

Galileo revolutionized astronomy when he applied the telescope to the study of extraterrestrial bodies in the early 17th century. Until then, magnification instruments had never been used for this purpose. Since Galileo's pioneering work, increasingly more powerful optical telescopes have been developed, as has a wide array of instruments capable of detecting and measuring invisible forms of radiation, such as radio, X-ray, and gamma-ray telescopes. Observational capability has been further enhanced by the invention of various kinds of auxiliary instruments (*e.g.*, the camera, spectrograph, and charge-coupled device) and by the use of electronic computers, rockets, and

spacecraft in conjunction with telescope systems. These developments have contributed dramatically to advances in scientific knowledge about the solar system, the Milky Way Galaxy, and the universe as a whole.

GENERAL CONSIDERATIONS

Observations at all wavelengths of the electromagnetic spectrum

Today, the telescope is used to explore every region of the electromagnetic spectrum from the shortest wavelengths (gamma rays) to the longest (radio waves; see Figure 8). The wavelengths of the spectrum are measured in three different units: angstroms (Å), micrometres (μ), and metres (m). Each of these units is customarily used for specific wavelength ranges, as shown in the figure. For example, the wavelengths for gamma rays and X rays are given in angstroms, those for infrared rays in micrometres, and those for intermediate radio waves in metres. (Centimetres are often used for short radio waves [microwaves] and kilometres for long radio waves.)

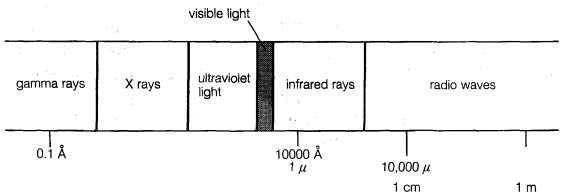


Figure 8: The electromagnetic spectrum.

Astronomical observations were restricted to visible wavelengths until the 1930s when Karl G. Jansky and Grote Reber of the United States opened the radio "window." Since the 1960s the use of Earth-orbiting telescope systems has enabled astronomers to make observations in all other spectral regions as well.

OPTICAL TELESCOPES

Refracting telescopes. Commonly known as refractors, telescopes of this kind are used to examine the visible-light region of the electromagnetic spectrum. Typical uses include viewing the Moon, other objects of the solar system such as Jupiter and Mars, and double stars. The name refractor is derived from the term refraction, which is the bending of light when it passes from one medium to another of different density—*e.g.*, from air to glass. The glass is referred to as a lens and may have one or more components. The physical shape of the components may be convex, concave, or plane-parallel. Figure 9 illustrates the principle of refraction and the term focal length. The focus is the point, or plane, at which light rays from infinity converge after passing through a lens and traveling a distance of one focal length. In a refractor, the first lens through which light from a celestial object passes is called the objective lens. It should be noted that the light will be inverted at the focal plane. A second lens, referred to as the eyepiece lens, is placed behind the focal plane and enables the observer to view the enlarged, or magnified, image. Thus, the simplest form of refractor consists of an objective and an eyepiece, as illustrated in Figure 10.

Aperture

The diameter of the objective is referred to as the aperture; it typically ranges from a few centimetres for small spotting telescopes up to one metre for the largest refractor in existence. The objective, as well as the eyepiece,

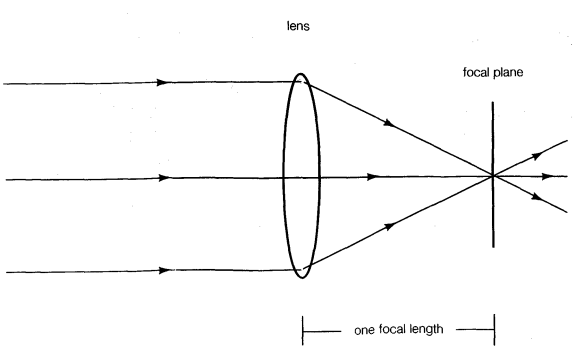


Figure 9: Focal length of a lens.

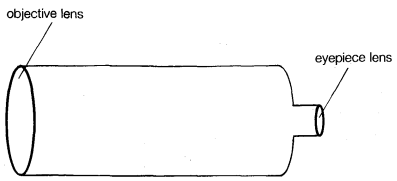


Figure 10: Refracting telescope.

may have several components. Small spotting telescopes may contain an extra lens behind the eyepiece to erect the image so that it does not appear upside-down. When an object is viewed with a refractor, the image may not appear sharply defined or it may even have a predominant colour in it. Such distortions, or aberrations, are sometimes introduced when the lens is polished into its design shape. The major kind of distortion in a refractor is chromatic aberration, which is the failure of the different coloured light rays to come to a common focus. Chromatic aberration can be minimized by adding additional components to the objective. In lens-design technology, the coefficients of expansion of different kinds of glass are carefully matched to minimize the aberrations that result from temperature changes of the telescope at night.

Eyepieces, which are used with both refractors and reflectors (see below *Reflecting telescopes*), have a wide variety of applications and provide observers with the ability to select the magnification of their instruments. The magnification, sometimes referred to as magnifying power, is determined by dividing the focal length of the objective by the focal length of the eyepiece. For example, if the objective has a focal length of 254 centimetres (100 inches) and the eyepiece has a focal length of 2.54 centimetres (1 inch), then the magnification will be 100. Large magnifications are very useful for observing the Moon and the planets; however, since stars appear as point sources due to their great distances, magnification provides no additional advantage when viewing them. Another important factor that one must take into consideration when attempting to view at high magnification is the stability of the telescope mounting. Any vibration in the mounting also will be magnified and may severely reduce the quality of the observed image. Thus, great care is usually taken to provide a stable platform for the telescope. This problem should not be associated with that of atmospheric seeing, which may introduce a disturbance to the image due to fluctuating air currents in the path of the light from a celestial or terrestrial object. Generally, most of the seeing disturbance arises in the first 30 metres of air above the telescope. Large telescopes are frequently installed on mountain peaks in order to get above the seeing disturbances.

The most important of all the powers of an optical telescope is its light-gathering power. This capacity is strictly a function of the diameter of the clear objective—that is, the aperture—of the telescope. Comparisons of different-sized apertures for their light-gathering power are calculated by the ratio of their diameters squared; for example, a 25-centimetre objective will collect four times the light of a 12.5-centimetre objective $[(25 \times 25) \div (12.5 \times 12.5)] = 4$. The advantage of collecting more light with a larger-aperture telescope is that one can observe fainter stars, nebulas, and very distant galaxies.

Resolving power is another important feature of a telescope. This is the ability of the instrument to distinguish clearly between two points whose angular separation is less than the smallest angle that the observer's eye can resolve. The resolving power of a telescope can be calculated by the formula

$$\text{resolving power} = 11.25 \text{ seconds of arc}/d,$$

where d is the diameter of the objective expressed in centimetres. Thus, a 25-centimetre-diameter objective has a theoretical resolution of 0.45 second of arc and a 250-centimetre telescope has one of 0.045 second of arc. An important application of resolving power is in the observation of visual binary stars. Here, one star is routinely observed as it revolves around a second star. Many observatories conduct extensive visual binary observing pro-

grams and publish catalogs of their observational results. One of the major contributors in this field is the United States Naval Observatory in Washington, D.C.

Most refractors currently in use at observatories have equatorial mountings. (The mounting describes the orientation of the physical bearings and structure that permits a telescope to be pointed at a celestial object for viewing.) In the equatorial mounting, the polar axis of the telescope is constructed parallel to the Earth's axis. The polar axis supports the declination axis of the instrument. Declination is measured on the celestial sky north or south from the celestial equator. The declination axis makes it possible for the telescope to be pointed at various declination angles as the instrument is rotated about the polar axis with respect to right ascension. Right ascension is measured along the celestial equator from the vernal equinox (*i.e.*, the position on the celestial sphere where the Sun crosses the celestial equator from south to north on the first day of spring). Declination and right ascension are the two coordinates that define a celestial object on the celestial sphere. Declination is analogous to latitude and right ascension is analogous to longitude. Graduated dials are mounted on the axis to permit the observer to point the telescope precisely. To track an object, the telescope's polar axis is driven smoothly by an electric motor at a sidereal rate—namely, at a rate equal to the rate of rotation of the Earth with respect to the stars. Thus, one can track or observe with a telescope for long periods of time if the sidereal rate of the motor is very accurate. High-accuracy, motor-driven systems have become readily available with the rapid advancement of quartz-clock technology. Most major observatories now rely on either quartz or atomic clocks to provide accurate sidereal time for observations as well as to drive telescopes at an extremely uniform rate.

A notable example of a refracting telescope is the 66-centimetre refractor of the U.S. Naval Observatory (see Figure 11). This instrument was used by the astronomer Asaph Hall to discover the two moons of Mars, Phobos and Deimos, in 1877. Today, the telescope is used primarily for observing double stars. The 91-centimetre refractor at Lick Observatory on Mount Hamilton, Calif., U.S., and the one-metre instrument at Yerkes Observatory in Williams Bay, Wis., U.S., are the largest refracting systems currently in operation (Table 3).

Official U.S. Navy photograph

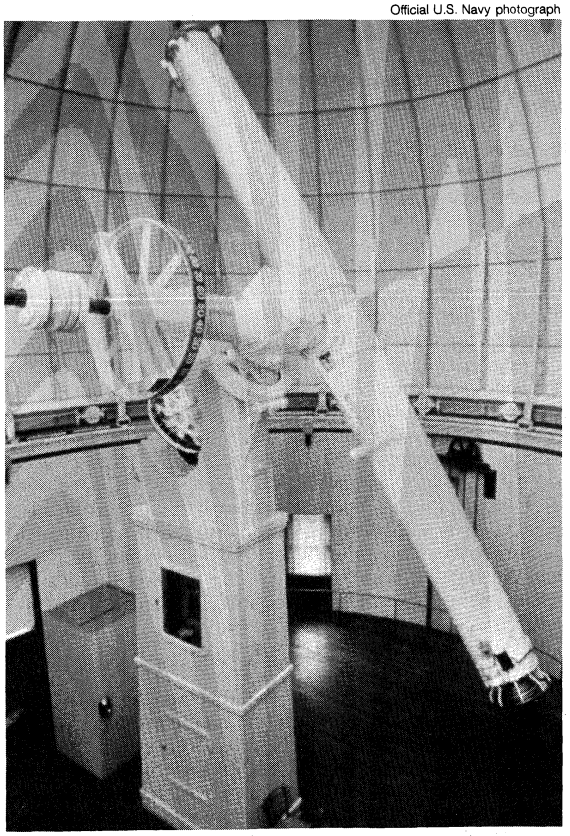


Figure 11: The 65-centimetre refractor of the U.S. Naval Observatory.

Another type of refracting telescope is the astrograph, which usually has an objective diameter of approximately 20 centimetres. The astrograph has a photographic plateholder mounted in the focal plane of the objective so that photographs of the celestial sphere can be taken. The photographs are usually taken on glass plates. The principal application of the astrograph is to determine the positions of a large number of faint stars. These positions are then published in catalogs such as the *AGK3* and serve as reference points for deep-space photogaphy. Astrographs

Reflecting telescopes. Reflectors are used to examine not only the visible region of the electromagnetic spectrum but also to explore both the shorter- and longer-wavelength regions adjacent to it (*i.e.*, the ultraviolet and the infrared). The name of this type of instrument is derived from the fact that the primary mirror reflects the light back to a focus instead of refracting it. The primary mirror usually has a concave spherical or parabolic shape and, as it reflects the light, it inverts the image at the focal plane. Figure 12 illustrates the principle of a concave reflecting mirror. The formulas for resolving power, magnifying power, and light-gathering power, as discussed for refractors, apply to reflectors as well.

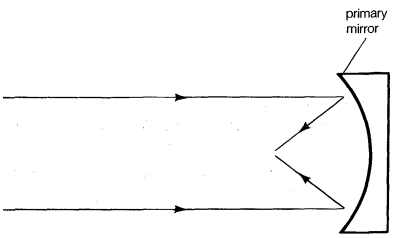


Figure 12: Concave mirror.

The primary mirror is located at the lower end of the telescope tube in a reflector and has its front surface coated with an extremely thin film of metal, such as aluminum. The back of the mirror is usually made of glass, although other materials have been used from time to time. Pyrex (trademark) was the principal glass of choice for many of the older large telescopes, but new technology has led to the development and widespread use of a number of glasses with very low coefficients of expansion. A low coefficient of expansion means that the shape of the mirror will not change significantly as the temperature of the telescope changes during the night. Since the back of the mirror serves only to provide the desired form and physical support, it does not have to meet the high optical quality standards required for a lens.

Reflecting telescopes have a number of other advantages over refractors. They are not subject to chromatic aberration because reflected light does not disperse according to wavelength. Also, the telescope tube of a reflector is shorter than that of a refractor of the same diameter, which reduces the cost of the tube. Consequently, the dome for housing a reflector is smaller and more economical to construct. So far only the primary mirror for the reflector has been discussed. In examining Figure 12, one might wonder about the location of the eyepiece. The primary mirror reflects the light of the celestial object to the prime focus near the upper end of the tube. Obviously, if an observer put his eye there to observe with a modest-sized reflector, he would block out the light from the primary mirror with his head. Isaac Newton placed a small plane mirror at an angle of 45° inside the prime focus and thereby brought the focus to the side of the telescope tube. The amount of light lost by this procedure is very small when compared to the total light-gathering power of the primary mirror. The Newtonian reflector, shown in Figure 13, is popular among amateur telescope makers.

A contemporary of Newton, N. Cassegrain of France, invented another type of reflector. Called the Cassegrainian telescope, this instrument employs a small convex mirror to reflect the light back through a small hole in the primary mirror to a focus located behind the primary. Figure 14 illustrates a typical Cassegrain reflector. Some large Cassegrainian telescope

telescopes of this kind do not have a hole in the primary mirror but use a small plane mirror in front of the primary to reflect the light outside the main tube and provide another place for observation. The Cassegrain design usually permits short tubes relative to their mirror diameter.

One more variety of reflector was invented by another of Newton's contemporaries, the Scottish astronomer James Gregory. Gregory placed a concave secondary mirror outside the prime focus to reflect the light back through a hole in the primary mirror. Notable is the fact that the Gregorian design was adopted for the Earth-orbiting space observatory, the Solar Maximum Mission (SMM), launched in 1980.

Most large reflecting telescopes currently in use have a cage at their prime focus that permits the observer to sit inside the telescope tube while operating the instrument. The five-metre reflector at Palomar Observatory, near San Diego, Calif., is so equipped. While most reflectors have equatorial mounts similar to refractors, the world's largest reflector, the six-metre instrument at the Special Astrophysical Observatory in Zelenchukskaya, Russian S.F.S.R., has an altitude-azimuth mounting. The significance of the latter design is that the telescope must be moved in both altitude and azimuth as it tracks a celestial object. Equatorial mountings, by contrast, require motion in only one coordinate while tracking since the declination coordinate is constant. Reflectors, like refractors, usually have small guide telescopes mounted parallel to their main optical axis to facilitate locating the desired object. These guide telescopes have low magnification and a wide field of view, the latter being a desirable attribute for finding stars or other remote cosmic objects.

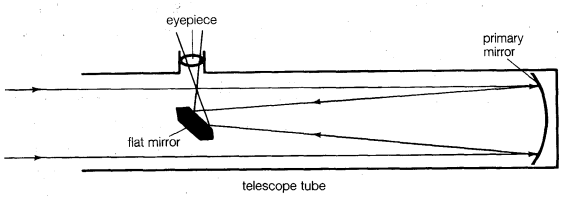


Figure 13: The Newtonian reflector.

The parabolic shape of a primary mirror has a basic failing in that it produces a narrow field of view. This can be a problem when one wishes to observe extended celestial objects. To overcome this difficulty, most large reflectors now have a modified Cassegrain design. The central area of the primary mirror has its shape deepened from that of a paraboloid, and the secondary mirror is configured to compensate for the altered primary. The result is the Ritchey-Chrétien design, which has a curved rather than a flat focus. Obviously, the photographic medium must be curved to collect high-quality images across the curved focal plane. The one-metre telescope of the U.S. Naval Observatory in Flagstaff, Ariz., was one of the early examples of this design. The 1.54-metre astrometric reflector of the Naval Observatory is shown in Figure 15.

Ritchey-Chrétien design

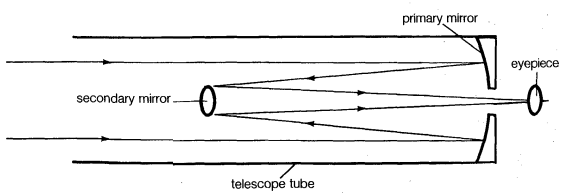


Figure 14: The Cassegrain reflector.

The Schmidt telescope. The above-mentioned Ritchey-Chrétien design has a good field of view of about 1°. For some astronomical applications, however, photographing larger areas of the sky is mandatory. In 1930 Bernhard Schmidt, an optician at the Hamburg Observatory in Bergedorf, Ger., designed a catadioptric telescope that satisfied the requirement of photographing larger celestial areas. A catadioptric telescope design incorporates the best features of both the refractor and reflector—*i.e.*, it has both reflective and refractive optics. The Schmidt telescope has a spherically shaped primary mirror. Since parallel light rays that are reflected by the centre of a spherical mirror

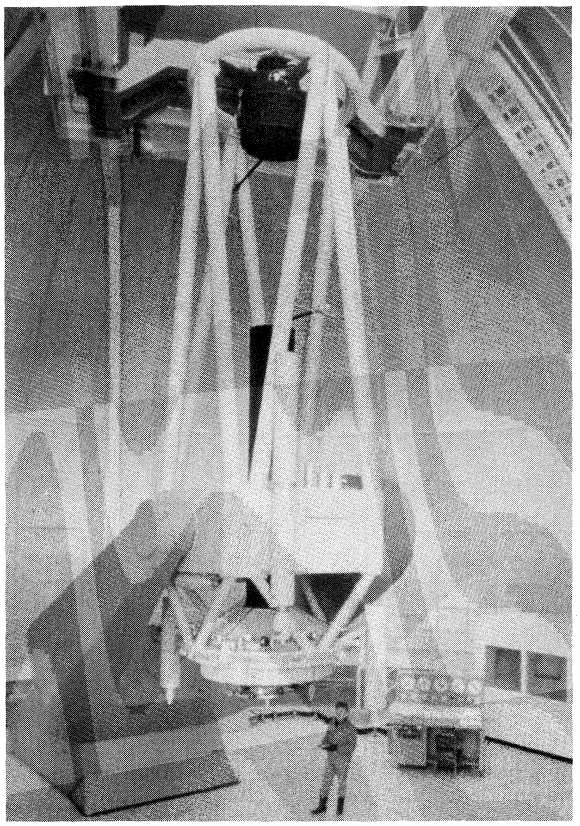


Figure 15: The astrometric reflector of the U.S. Naval Observatory.
Official U.S. Navy photograph

are focussed farther away than those reflected from the outer regions, Schmidt introduced a thin lens (called the correcting plate) at the radius of curvature of the primary mirror. Since this correcting plate is very thin, it introduces little chromatic aberration. The resulting focal plane has a field of view several degrees in diameter. Figure 16 illustrates a typical Schmidt design.

The *National Geographic Society–Palomar Observatory Sky Survey* made use of a 1.2-metre Schmidt telescope (Figure 17) to photograph the northern sky in both the red and blue regions of the visible spectrum. The survey produced 900 pairs of photographic plates (about 7° by 7° each) taken between 1949 and 1956. Schmidt telescopes of the European Southern Observatory in Chile and of the observatory at Siding Spring Mountain in Australia have photographed the remaining part of the celestial sky that cannot be observed from Palomar Mountain. (The survey undertaken at the latter included photographs in the infrared as well as in the red and blue regions of the spectrum.)

Multimirror telescopes. The main reason astronomers build larger telescopes is to increase light-gathering power so that they can see deeper into the universe. Unfortunately, the cost of constructing larger single-mirror telescopes increases rapidly—approximately with the cube of the diameter of the aperture. Thus, in order to achieve the goal of increasing light-gathering power while keeping

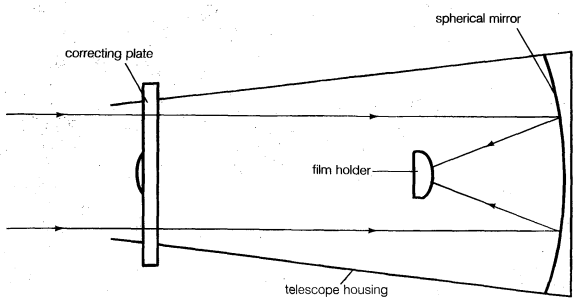


Figure 16: The Schmidt telescope.

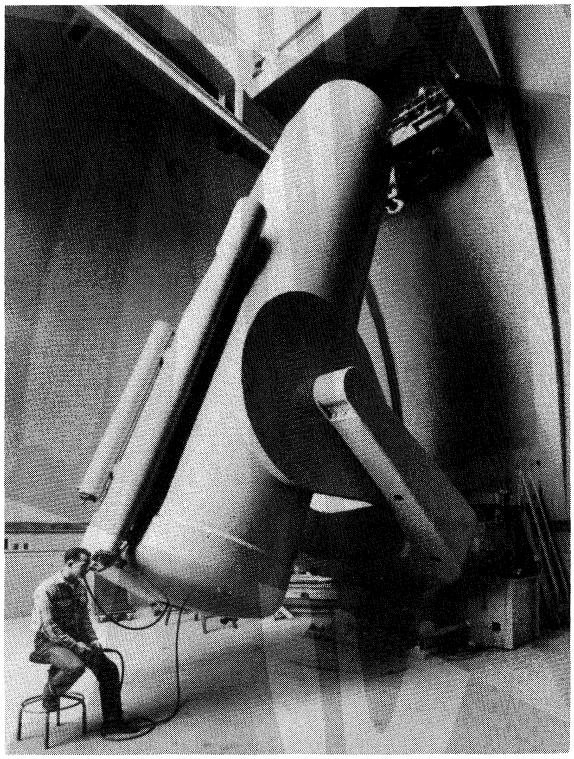


Figure 17: The 1.2-metre Schmidt telescope of the Palomar Observatory near San Diego, Calif.
Palomar Observatory photograph

costs down, it has become necessary to explore new, more economical and nontraditional telescope designs.

The Multiple Mirror Telescope

The American-built Multiple Mirror Telescope (MMT), located at the Whipple Observatory in Arizona (Figure 18), represents such an effort. The MMT has six 1.8-metre paraboloid mirrors mounted in a single framework; the light from all the mirrors is concentrated into a single focus. The mirrors, under computer control, are automatically aligned at regular intervals during an observing tour. A 10-metre multimirror telescope is expected to be installed on Mauna Kea on the island of Hawaii by the early 1990s. When completed, this instrument, called the Keck Telescope, will comprise 36 contiguous, adjustable mirror segments, all under computer control. Even larger multimirror instruments are currently being planned by American, western European, and Soviet astronomers.

Special types of optical telescopes. *Solar telescopes.* Either a refractor or reflector may be used for visual observations of solar features, such as sunspots or solar prominences. The main precaution that needs to be taken is to reduce the intensity of the image so that the observer's eye will not be damaged. Generally, this is done with a tinted eyepiece. Special solar telescopes have been constructed, however, for investigations of the Sun that require the use of such ancillary instruments as spectroheliographs and coronagraphs. These telescopes are mounted in towers and have very long focus objectives. Typical examples of tower solar telescopes are found at the Mount Wilson Observatory in California and the McMath-Hulbert Observatory in Michigan. The long focus objective produces a very good scale factor, which in turn makes it possible to look at individual wavelengths of the solar electromagnetic spectrum in great detail. A tower telescope has an equatorially mounted plane mirror at its summit to direct the sunlight into the telescope objective. This plane mirror is called a coelostat. Bernard Lyot constructed another type of solar telescope in 1930 at Pic du Midi Observatory in France. This instrument was specifically designed for photographing the Sun's corona (the outer layer), which up to that time had only been successfully photographed during solar eclipses. The coronagraph, as this special telescope is called, must be located at a high altitude to be effective. The high altitude is required to reduce the scattered sunlight, which would reduce the quality of the photograph. The High Altitude Observatory in Colorado and the Sacramento Peak Observatory in New Mexico have coronagraphs.

Coronagraph

Earth-orbiting space telescopes. While astronomers continue to seek new technological breakthroughs with which to build larger ground-based telescopes, it is readily apparent that the only solution to some scientific problems is to make observations from above the Earth's atmosphere. A series of Orbiting Astronomical Observatories (OAOs) has been launched by the National Aeronautics and Space Administration (NASA); the OAO launched in 1972 (later named Copernicus) had an 81-centimetre telescope on board. The most significant observational system so far placed in Earth orbit is, however, the Hubble Space

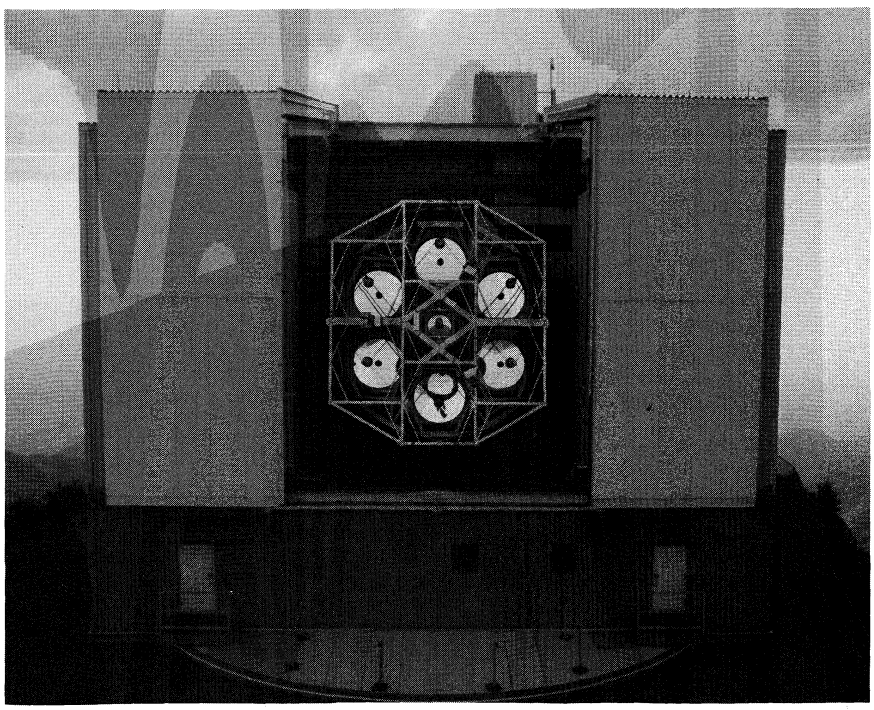


Figure 18: The Multiple Mirror Telescope of the Whipple Observatory, Mount Hopkins, Ariz.

Table 3: Some Important Ground-Based Optical Telescopes

name	aperture and type	observatory and location
Bolshoi Teleskop	6-metre reflector	Special Astrophysical Observatory, Zelenchukskaya, Russian S.F.S.R.
George Ellery Hale Telescope	5-metre reflector	Palomar Observatory, Palomar Mountain, Calif., U.S.
William Herschel Telescope	4.2-metre reflector	Roque de los Muchachos Observatory, La Palma, Canary Islands, Spain
	4-metre reflector	Cerro Tololo Inter-American Observatory, Cerro Tololo, Chile
Anglo-Australian Telescope	3.9-metre reflector	Anglo-Australian Observatory, Siding Spring Mountain, N.S.W., Australia
Nicholas U. Mayall Telescope	3.8-metre reflector	Kitt Peak National Observatory, Kitt Peak, Ariz., U.S.
Canada-France-Hawaii Telescope	3.6-metre reflector	Canada-France-Hawaii Telescope Corporation, Mauna Kea, Hawaii, U.S.
	3.6-metre reflector	European Southern Observatory, La Silla Hill, Chile
Multiple Mirror Telescope	six 1.8-metre paraboloid mirrors mounted in a single framework	Whipple Observatory, Mount Hopkins, Ariz.
	1.2-metre Schmidt	Palomar Observatory, Palomar Mountain, Calif.
	1-metre refractor	Yerkes Observatory, Williams Bay, Wis., U.S.
	91-centimetre refractor	Lick Observatory, Mount Hamilton, Calif.
	83-centimetre refractor	Meudon Observatory, Meudon, France

Telescope (HST; see Figure 19). Launched in 1990, the HST is essentially an optical-ultraviolet telescope with a 2.4-metre primary mirror. It has been designed to enable astronomers to see into a volume of space 300 to 400 times larger than that permitted by other systems. At the same time, the HST is not impeded by any of the problems caused by the atmosphere. It is equipped with five principal scientific instruments: (1) a wide-field and planetary camera, (2) a faint-object spectrograph, (3) a high-resolution spectrograph, (4) a high-speed photometer, and (5) a faint-object camera. The HST was launched into orbit from the U.S. Space Shuttle at an altitude of more than 570 kilometres above the Earth. The European Space Agency (ESA) provided some of the scientific instrumentation for the telescope. The increased sensitivity and resolution expected from the HST should enhance scientific knowledge of the universe substantially.

Astronomical transit instruments. These small but extremely important telescopes play a vital role in mapping the celestial sphere. Without the transit instrument's very accurate determination of stellar and planetary positions, the larger deep-space telescopes would not be able to find their desired celestial object.

Astronomical transit instruments are usually refractors with apertures of 15 to 20 centimetres. (Ole Rømer, a Danish astronomer, is credited with having invented this type of telescope system.) The main optical axis of the instrument is aligned on a north–south line such that its motion is restricted to the plane of the meridian of the observer. The observer's meridian is a great circle on the celestial sphere that passes through the north and south points of the horizon as well as through the zenith of the observer. Restricting the telescope to motion only in the meridian provides an added degree of stability, but

By courtesy of the Hughes Aircraft Company

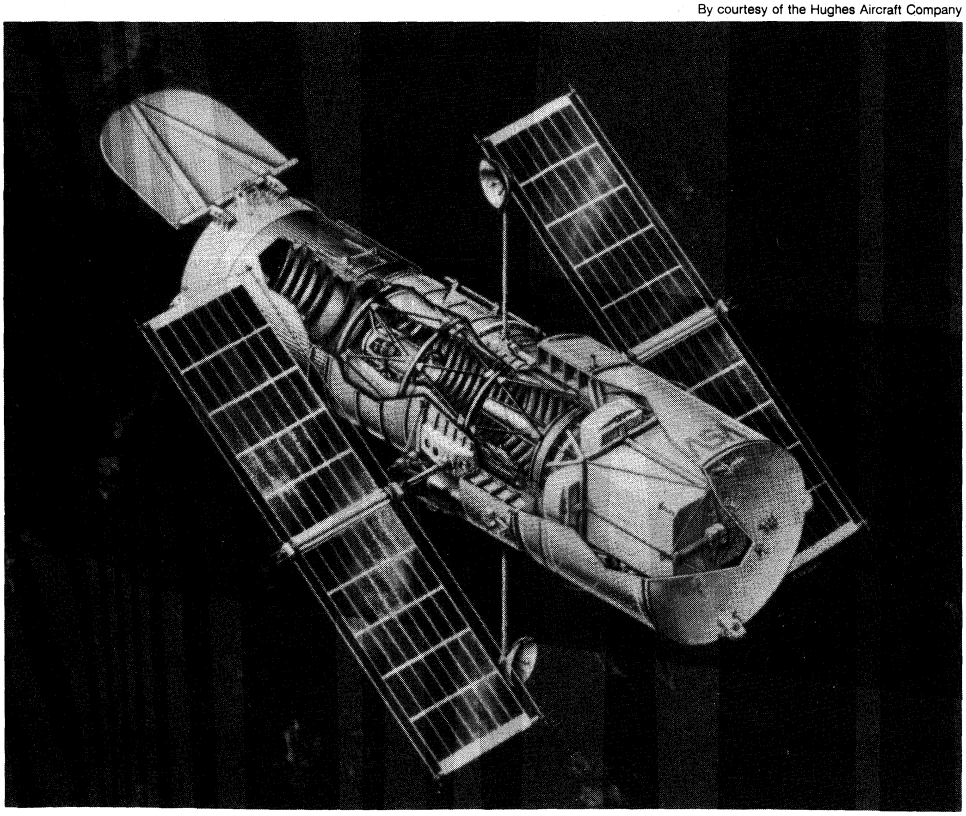


Figure 19: Cutaway of the Hubble Space Telescope, revealing the Optical Telescope Assembly (OTA), the heart of this orbiting observational system.

Principal types of transit telescopes

it requires the observer to wait for the celestial object to rotate across his meridian. The latter process is referred to as transiting the meridian, from which the name of the telescope is derived. There are various types of transit instruments, as, for example, the transit circle telescope, the vertical circle telescope, and the horizontal meridian circle telescope. The transit circle telescope determines the right ascension of celestial objects, while the vertical circle instrument measures only the declinations of such objects. Transit circles and horizontal meridian circles measure both right ascension and declination at the same time. The final output data of all transit instruments are included in star or planetary catalogs.

One of the most accurate astronomical transit instruments in the world is the U.S. Naval Observatory's 15-centimetre transit circle telescope (see Figure 20). Other notable examples of this class of telescopes include the transit circle of the National Astronomical Observatory in Tokyo, the meridian circle of the Bordeaux Observatory, and the automatic meridian circle of the Roque de los Muchachos Observatory in the Canary Islands.

Official U.S. Navy photograph

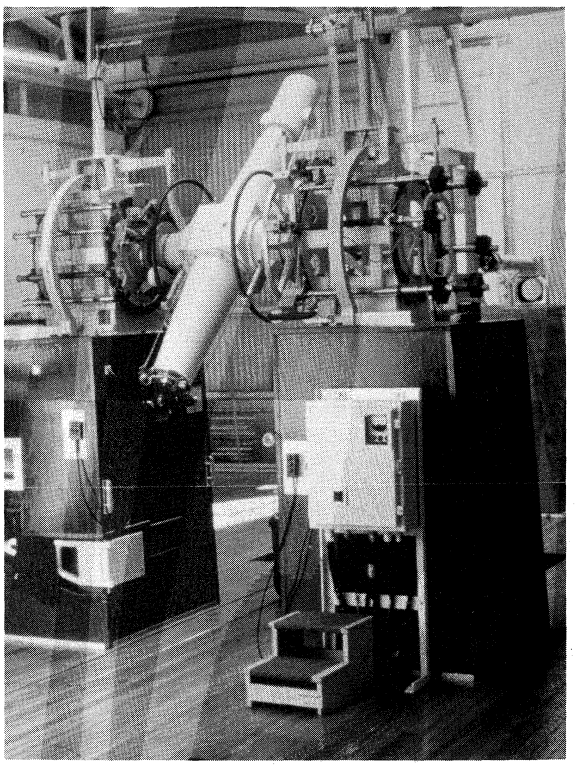


Figure 20: The 15-centimetre transit circle instrument of the U.S. Naval Observatory.

Astrolabes. Another special type of telescopic instrument is the modern version of the astrolabe. Known as a prismatic astrolabe, it too is used for making precise determinations of the positions of stars and planets. It may sometimes be used inversely to determine the latitude and longitude of the observer, assuming the star positions are accurately known. The aperture of a prismatic astrolabe is small, usually only 8 to 10 centimetres. A small pool of mercury and a refracting prism make up the other principal parts of the instrument. An image reflected off the mercury is observed along with a direct image to give the necessary position data. The most notable example of this type of instrument is the French-constructed Danjon astrolabe. During the 1970s, however, the Chinese introduced various innovations that resulted in a more accurate and automatic kind of astrolabe, which is now in use at the Beijing Observatory. (B.L.Kl.)

RADIO TELESCOPES

Radio telescopes are used to study naturally occurring radio emissions from stars, galaxies, quasars, and other astronomical objects between wavelengths of about 10 metres (30 megahertz [MHz]) and 1 millimetre (300 gigahertz [GHz]). At wavelengths longer than about 20 centimetres (1.5 GHz), irregularities in the ionosphere distort the incoming signals. This causes a phenomenon known as scintillation, which is analogous to the twinkling of stars seen at optical wavelengths. The absorption of cosmic radio waves by the ionosphere becomes more and more important as wavelength increases. The ionosphere becomes opaque to incoming signals of wavelengths longer than about 10 metres. Radio observations of the universe at these wavelengths are difficult from ground-based radio telescopes. Below wavelengths of a few centimetres, absorption in the atmosphere becomes increasingly critical. At wavelengths shorter than 1 centimetre (30 GHz), observations from the ground are possible only in a few specific wavelength bands that are relatively free of atmospheric absorption.

Principles of operation. Radio telescopes vary widely, but they all have two basic components: (1) a large radio antenna and (2) a radiometer or radio receiver. The sensitivity of a radio telescope—*i.e.,* the ability to measure weak sources of radio emission—depends on the area and efficiency of the antenna, the sensitivity of the radio receiver used to amplify and detect the signals, and the duration of the observation. For broadband continuum emission the sensitivity also depends on the receiver bandwidth. Because some astronomical radio sources are extremely weak, radio telescopes are usually very large and only the most sensitive radio receivers are used. Moreover, weak cosmic signals can be easily masked by terrestrial radio interference, and great effort is taken to protect radio telescopes from man-made interference.

Power gain

The power gain of a radio antenna is usually measured in terms of the improved sensitivity over a simple dipole antenna. For any given antenna the gain varies with the surface area of the antenna and inversely with the square of the wavelength of operation. In a high-gain antenna, radiation is accepted only from a very narrow beam in the sky; the width of the beam depends on the ratio of the wavelength of operation to the diameter of the antenna.

The most familiar type of radio telescope is the radio reflector consisting of a parabolic antenna—the so-called dish—which operates in the same manner as a television-satellite receiving antenna to focus the incoming radiation onto a small antenna referred to as the feed, a term that first originated with antennas used for radar transmissions. In a radio telescope the feed is typically a waveguide horn and is connected to a sensitive radio receiver. Cryogenically cooled solid-state amplifiers with very low internal noise are used to obtain the best possible sensitivity.

In some radio telescopes the parabolic surface is equatorially mounted, with one axis parallel to the rotation axis of the Earth. Equatorial mounts are attractive because they allow the telescope to follow a position in the sky as the Earth rotates by compensating motion about a single axis, but they are difficult and expensive to build. In most cases, a digital computer is used to drive the telescope about the azimuth and elevation axes to follow the motion of a radio source across the sky.

Observing times up to many hours are expended and sophisticated signal-processing techniques are used to detect astronomical radio signals that are as much as one million times weaker than the noise generated in the receiver. Signal-processing and analysis are usually done in a digital computer. Although some of the computations may be carried out by microcomputers (*i.e.,* those of the personal-computer class), other tasks require large, high-speed machines to translate the raw data into a form useful to the astronomer.

In the simplest form of radio telescope, the receiver is placed directly at the focal point of the parabolic reflector, and the detected signal is carried by cable along the feed support structure to a point near the ground where it can be recorded and analyzed. However, it is difficult in this type of system to access the instrumentation for maintenance and repair, and weight restrictions limit the size and number of individual receivers that can be installed on the telescope. More often, a secondary hyperbolic reflector is placed near the focal point of the paraboloid to focus the

radiation to a point near the vertex of the main reflector where multiple receivers may be more readily accommodated with less stringent weight restrictions and access is more straightforward. Secondary focus systems also have the advantage that both the primary and secondary reflecting surfaces may be carefully shaped so as to improve the gain over that of a simple parabolic antenna.

Performance-limiting factors

The performance of a radio telescope is limited by various factors: the accuracy of a reflecting surface that may depart from the ideal shape because of manufacturing irregularities; the effect of wind load; thermal deformations that cause differential expansion and contraction; and deflections due to changes in gravitational forces as the antenna is pointed to different parts of the sky. Departures from a perfect parabolic surface become important when they are a few percent or more of the wavelength of operation. Since small structures can be built with greater precision than larger ones, radio telescopes designed for operation at millimetre wavelength are typically only a few tens of metres across, whereas those designed for operation at centimetre wavelengths range up to 100 metres in diameter.

Traditionally, the effect of gravity has been minimized by designing the movable structure to be as stiff as possible in order to reduce the deflections resulting from gravity. A more effective technique, based on the principle of homology, allows the structure to deform under the force of gravity, but, if the cross section and weight of each member of the movable structure are properly chosen, the gravitational forces will deform the reflecting structure into a new paraboloid with a slightly different focal point. It is then necessary only to move the feed or secondary reflector to maintain optimum performance. Homologous designs have become possible only since the development of computer-aided structural analysis.

Some radio telescopes, particularly those designed for operation at very short wavelengths, are placed in protective radomes that can nearly eliminate the effect of both wind loading and temperature differences throughout the structure. Special materials that exhibit very low absorption and reflection of radio waves have been developed for such structures, but the cost of enclosing a large antenna in a suitable temperature-controlled radome may be almost as much as the cost of the movable antenna itself.

The cost of constructing a very-large-aperture antenna can be greatly reduced by fixing the structure to the ground and moving either the feed or the secondary reflector to steer the beam in the sky. For parabolic reflecting surfaces, the beam can be steered in this way over only a limited range of angle without introducing aberration and a loss of power gain. For operation at relatively long metre wavelengths where the reflecting surface need not have an accuracy better than a few centimetres, it becomes practical to build very large fixed structures in which the reflecting surface is made of simple "chicken wire" fencing or even parallel rows of wires.

Radio telescopes are used to measure broad-bandwidth continuum radiation as well as spectroscopic features due to atomic and molecular lines found in the radio spectrum of astronomical objects. In early radio telescopes, spectroscopic observations were made by tuning a receiver across a sufficiently large frequency range to cover the various frequencies of interest. This procedure, however, was extremely time-consuming and greatly restricted obser-

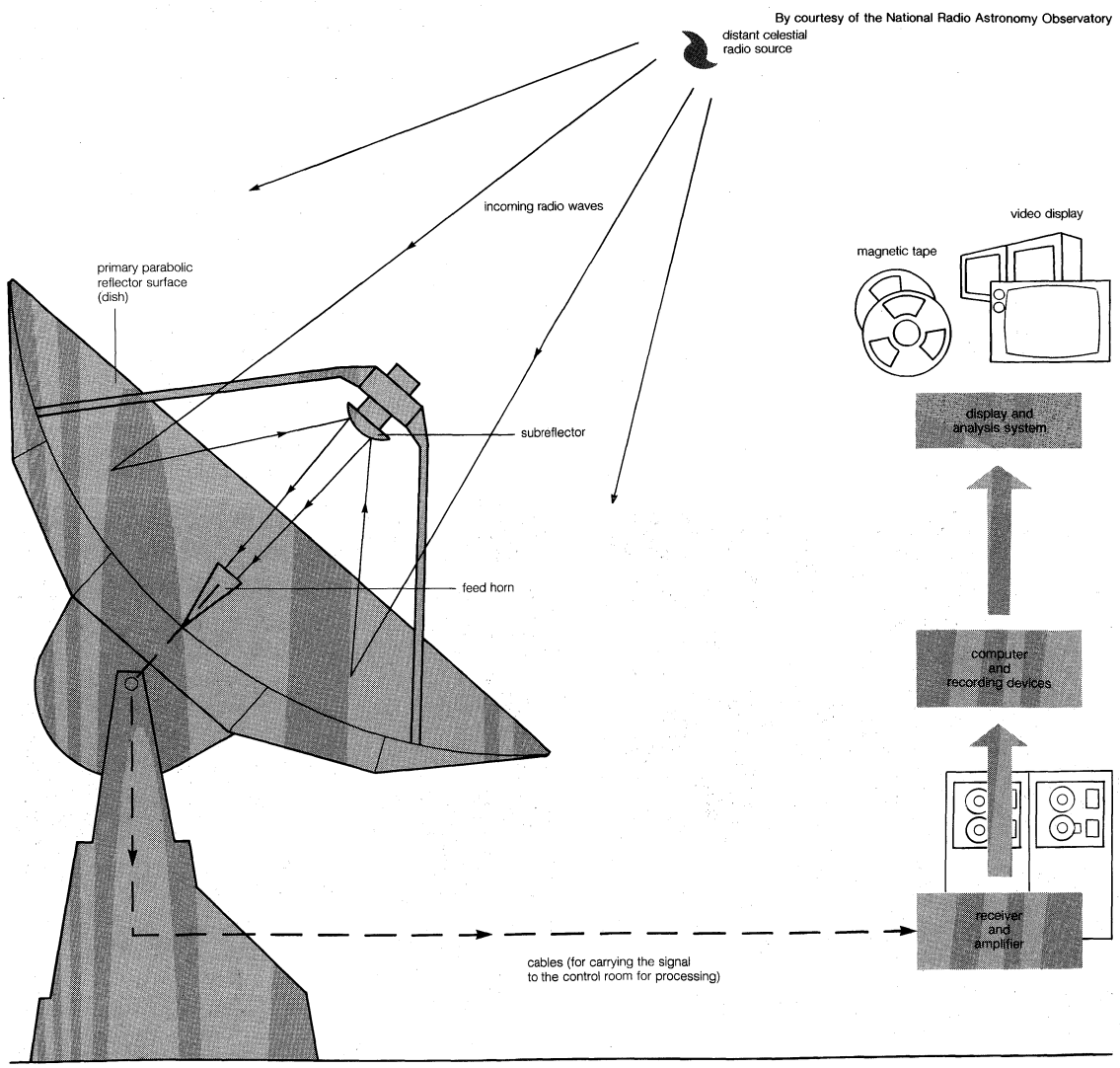


Figure 21: Radio telescope system.

vations. Modern radio telescopes observe simultaneously at a large number of frequencies by dividing the signals up into as many as several thousand separate frequency channels that may range over a total bandwidth of tens to hundreds of megahertz.

The most straightforward type of radio spectrometer employs a large number of filters, each tuned to a separate frequency and followed by a separate detector to produce a multichannel, or multifrequency, receiver. Alternatively, a single broad-bandwidth signal may be converted into digital form and analyzed by the mathematical process of autocorrelation and Fourier transformation (see below). In order to detect faint signals, the receiver output is often averaged over periods of up to several hours to reduce the effect of noise generated in the receiver.

Angular resolution

Radio interferometry and aperture synthesis. Angular resolution, or ability of a radio telescope to distinguish fine detail in the sky, depends on the wavelength of observations divided by the size of the instrument. Yet even the largest antennas, when used at their shortest operating wavelength, have an angular resolution of only about one arc minute, which is comparable to that of the unaided human eye at optical wavelengths. Because radio telescopes operate at much longer wavelengths than do optical telescopes, it was thought for many years that the resolution of radio telescopes must be much poorer than that of optical instruments. In actuality, this is not the case, for several reasons.

First, although the theoretical resolution of an optical telescope may be as good as a few hundredths of a second of arc, distortions of the incoming light signal by the Earth's atmosphere, known as seeing, diffuse the image, so that even at a good mountain site under good observing conditions the best angular resolution is only a little better than one arc second. At the much longer radio wavelengths, the distortions introduced by the atmosphere are less important, and so the theoretical angular resolution of a radio telescope can in practice be achieved (Figure 22). Because radio signals are easier than light signals to distribute over large distances without distortion, it is possible to build radio telescopes of essentially unlimited dimensions. In fact, the history of radio astronomy has been one of solving engineering problems to construct radio telescopes of continually increasing angular resolution.

The high angular resolution of radio telescopes is achieved by using the principles of interferometry to synthesize a very large effective aperture from a number of small elements. In a simple two-element radio interferometer, the signals from an unresolved, or "point," source alternately arrive in phase and out of phase as the Earth rotates and causes a change in the difference in path from the radio source to the two elements of the interferometer. This produces interference fringes in a manner similar to that in an optical interferometer. If the radio source has finite angular size, then the difference in path length to the elements of the interferometer varies across the source. The measured interference fringes from each interferometer pair thus depend on the detailed nature of the radio "brightness" distribution in the sky.

During the late 1940s Australian radio astronomers realized that each interferometer measurement is one "Fourier component" of the brightness distribution of the radio source. Further developments during the 1950s by Martin Ryle and his colleagues in Cambridge, Eng., involved the use of movable-element interferometers and the rotation of the Earth to sample a sufficient number of Fourier components with which to synthesize the effect of a large aperture and thereby reconstruct high-resolution images of the radio sky. The laborious computational task of doing Fourier transforms to obtain images from the interferometer data is accomplished with high-speed computers and the fast Fourier transform (FFT), a mathematical technique that entails the application of a group of algorithms specially suited for computing discrete Fourier transforms (see ANALYSIS: *The Fourier transform*).

Use of fast Fourier transform

In recognition of their contributions to the development of the Fourier synthesis technique, more commonly known as aperture synthesis, Ryle and Antony Hewish were awarded the 1974 Nobel Prize for Physics. During the 1960s the Swedish physicist Jan Hogbom developed a technique called "clean," which can be used to remove spurious responses from a celestial radio image caused by the use of discrete, rather than continuous, spacings in deriving the radio image. Further developments, based on a technique introduced in the early 1950s by British scientists Roger Jennison and Francis Graham Smith, led to the concept of self-calibration, which is used to remove errors in a radio image due to uncertainties in the response of individual antennas as well as small errors introduced by the propagation of radio signals through the terrestrial atmosphere.

The combination of up to millions of data points to form a single image, together with the lengthy calculations required to clean and self-calibrate, is a formidable computational task. For more complex images, such calculations are made practical only by using large, high-speed computers.

Very long baseline interferometry. In conventional interferometers and arrays, coaxial-cable, waveguide, or even fibre-optic links are used to distribute a common local oscillator reference signal to each antenna and also to return the received signal from an individual antenna to a central laboratory where it is correlated with the signals from other antennas. For antenna spacings greater than a few tens of kilometres, however, it becomes prohibitively expensive to employ real physical links to distribute the

By courtesy of (left) the National Radio Astronomy Observatory/Associated Universities, Inc., and (right) the National Optical Astronomy Observatories

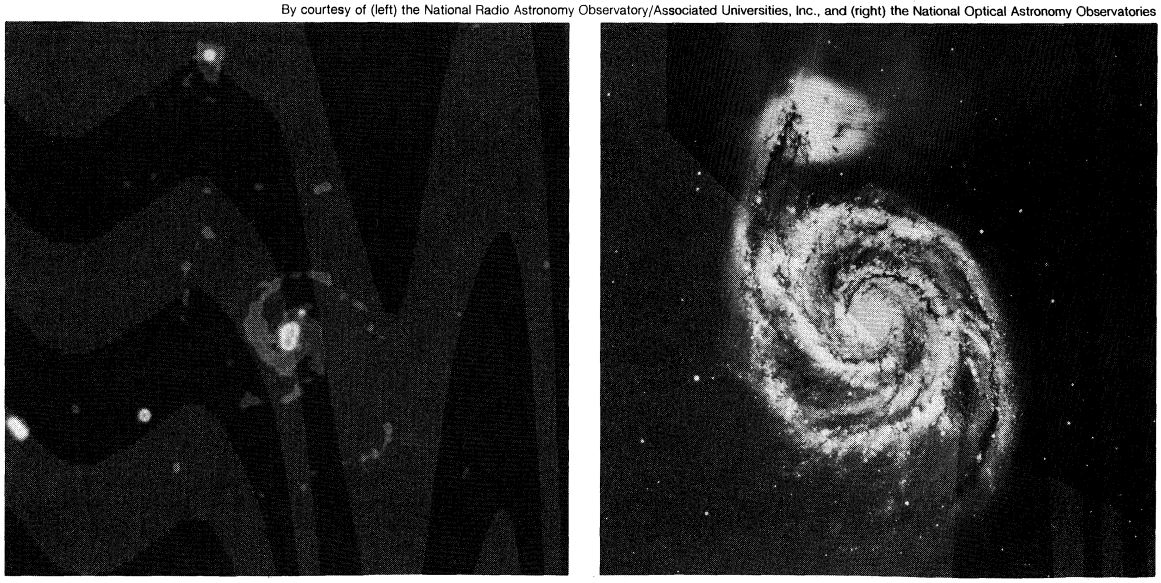


Figure 22: Comparison of a radiograph (left) and a photograph (right) of a spiral galaxy.

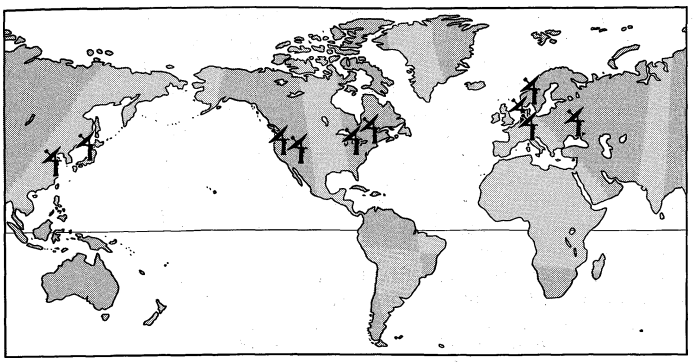


Figure 23: A portion of a gigantic interferometer system formed by linking together individual radio telescopes located many hundreds or even thousands of kilometres apart.

signals. Very high frequency (VHF) or ultrahigh frequency (UHF) radio links can be used, but the need for a large number of repeater stations makes this impractical for spacings greater than a few hundred kilometres.

Interferometer systems of essentially unlimited element separation can be formed by using the technique of very long baseline interferometry, or VLBI (see Figure 23). In a VLBI system the signals received at each element are recorded by broad-bandwidth videotape recorders located at the element. The recorded tapes are then transported to a common location where they are replayed and the signals combined to form interference fringes. The successful operation of a VLBI system requires that the tape recordings be synchronized within a few millionths of a second and that the local oscillator reference signal be stable to better than one part in a trillion. For the most precise work, hydrogen maser frequency standards are used to give a timing accuracy of only a few billionths of a second and a frequency stability of one part in a quadrillion.

For many VLBI applications, modified consumer-type videocassette recorders (VCRs) provide adequate performance, and the low cost and widespread availability of these devices have allowed as many as 18 radio telescopes throughout the world to be used simultaneously to obtain high-resolution images. The most sensitive observations, however, require the use of special recorders that are able to record up to several hundred megabits of data per second. A single magnetic tape capable of recording for several hours can contain one trillion bits of information, which is roughly equivalent to storing the entire contents of a modest-sized library.

Planetary radar astronomy

Radar techniques. Techniques analogous to those used in military and civilian radar applications are employed with radio telescopes to study the relatively nearby objects in the solar system. By measuring the spectrum and the time of flight of signals reflected from planetary surfaces, it is possible to examine topographical features, deduce rates of rotation, and determine with great accuracy the distance to the planets. Nonetheless, radio signals reflected from the planets are weak, and high-power radar transmitters are needed in order to obtain measurable signal detections. The time it takes for a radar signal to travel to Venus and back, even at the closest approach of the planet to the Earth, is about five minutes. For Saturn, it is more than two hours.

Major applications of radio telescopes. Radio telescopes permit astronomers to study many kinds of extraterrestrial radio sources. These astronomical objects emit radio waves by one of several processes, including (1) thermal radiation from solid bodies such as the planets, (2) thermal, or bremsstrahlung, radiation from hot gas in the interstellar medium, (3) synchrotron radiation from relativistic electrons in weak magnetic fields, (4) line radiation from atomic or molecular transitions that occur in the interstellar medium or in the gaseous envelopes around stars, and (5) pulsed radiation resulting from the rapid rotation of neutron stars surrounded by an intense magnetic field and energetic electrons.

Radio telescopes enabled investigators to discover intense radio emissions from Jupiter and have been used to measure the temperature of all the planets. Astronomers have relied on radar observations to map the large-scale features on the surface of Venus, which is completely obscured from visual scrutiny by the heavy cloud cover that permanently enshrouds the planet (Figure 24). In addition, radar studies have shown that Venus is rotating in the retrograde, or reverse, direction from that of the other planets. Radar measurements also have revealed the rotation of Mercury, which was previously thought to keep the same side toward the Sun. Accurate measurements of the travel time of radar signals reflected from Venus

By courtesy of D.B. Campbell, National Astronomy and Ionosphere Center, Cornell University, and *Science*, vol. 246, p. 375, Oct. 20, 1989, copyright 1989 by the American Association for the Advancement of Science

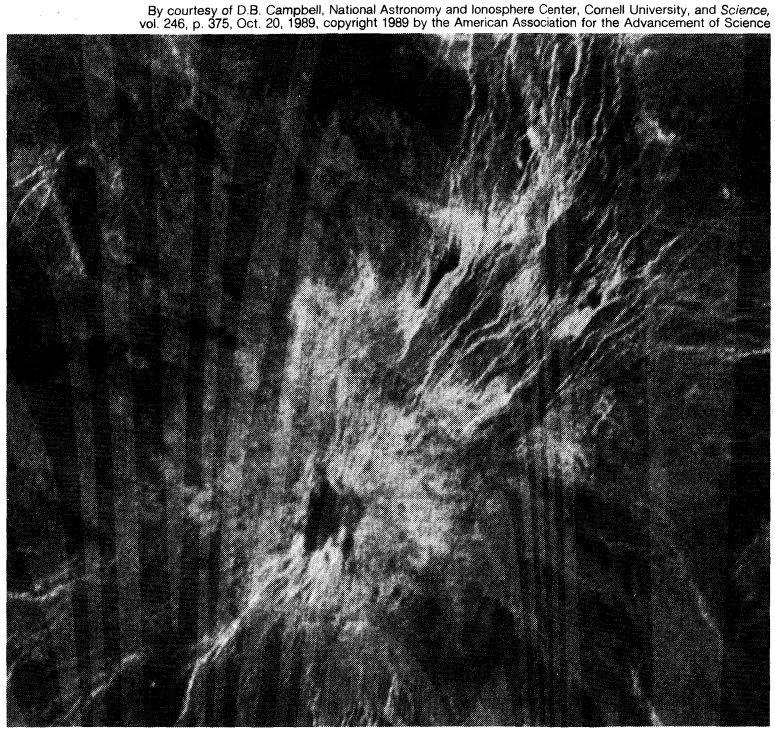


Figure 24: High-resolution radar image of a large topographic rise on Venus' surface known as Beta Regio. The summit of the volcano Theia Mons can be seen at the lower centre of the image.

near superior conjunction have indicated that radio waves passing close to the Sun slow down owing to gravity and have thereby provided a new independent test of Albert Einstein's general theory of relativity.

Broadband continuum emission throughout the radio-frequency spectrum is observed from a variety of stars (especially binary, X-ray, and other active stars), from supernova remnants, and from magnetic fields and relativistic electrons in the interstellar medium. The discovery of pulsars (from *pulsating radio stars*) in 1967 revealed the existence of rapidly rotating neutron stars throughout the Milky Way Galaxy and led to the first observation of the effect of gravitational radiation.

Discovery of molecules in interstellar space

Utilizing radio telescopes equipped with sensitive spectrometers, researchers have discovered more than 50 separate molecules, including familiar chemical compounds like water vapour, formaldehyde, ammonia, methanol, ethanol, and carbon dioxide. The important spectral line of atomic hydrogen at 1421.405 MHz (21 centimetres) is used to determine the motions of hydrogen clouds in the Milky Way Galaxy and other spiral systems. This is done by measuring the change in the wavelength of the observed lines arising from Doppler shift. It has been established from such measurements that the rotational velocities of the hydrogen clouds vary with distance from the galactic centre. The mass of a spiral galaxy can, in turn, be estimated using this velocity data (see also COSMOS: *Galaxies*).

Radio telescopes have discovered powerful radio galaxies and quasars beyond the Milky Way system. These cosmic objects have intense clouds of radio emission that extend hundreds of thousands of light-years away from a central energy source located in an active galactic nucleus (Figure 25) or quasar. VLBI observations made with worldwide networks of radio telescopes have revealed apparent faster-than-light motion in many quasars. (For more specific information about quasars and other extragalactic radio sources, see COSMOS: *Quasars and related objects.*)

Measurements made by Arno A. Penzias and Robert W. Wilson with an experimental communications antenna at Bell Telephone Laboratories detected the existence of cosmic background radiation at a temperature of 3 K. This radiation, which comes from all parts of the sky, is thought to be the remaining radiation from the hot big bang, the primeval explosion from which the universe presumably originated some 15 billion years ago.

Important radio telescopes. The first really large fully steerable radio telescope was completed in 1957 at Jodrell Bank, Eng. This 76-metre instrument is still used for a number of research programs (Table 3). The world's largest fully steerable radio telescope is the 100-metre diameter antenna operated by the Max Planck Institute for Radio Astronomy at Effelsberg, near Bonn, W.Ger. (Figure 26). It is used in a number of different wavelength bands as short as one centimetre for atomic and molecular spectroscopy and for other galactic and extragalactic studies. Because of its large collecting area and full sky coverage, the Effelsberg radio telescope is frequently used for worldwide VLBI observations.

The Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia maintains near Parkes, N.S.W., a 64-metre radio telescope that is the largest of its kind in the Southern Hemisphere. The world's largest radome-enclosed radio telescope is the 36-metre Haystack antenna operated by the Northeast Radio Observatory Corporation under agreement with the Massachusetts Institute of Technology (MIT).

A 91-metre fixed-azimuth radio telescope with limited elevation motion was operated by the National Radio Astronomy Observatory in Green Bank, W.Va., U.S., until its unexpected collapse in late 1988. A 43-metre equatorially mounted radio instrument, however, remains in operation in Green Bank; this telescope is used primarily for molecular spectroscopy at wavelengths as short as one centimetre. Green Bank is located in the national Radio Quiet Zone, which offers unique protection for radio telescopes from sources of man-made interference.

The largest single radio telescope in the world is the 305-metre fixed spherical reflector operated by Cornell University near Arecibo, P.R. (Figure 27). The 305-metre antenna has an enormous collecting area, but the beam can be moved through only a limited angle of about 20° from the zenith. It is used for planetary radar astronomy as well as for studying pulsars and other galactic and extragalactic phenomena. The Soviet RATAN-600 telescope (RATAN stands for Radio Telescope of the Academy of Sciences), located near Zelenchukskaya in the Caucasus Mountains, has 895 reflecting panels, each 7.4 metres high, arranged in a ring 576 metres in diameter. Using long parabolic cylinders or dipole elements, researchers in Australia, France, India, Italy, and the Ukrainian S.S.R. have also built antennas with very large collecting areas.

Several smaller, more precise radio telescopes for observing at millimetre wavelength have been installed high atop mountains, where clear skies and high altitudes minimize

By courtesy of the National Radio Astronomy Observatory/Associated Universities, Inc.



Figure 25: VLA (Very Large Array) image of an interacting twin-jet radio galaxy. The two black dots (at the bottom centre) are each associated with one of the twin nuclei of a distant galaxy. The jets appear to interact and wrap around one another.

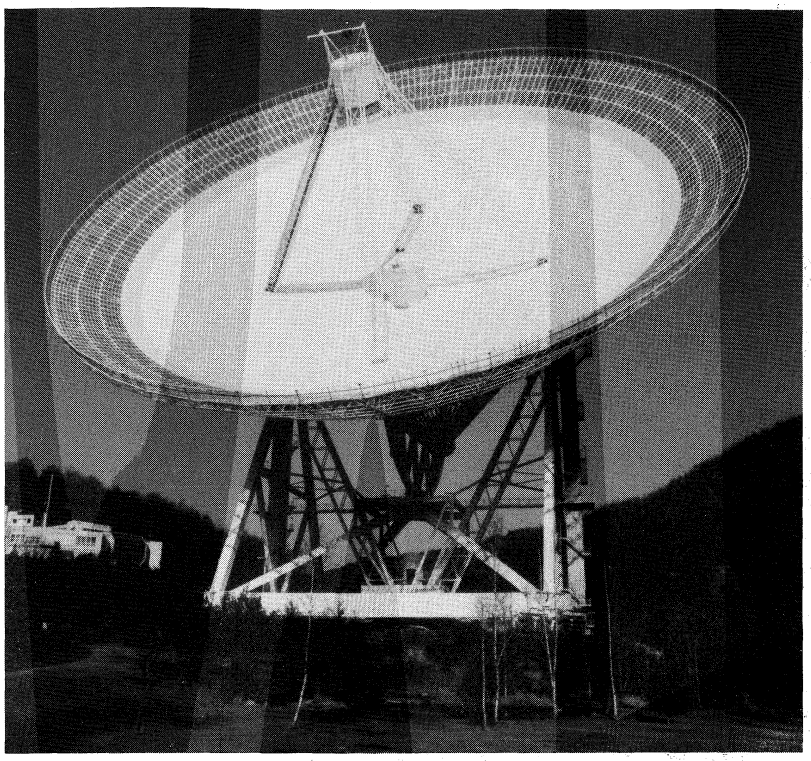


Figure 26: The 100-metre radio telescope at Effelsberg, near Bonn.
By courtesy of the Max-Planck-Institut fur Radioastronomie; photographer, G. Hutschenreiter

absorption from the terrestrial atmosphere. A 45-metre radio dish near Nobeyama Plateau, Japan, is used for observations at wavelengths as short as a few millimetres. The French-German Institut de Radio Astronomie Millimetrique (IRAM) in Grenoble, Fr., operates a 30-metre antenna at an altitude of 2,850 metres on Veleta Peak in the Spanish Sierra Nevada for observations at wavelengths as short as one millimetre. Radio telescopes having a diameter of 10 to 15 metres that operate at submillimetre wavelengths are positioned near the summit of Mauna Kea, Hawaii, U.S., at an elevation of about 4,250 metres and on La Silla Hill in Chile at an elevation of 2,350 metres. Millimetre interferometers and arrays are operated at the Owens Valley Radio Observatory of the California Institute of Technology, the Hat Creek Observatory Laboratories of the University of California at Berkeley, the IRAM Plateau de Bure facility in France, and the Nobeyama Observatory.

Very Large Array

The world's most powerful radio telescope is the Very Large Array (VLA) located on the Plains of San Agustin near Socorro, N.M., U.S. The VLA consists of 27 parabolic antennas, each measuring 25 metres in diameter. The total collecting area is equivalent to a single 130-metre antenna. Each element of the VLA can be moved by a transporter along a Y-shaped railroad track; it is possible to change the length of the arms between 600 metres and 21 kilometres to vary the resolution of the system (Figure 28). Each antenna is equipped with receivers that operate in six different bands from wavelengths of approximately one centimetre to one metre. When used at the shorter wavelength in the largest antenna configuration, the angular resolution of the VLA is several tenths of one arc second. The VLA is operated by the U.S. National Radio Astronomy Observatory as a national facility and is used by more than 500 astronomers each year for a wide variety of research programs devoted to the study of the solar system, Milky Way Galaxy, and extragalactic systems.

There are a number of other large radio telescope arrays around the world. The Netherlands Foundation for Radio Astronomy operates the Westerbork Synthesis Radio Telescope in continental Europe (see Table 3). The Commonwealth Scientific and Industrial Research Organization maintains the six-element Australia Telescope at Culgoora, N.S.W., for studies of the southern skies. A number of smaller arrays are operated by the Mullard Radio Astronomy Observatory near Cambridge, Eng., including the pioneering One-Mile Radio Telescope and a simple yet very powerful array of Yagi antennas operating at 151 MHz.

The Multi-Element Radio-Linked Interferometer Network (MERLIN), operated by the Nuffield Radio Astronomy Laboratories at Jodrell Bank, employs microwave radio links to connect seven antennas separated by up to 200 kilometres in the southern part of England. It is used primarily to study compact radio sources associated with quasars and active galactic nuclei.

The Very Long Baseline Array (VLBA), which will consist of ten 25-metre dishes spread across the United States from the Virgin Islands to Hawaii upon completion in the early 1990s, is expected to yield radio images of quasars and other compact radio sources of unprecedented angular resolution and quality. Other radio telescopes being constructed in such countries as Italy and Australia will be dedicated to VLBI research programs. When used together with the VLBA and other radio telescopes throughout the world, the effective resolution of the system will be comparable to that of a single antenna whose diameter is roughly equivalent to the Earth's. Future space-based radio antennas are expected to increase the resolution still further to produce images of cosmic radio sources in even finer detail. (K.I.K.)

OTHER TYPES OF TELESCOPES

Infrared telescopes. Telescopic systems of this type do not really differ significantly from reflecting telescopes designed to observe in the visible region of the electromagnetic spectrum. The main difference is in the physical location of the infrared telescope, since infrared photons have lower energies than those of visible light. The infrared rays are readily absorbed by the water vapour in the Earth's atmosphere, and most of this water vapour is located at the lower atmospheric regions—*i.e.*, near sea level. Earth-bound infrared telescopes have been successfully located on high mountaintops, as, for example, Mauna Kea in Hawaii. The other obvious placement of infrared instruments is in a satellite such as the Infrared Astronomical Satellite (IRAS), which mapped the celestial sky in the infrared in 1983. The Kuiper Airborne Observa-

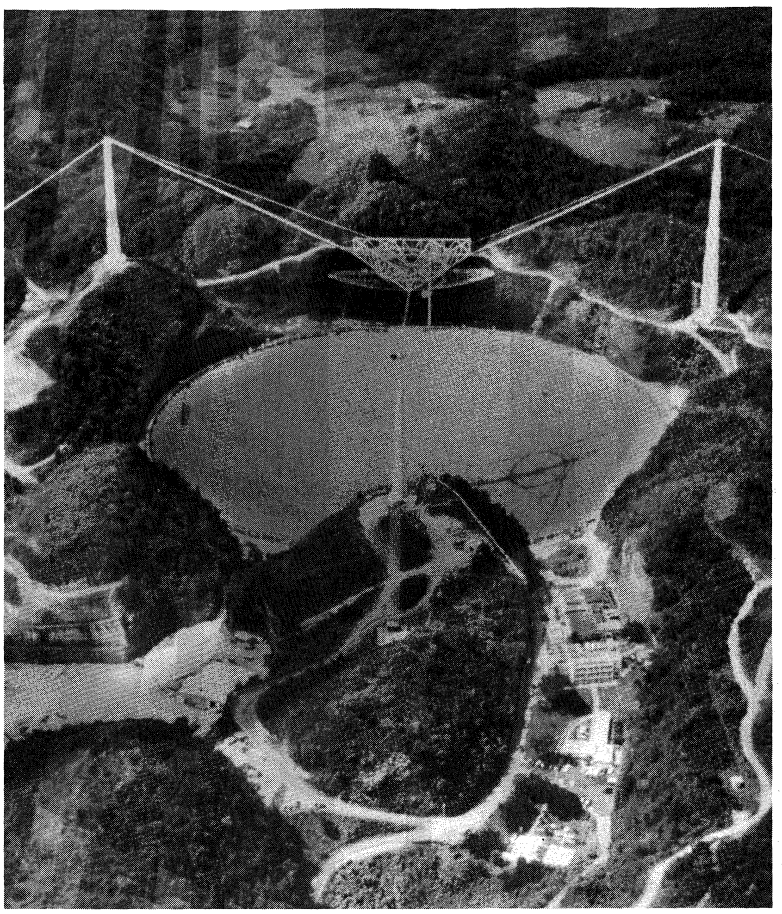


Figure 27: The 305-metre rigid dish of the Arecibo Observatory in Puerto Rico, operated by Cornell University under a cooperative agreement with the National Science Foundation.
By courtesy of the National Astronomy and Ionosphere Center, Cornell University

tory, operated by NASA, consists of a 0.9-metre telescope that is flown in a special airplane above the water vapour to collect infrared data. The majority of the data collection in the infrared is achieved with some type of electronic camera, since ordinary film is unable to register the low-energy photons.

Use of electronic cameras

Another example of an infrared telescope is the United Kingdom Infrared Telescope (UKIRT), which has a 3.8-metre mirror made of Cer-Vit (trademark), a glass ceramic that has a very low coefficient of expansion (see above; Figure 29). This instrument is configured in a Cassegrain design and employs a thin monolithic primary mirror with a lightweight support structure. This telescope is located at Mauna Kea Observatory. The 3-metre Infrared Telescope Facility (IRTF), also located at Mauna Kea, is sponsored by NASA and operated by the University of Hawaii.

Ultraviolet telescopes. These telescopes are used to examine the shorter wavelengths of the electromagnetic spectrum immediately adjacent to the visible portion. Like the infrared telescopes, the ultraviolet systems also employ reflectors as their primary collectors. Ultraviolet radiation is composed of higher-energy photons than infrared radiation, which means that photographic techniques as well as electronic detectors can be used to collect astronomical data. The Earth's stratospheric ozone layer, however, blocks all wavelengths shorter than 3000 angstroms from reaching ground-based telescopes. As this ozone layer lies at an altitude of 20 to 40 kilometres, astronomers have to resort to rockets and satellites to make observations from above it. Since 1978 an orbiting observatory known as the International Ultraviolet Explorer (IUE) has studied celestial sources of ultraviolet radiation. The IUE telescope is equipped with a 45-centimetre mirror and records data electronically down to 1000 angstroms. The IUE is in a synchronous orbit (*i.e.*, its period of revolution around the Earth is identical to the period of the planet's rotation) in view of NASA's Goddard Space Flight Center in Greenbelt, Md., and so data can be transmitted to the ground

International Ultraviolet Explorer

Table 4: Some Notable Radio Telescope Systems

size and type	designation or institution	location
305-metre fixed spherical dish	Arecibo Observatory	Arecibo, P.R.
27 25-metre dishes in 36-kilometre array	Very Large Array (VLA)	near Socorro, N.M., U.S.
100-metre steerable dish	Max Planck Institute for Radio Astronomy	Effelsberg, near Bonn, W.Ger.
76-metre steerable dish	Jodrell Bank Experimental Station (also called Nuffield Radio Astronomy Laboratories)	Jodrell Bank, Cheshire, Eng.
64-metre steerable dish	Australia Telescope National Facility	Parkes, N.S.W., Australia
14 25-metre dishes in 3.2-kilometre array	Westerbork Synthesis Radio Telescope	Westerbork, Neth.
30-metre fully steerable precision dish	Institut de Radio Astronomie Millimetrique (IRAM)	Veleta Peak, Spain
45-metre steerable dish	National Astronomy Observatory	Nobeyama, Japan
36-metre, radome-enclosed, fully steerable dish	Haystack Observatory	Westford, Mass., U.S.

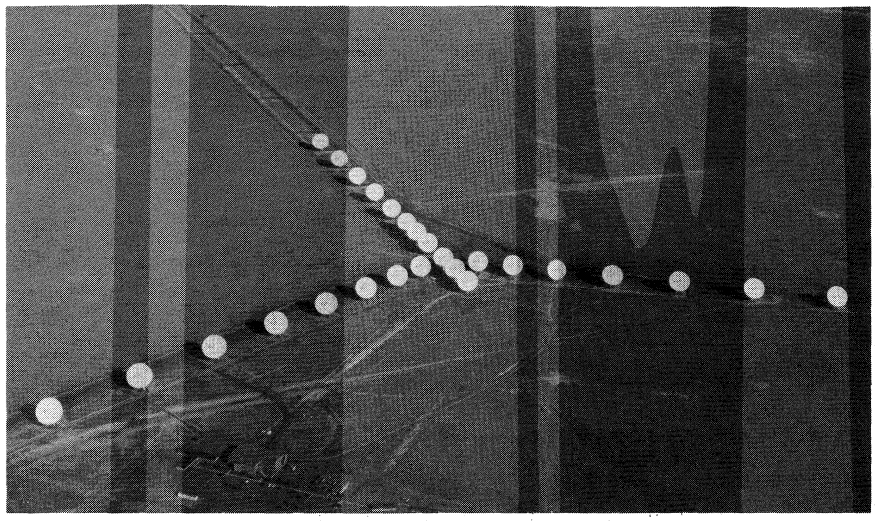


Figure 28: The Very Large Array near Socorro, N.M.
By courtesy of the National Radio Astronomy Observatory/Associated Universities, Inc.

station at the end of each observing tour and examined immediately on a television monitor.

Another Earth-orbiting spacecraft, the Extreme Ultraviolet (EUV) Explorer satellite, which is scheduled to be launched in the early 1990s, is designed to survey the sky in the extreme ultraviolet region between 400 and 900 angstroms. It has four telescopes with gold-plated mirrors, the design of which is critically dependent on the transmission properties of the filters used to define the EUV band passes. The combination of the mirrors and filters has been selected to maximize the telescope's sensitivity to detect faint EUV sources. Three of the telescopes have scanners that are pointed in the satellite's spin plane. The fourth telescope, the Deep Survey/Spectrometer Telescope, is directed in an anti-Sun direction. Its function is to conduct a photometric deep-sky survey in the ecliptic plane for part of the mission and then to collect spectroscopic observations in the final phase of the mission.

X-ray telescopes. The X-ray telescope is used to examine the shorter-wavelength region of the electromagnetic spectrum adjacent to the ultraviolet region. The design of this type of telescope must be radically different from that of a conventional reflector. Since X-ray photons have so much energy, they would pass right through the mirror of a standard reflector. X rays must be bounced off a mirror at a very low angle if they are to be captured. (This technique is referred to as grazing incidence.) For this reason, the mirrors in X-ray telescopes are mounted with their surfaces only slightly off a parallel line with the incoming X rays, as seen in Figure 30. Application of the grazing-incidence principle makes it possible to focus X rays from a cosmic object into an image that can be recorded electronically (see Figure 31).

Space-based X-ray telescopes

NASA launched a series of three High-Energy Astronomy Observatories (HEAOs) during the late 1970s to explore cosmic X-ray sources. HEAO-1 mapped the X-ray sources with high sensitivity and high resolution. Some of the more interesting of these objects were studied in detail by HEAO-2 (named the Einstein Observatory). HEAO-3, on the other hand, was used primarily to investigate cosmic rays and gamma rays.

The European X-ray Observatory Satellite (EXOSAT), developed by the European Space Agency (ESA), was capable of greater spectral resolution than the Einstein Observatory and was more sensitive to X-ray emissions at shorter wavelengths. EXOSAT remained in orbit from 1983 to 1986.

A much larger X-ray astronomy satellite is scheduled to be launched in 1990 as part of a cooperative program involving the United States, West Germany, and the United Kingdom. This satellite, called the Röntgenstrahlen Satellit (ROSAT), has two parallel grazing-incidence telescopes. One of them, the X-ray telescope (XRT), bears many similarities to the equipment of the Einstein Observatory,

Copyright © Royal Observatory, Edinburgh

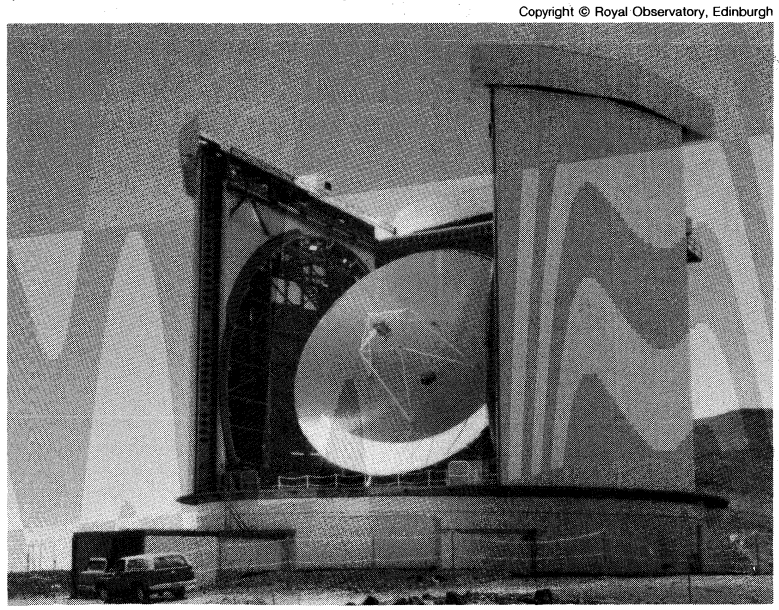


Figure 29: The United Kingdom Infrared Telescope located on Mauna Kea, Hawaii.

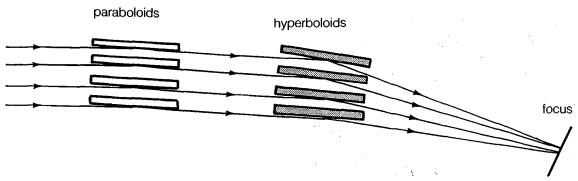


Figure 30: The grazing-incidence principle of the X-ray telescope.

except that it has a larger geometric area and better mirror resolution. One of the focal plane instruments, the high-resolution imager (HRI)—supplied by the Smithsonian Astrophysical Observatory—is expected to show considerable improvement over similar Einstein Observatory instrumentation. The other telescope, the extended ultraviolet wide-field camera, has an imaging detector much like the X-ray HRI. A positive sensitive proportional counter is to be used for an X-ray sky survey to produce a catalog of 100,000 sources with a positional accuracy of better than 30 arc seconds. A wide-field camera with a 5°-diameter field of view is also part of the ROSAT instrument package. It is designed to produce an extended ultraviolet survey with arc minute source positions in this wavelength region, making it the first instrument with such capability. The ROSAT mirrors are gold-coated and will permit detailed examination of the sky from 6 to 100 angstroms.

Gamma-ray telescopes. These instruments require the use of grazing-incidence techniques similar to those employed with X-ray telescopes. Gamma rays are the shortest (about 0.1 angstrom or less) known waves in the electromagnetic spectrum. As mentioned above, HEAO-3 was developed to collect data from cosmic gamma-ray sources. NASA and collaborative international agencies have numerous ongoing and planned projects in the area of gamma-ray astronomy. The scientific objectives of the programs include determining the nature and physical parameters of high-energy (up to 10 gigaelectron volts) astrophysical systems. Examples of such systems include stellar coronas, white dwarfs, neutron stars, black holes, supernova remnants, clusters of galaxies, and diffuse gamma-ray background. In addition to satellite investigations of these cosmic high-energy sources, NASA has an extensive program that involves the design and development of gamma-ray telescope systems for deployment in high-altitude balloons. All mirrors in gamma-ray telescopes have gold coatings similar to those in X-ray telescope mirrors.

THE DEVELOPMENT OF THE TELESCOPE AND AUXILIARY INSTRUMENTATION

Evolution of the optical telescope. Galileo is credited with having developed telescopes for astronomical observation in 1609. While the largest of his instruments was only about 120 centimetres long and had an objective diameter of 5 centimetres, it was equipped with an eyepiece that provided an upright (*i.e.,* erect) image. Galileo used his modest instrument to explore such celestial phenomena as the valleys and mountains of the Moon, the phases of Venus, the Milky Way, and the four largest satellites of Jupiter, which had never been systematically observed before.

The reflecting telescope was developed in 1668 by Newton, though John Gregory had independently conceived of an alternative reflector design in 1663. Cassegrain introduced another variation of the reflector in 1672. Near the end of the century, others attempted to construct refractors as long as 61 metres, but these instruments were too awkward to be effective.

The most significant contribution to the development of the telescope in the 18th century was that of Sir William Herschel. Herschel, whose interest in telescopes was kindled by a modest 5-centimetre Gregorian, persuaded the king of England to finance the construction of a reflector with a 12-metre focal length and a 120-centimetre mirror. Herschel is credited with having used this instrument to lay the observational groundwork for the concept of extragalactic "nebulas"—*i.e.,* galaxies outside the Milky Way system.

The reflectors of Rosse and Lassell

Reflectors continued to evolve during the 19th century with the work of William Parsons, 3rd Earl of Rosse, and William Lassell. In 1845 Lord Rosse constructed in Ireland a reflector with a 185-centimetre mirror and a focal length of about 16 metres. For 75 years this telescope ranked as the largest in the world and was used to explore thousands of nebulas and star clusters. Lassell built several reflectors, the largest of which was on Malta; this instrument had a 124-centimetre primary mirror and a focal length of more than 10 metres. His telescope had greater reflecting power than that of Rosse, and it enabled him to catalog 600 new nebulas as well as to discover several satellites of the outer planets—Triton (Neptune's largest moon), Hyperion (Saturn's 8th moon), and Ariel and Umbriel (two of Uranus' moons).

Refractor telescopes, too, underwent development during the 18th and 19th centuries. The last significant one to be built was the 1-metre refractor at Yerkes Observatory. Installed in 1897, it remains the largest refracting system in the world. Its objective was designed and constructed by the optician Alvan Clark, while the mount was built by the firm of Warner & Swasey.

The reflecting telescope has predominated in the 20th century. The rapid proliferation of larger and larger instruments of this type began with the installation of the 2.5-metre reflector at the Mount Wilson Observatory near Pasadena, Calif., U.S. The technology for mirrors under-

By courtesy of Dr. Frederick Seward, Harvard-Smithsonian Center for Astrophysics

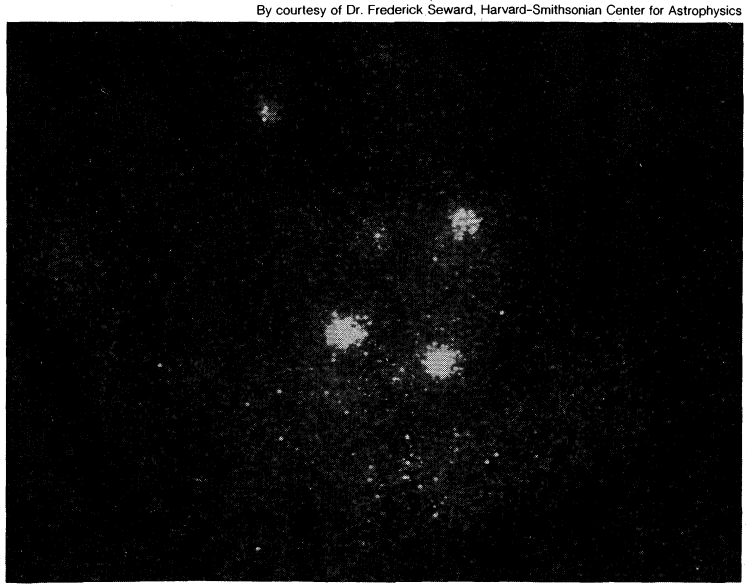


Figure 31: X-ray image of the Eta Carinae nebula.

went a major advance when the Corning Glass Works (in Steuben county, N.Y., U.S.) developed Pyrex. This borosilicate glass, which undergoes substantially less expansion than ordinary glass, was used in the 5-metre Hale reflector built in 1950 at the Palomar Observatory. Pyrex also was utilized in the main mirror of the world's largest telescope, the 6-metre reflector of the Special Astrophysical Observatory in Zelenchukskaya. In recent years, much better materials for mirrors have become available. Cer-Vit, for example, was used for the 4.2-metre William Herschel Telescope of the Roque de los Muchachos Observatory in the Canary Islands, and Zerodur (trademark) for the 3.5-metre reflector at the German-Spanish Astronomical Center in Calar Alto, Spain. (B.L.Kl.)

Development of new materials for mirrors

Development of the radio telescope. Extraterrestrial radio emission was first reported in 1933 by Karl Jansky, an engineer at the Bell Telephone Laboratories, while he was searching for the cause of shortwave interference. Jansky had mounted a directional radio antenna on a turntable so that he could point it at different parts of the sky in order to determine the direction of the interfering signals. He not only detected interference from distant thunderstorms but also located a source of such radio "noise" from the centre of the Milky Way Galaxy. This first detection of cosmic radio waves received considerable attention from the general public but only passing notice from the astronomical community.

Grote Reber, a radio engineer and amateur radio operator, built a 9.5-metre parabolic reflector in his backyard in Wheaton, Ill., U.S., to continue Jansky's investigation of cosmic radio noise. In 1944 he published the first radio map of the sky. After World War II ended, the technology that had been developed for military radar was applied to astronomical research. Radio telescopes of increasing size and sophistication were built first in Australia and Great Britain and later in the United States and other countries (see above *Important radio telescopes*). (K.I.K.)

Advances in auxiliary instrumentation. Almost as important as the telescope itself are the auxiliary instruments that the astronomer uses to exploit the light received at the focal plane. Examples of such instruments are the camera, spectrograph, photomultiplier tube, charge-coupled device (CCD), and charge injection device (CID). Each of these instrument types is discussed below.

Cameras. John Draper of the United States photographed the Moon as early as 1840 by applying the daguerreotype process. The French physicists A.-H.-L. Fizeau and J.-B.-L. Foucault succeeded in making a photographic image of the Sun in 1845. Five years later, astronomers at Harvard Observatory took the first photographs of the stars.

The use of photographic equipment in conjunction with telescopes has benefited astronomers greatly, giving them two distinct advantages: first, photographic images provide a permanent record of celestial phenomena and, second, photographic plates integrate the light from celestial sources over long periods of time and thereby permit astronomers to see much fainter objects than they would be able to observe visually. Typically, the camera's photographic plate (or film) is mounted in the focal plane of the telescope. The plate (or film) consists of glass or of a plastic material that is covered with a thin layer of a silver compound. The light striking the photographic medium causes the silver compound to undergo a chemical change. When processed, a negative image results—*i.e.,* the brightest spots (the Moon and the stars, for example) appear as the darkest areas on the plate or the film.

Spectrographs. Newton noted the interesting way in which a piece of glass can break up light into different bands of colour, but it was not until 1814 that the German physicist Joseph von Fraunhofer discovered the lines of the solar spectrum and laid the basis for spectroscopy. The spectrograph, as illustrated in Figure 32, consists of a slit, a collimator, a prism for dispersing the light, and a focusing lens. The collimator is an optical device that produces parallel rays from a focal plane source—*i.e.,* gives the appearance that the source is located at an infinite distance. The spectrograph enables astronomers to analyze the chemical composition of planetary atmospheres, stars, nebulas, and other celestial objects. A bright line in the spectrum indicates the presence of a glowing gas radiating at a wavelength characteristic of the chemical element in the gas. A dark line in the spectrum usually means that a cooler gas has intervened and absorbed the lines of the element characteristic of the intervening material. The lines also may be displaced to either the red end or the blue end of the spectrum. This effect was first noted in 1842 by the Austrian physicist Christian Johann Doppler. When a light source is approaching, the lines are shifted toward the blue end of the spectrum, and when the source is receding, the lines are shifted toward its red end. This effect, known as the Doppler effect, permits astronomers to study the relative motions of celestial objects with respect to the Earth's motion.

Studying spectral lines

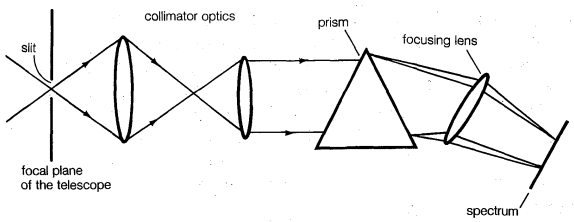


Figure 32: A spectrograph and its components.

The slit of the spectrograph is placed at the focal plane of the telescope. The resulting spectrum may be recorded photographically or with some kind of electronic detector such as a photomultiplier tube, CCD, or CID. If no recording device is used, then the optical device is technically referred to as a spectroscope.

Photomultiplier tubes. The photomultiplier tube is an enhanced version of the photocell, which was first used by astronomers to record data electronically. The photocell contains a photosensitive surface that generates an electric current when struck by light from a celestial source. The photosensitive surface is positioned just behind the focus. A diaphragm of very small aperture is usually placed in the focal plane to eliminate as much of the background light of the sky as possible. A small lens is used to focus the focal plane image on the photosensitive surface, which, in the case of a photomultiplier tube, is referred to as the photocathode. In the photomultiplier tube a series of special sensitive plates are arranged geometrically to amplify or multiply the electron stream. Frequently, magnifications of a million are achieved by this process.

The photomultiplier tube has a distinct advantage over the photographic plate. With the photographic plate the relationship between the brightness of the celestial source and its registration on the plate is not linear. In the case of the photomultiplier tube, however, the release of electrons in the tube is directly proportional to the intensity of light from the celestial source. This linear relationship is very useful for working over a wide range of brightness. A disadvantage of the photomultiplier tube is that only one object can be recorded at a time. The output from such a device is sent to a recorder or digital storage device to produce a permanent record.

Charge-coupled devices. The charge-coupled device uses a light-sensitive material on a silicon chip to electronically detect photons in a way similar to the photomultiplier tube. The principal difference is that the chip also contains integrated microcircuitry required to transfer the detected signal along a row of discrete picture elements (or pixels) and thereby scan a celestial object or objects very rapidly. When individual pixels are arranged simply in a single row, the detector is referred to as a linear array. When the pixels are arranged in rows and columns, the assemblage is called a two-dimensional array.

Pixel arrays

Pixels can be assembled in various sizes and shapes. The Hubble Space Telescope has a CCD detector with a $1{,}600 \times 1{,}600$ pixel array. Actually, there are four 800×800 pixel arrays mosaicked together. The sensitivity of a CCD is 100 times greater than a photographic plate and so has the ability to quickly scan objects such as planets, nebulas, and star clusters and record the desired data. Another feature of the CCD is that the detector material may be altered to provide more sensitivity at different

wavelengths. Thus, some detectors are more sensitive in the blue region of the spectrum than in the red region.

Today, most large observatories use CCDs to record data electronically. Another similar device, the charge injection device, is sometimes employed. The basic difference between the CID and the CCD is in the way the electric charge is transferred before it is recorded; however, the two devices may be used interchangeably as far as astronomical work is concerned.

Impact of technological developments. *Computers.* Besides the telescope itself, the electronic computer has become the astronomer's most important tool. Indeed, the computer has revolutionized the use of the telescope to the point where the collection of observational data is now completely automated. The astronomer need only identify the object to be observed, and the rest is carried out by the computer and auxiliary electronic equipment.

A telescope can be set to observe automatically by means of electronic sensors appropriately placed on the telescope axis. Precise quartz or atomic clocks send signals to the computer, which in turn activates the telescope sensors to collect astronomical data at the proper time. The computer not only makes possible more efficient use of telescope time but also permits a more detailed analysis of the data collected than could have been done manually. Data analysis that would have taken a lifetime or longer to complete with a mechanical calculator can now be done in a matter of hours or even minutes with a high-speed computer.

Improved means of recording and storing computer data also have contributed to astronomical research. Optical disc data storage technology, such as the compact disc-read only memory (CD-ROM) or the write once-read many times memory (WORM), has provided astronomers with the ability to store and retrieve vast amounts of telescopic and other astronomical data. A 12-centimetre CD-ROM disc, for example, may hold up to 600 megabytes of data—the equivalent of 20 nine-track magnetic tapes or 1,500 floppy discs. A 13-centimetre WORM disc typically holds about 300 to 400 megabytes of data.

Optical disc data storage

Rockets and spacecraft. As noted earlier, the quest for new knowledge about the universe has led astronomers to study electromagnetic radiation other than just visible light. Such forms of radiation, however, are blocked for the most part by the Earth's atmosphere, and so their detection and analysis can only be achieved from above this gaseous envelope.

During the late 1940s, single-stage sounding rockets were sent up to 160 kilometres or more to explore the upper layers of the atmosphere. From 1957, more sophisticated multistage rockets were launched as part of the International Geophysical Year; these rockets carried artificial satellites equipped with a variety of scientific instruments. Beginning in 1959, the Soviet Union and the United States, engaged in a "space race," intensified their efforts and launched a series of unmanned probes to explore the Moon. Lunar exploration culminated with the first manned landing on the Moon by the U.S. Apollo 11 astronauts on July 20, 1969. Numerous other U.S. and Soviet spacecraft were sent to further study the lunar environment until the mid-1970s.

Since the early 1960s both the United States and the Soviet Union have also launched a multitude of unmanned deep-space probes to learn more about the other planets and satellites of the solar system. Carrying television cameras, detectors, and an assortment of other instruments, these planetary probes have sent back impressive amounts of scientific data and close-up pictures. Among the most successful missions have been those involving the Soviet Venera probes to Venus and the U.S. Viking 1 and 2 landings on Mars (see Figure 33) and Voyager 2 flybys of Jupiter, Saturn, Uranus, and Neptune. When the Voyager 2 probe flew past Neptune and its moons in August 1989 (Figure 34), every one of the known planets except Pluto had been explored by spacecraft. Many long-held views, particularly those about the outer planets, were altered by the findings of the Voyager probe. These findings included the discovery of several rings and six additional satellites around Neptune, all of which are undetectable to ground-based telescopes.

Planetary exploration with unmanned probes

Specially instrumented spacecraft have enabled astronomers to investigate other celestial phenomena as well. The Orbiting Solar Observatories (OSOs) and Solar Maximum Mission, Earth-orbiting U.S. satellites equipped with ultraviolet detector systems, have provided a means for studying solar activity (see also above *Earth-orbiting space telescopes*). Another example is the Giotto probe of the European Space Agency, which enabled astronomers to obtain detailed photographs of the nucleus of Comet Halley during the 1986 passage of the comet. (For further details concerning the use of sounding rockets and spacecraft for space exploration, see EXPLORATION: *Space programs* and *Contributions to the physical sciences.*)

(B.L.Kl.)

By courtesy of the National Aeronautics and Space Administration

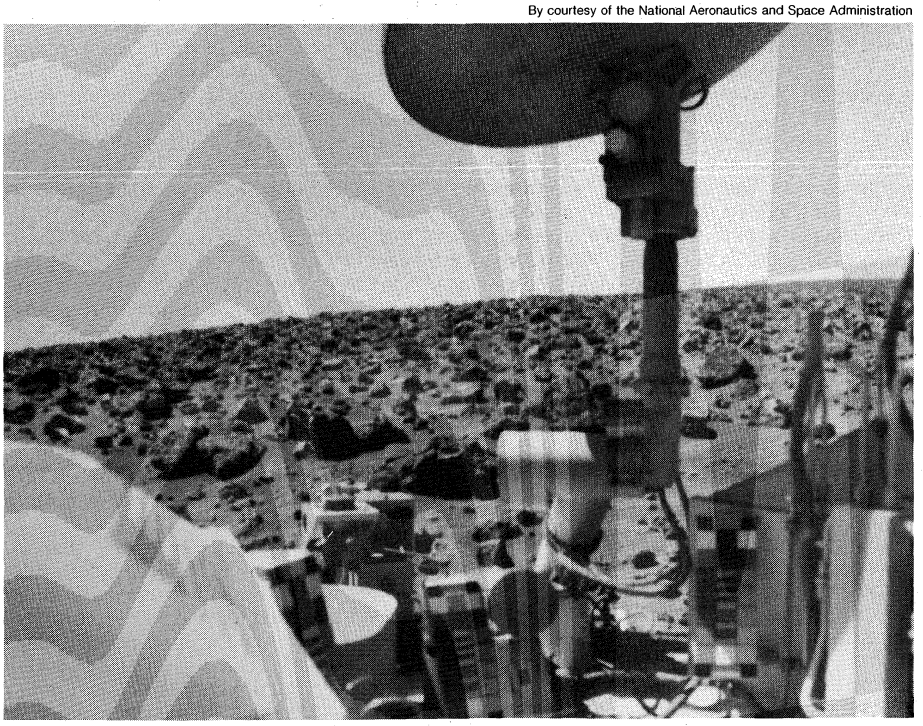


Figure 33: Mars's Utopian Plain photographed by the Viking 2 lander.

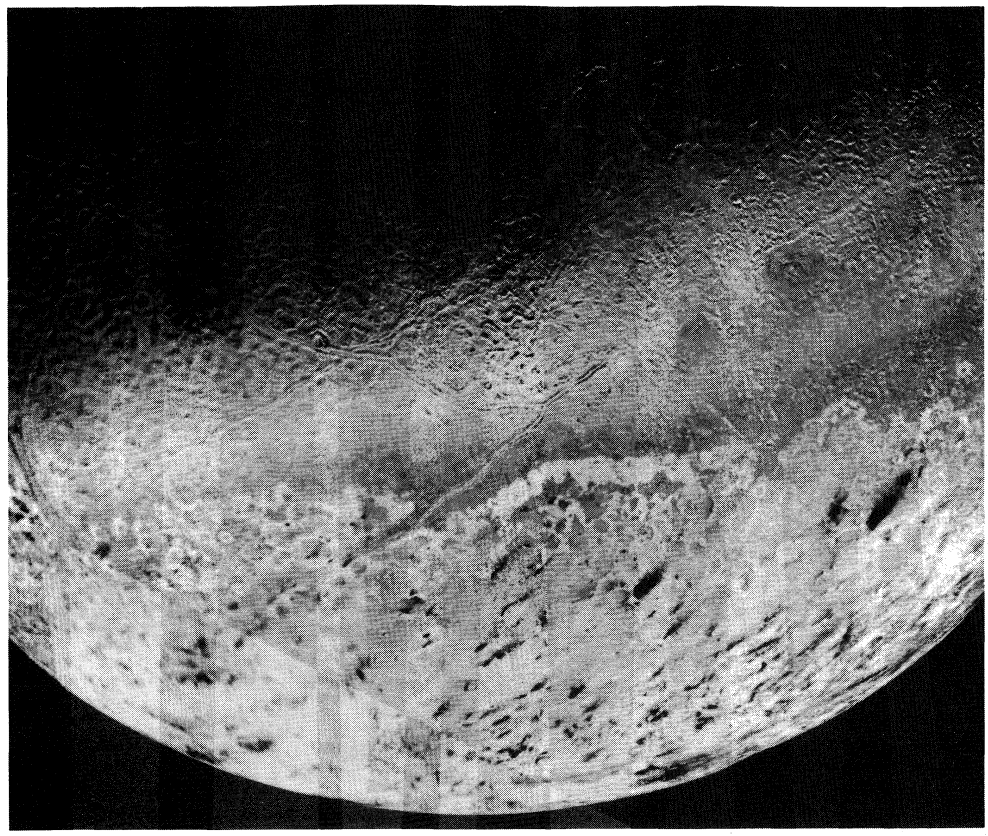


Figure 34: Close-up view of Neptune's largest satellite, Triton, produced from images transmitted by Voyager 2 during its flyby on Aug. 25, 1989.
By courtesy of the National Aeronautics and Space Administration

Sonar

Sonar is a system for detecting the presence of objects underwater by acoustical echo. "Active sonar" employs an apparatus for radiating acoustical energy to bounce off underwater objects; "passive sonar" consists merely of the passive reception of acoustical energy generated by another source.

DEVELOPMENT OF SONAR

Interest in a means of detecting underwater objects was originally aroused by the problem of icebergs, dramatized by the sinking of the *Titanic* in 1912. The first proposal was put forward by a British meteorologist, Lewis Fry Richardson, and the first successful application used in iceberg detection was made by the American radio pioneer Reginald Aubrey Fessenden. Development was stimulated by the outbreak of World War I and the impact of submarine warfare. A French physicist, Paul Langevin, played the leading role in research in which first British and then American scientists joined. A passive system of submarine detection, operational by 1916, employed a hydrophone (underwater microphone) and amplifier to pick up the noise emitted by submarine engines. By 1918 scientists had developed an active system in which a pulse of sound was transmitted and its rebounding echo used to detect a submarine even when its engines were shut down. The original term, "asdics," is said to have been derived from "*A*nti-*S*ubmarine *D*ivision-*ics*," although other explanations have been given. The name was long retained in the United Kingdom; the term sonar, from *so*und *na*vigation and *r*anging, an American acronym from World War II, is now used widely.

In the years between the wars, British and American researchers refined techniques to such a point that the Allies enjoyed a substantial advantage over Germany in World War II underwater detection. They developed two types of beams, one vertical to bounce off the seafloor for depth determination, the other horizontal and capable of detecting underwater objects and obstacles near the surface. The depth detector, or echo sounder, was in widespread navigational use in the 1930s. A major American wartime contribution was the development of a system that could rapidly scan with a narrow beam either a sector or all around, without mechanical motion of the acoustic transmitters or receivers, making possible swift and methodical search for submarines.

In the years since 1945 a great deal of work has been done in naval acoustics by the United States, the United Kingdom, and the Soviet Union, and other maritime nations, but virtually all of it is secret and so has had little influence on peaceful applications. The most important of these is the use of echo sounders to detect shoals of fish, a potential discovered in the 1930s. By 1950 specialized equipment was being installed on fishing vessels, with notable success.

The echo sounder

In the early 1970s the simple echo sounder remained the basic sonar device used by fishing fleets. A type of sonar has been developed with a horizontal scan that locates fish at a distance of one kilometre, greatly facilitating purse-seining (large net) operations by trawlers. The beam's direction can be changed by mechanical rotation of the underwater transmitting and receiving equipment, permitting a thorough search of a fair-sized area. The technique is slow, owing to the slow rate at which sound travels through water (1.5 kilometres per second).

Another problem in fish sonar is that there is as yet no technique for detecting fish close to the seabed in front of the ship, as the bottom-trawling method of fishing used by the British and other fishermen requires. There are, however, systems that accurately indicate the depth of a net towed behind the trawler, so that adjustments can be made to take advantage of any fish shoals detected on the main sonar.

Known military applications include the detection and location of submarines, control of antisubmarine weapons, sonar-equipped homing torpedoes, and mine hunting.

Detection range

Detection ranges for civil and military sonar systems vary from 100 metres to 10 kilometres. Wavelengths for the acoustic signal range between 0.5 and 30 centimetres, cor-

responding to frequencies of approximately 300 kilohertz and five kilohertz.

FUNDAMENTAL PRINCIPLES OF OPERATION

Principles of a simple sonar system. As with most radar systems (see RADAR), sonar systems generally use the transmission and reflection of a pulse of energy as their basis of operation. The arrangement in Figure 35 is typical. Individual systems vary, with some, for example, using a single transducer both for generating the acoustic waves in the water and for detecting the reflected waves. The pulses they send out are always bursts of the transmitted, or carrier, frequency, although the term sonar may also include marine seismic work, using the sound pulses generated by a small explosion.

From B.K. Gazey and D.G. Tucker, *Applied Underwater Acoustics* (1966); Pergamon Press; reprinted by permission

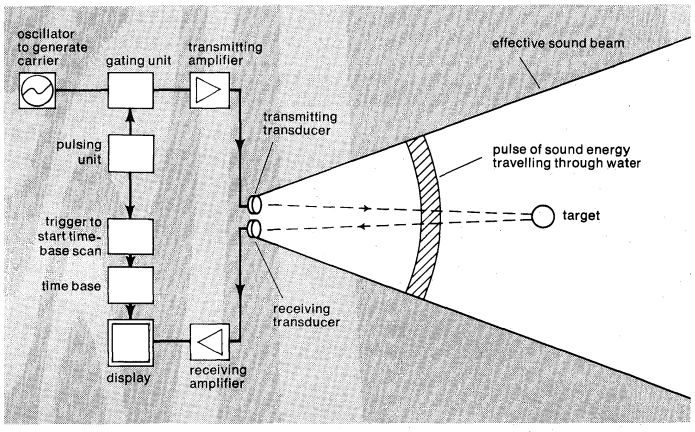


Figure 35: Schematic arrangement of a typical pulsed sonar system.

Display. The distance from the transducers to a reflecting target is indicated by the time elapsing between transmission of the pulse and the reception of an echo. The display, usually a pattern on a moving strip of paper, always includes a time base, the traverse of which is initiated by the transmitted pulse. For simple systems, the commonest display is a chemical recorder that records the received information on sensitized paper by means of a recording stylus drawn across paper. Since the echo actuates the marking mechanism, the position of the mark across the traverse indicates the range of the target. The paper is moved slowly in a direction perpendicular to the traverse of the marking stylus to make successive traverses side by side. If the range of the target in relation to the transducers does not vary, the line produced by the echo is parallel to the direction of motion of the paper. If the range changes (due to the motion of the target or ship), the line slopes with respect to the paper motion.

Resolution. The resolution of the system in range—*i.e.*, the fineness of detail it can show in a direction outward from the transmitter—improves as the transmitted pulse is made shorter. The carrier frequency may have to be increased to permit this shortening. Greater bandwidth is then required in the receiving equipment. Improved angular resolution—*i.e.*, the ability to distinguish two targets at slightly different angles—is generally obtained by making the beams narrow (more directional).

Doppler sonar. Another sonar method, Doppler detection, relies upon the relative speed of the target and the observing station to provide an indication of target speed. It employs the Doppler effect. Detection by this means requires a long pulse to give an aural recognition of tone. The received signal is mixed with a local signal of slightly different frequency, the frequencies being so arranged that mixing the two produces an audible signal. A shift of pitch between the transmitted and received signals can easily be detected by a trained ear; this method gives not only very sensitive detection but also valuable information regarding the speed of the target.

Doppler detection can be used on fast-moving ocean fish, but most fish move too slowly for the method to be of value for ordinary fishing.

The propagation of acoustic waves in water. If water did not introduce any losses, if it extended indefinitely in all directions, and if it were uniform in all respects, then the spreading of sound would be spherical. Under these conditions the intensity of the sound would decrease inversely to the square of the increasing distance from the transmitter. This relationship is usually expressed as a loss of 6 decibels per doubling of range for one-way transmission, or 12 decibels per doubling of range for the echo signal. The decibel is a logarithmic unit. If the intensity at one point is I_1 and at a farther point is I_2, then loss in decibels between the points is 10 times the logarithm of the ratio of I_1 to I_2.

Sound intensity in the sea

The sea, however, is not infinite in all directions. It is bounded by the surface and the seabed, so that spreading does not follow the inverse square law. The sound intensity, channeled between surface and bottom, falls off less rapidly than the inverse square law implies. In the extreme case, called cylindrical spreading, the loss increases only three decibels for doubling of the range for one-way transmission.

There is also some loss due to absorption of the sound energy by the water, which converts it into heat. This loss is variable; it is small at low frequencies but rises very rapidly with an increase in frequency. In fresh water, the loss expressed in decibels per kilometre is proportional to the square of the frequency and is about three at 100 kilohertz. In seawater, there is an additional loss at frequencies below one megahertz due to the dissolved salts. Below 100 kilohertz, this additional loss is approximately constant at around 15 decibels per kilometre.

Further propagation effects of importance result from variations of the velocity of sound in seawater. Under normal conditions the velocity is about 1,500 metres per second. The main causes of varying velocity are, in order of their importance, temperature, pressure due to depth, and salinity. The magnitudes of the effects cannot be expressed by any fundamental equation, though several empirical, or rule-of-thumb, equations have been proposed. Increased salinity, temperature, and depth all increase velocity over all normal ranges.

Perhaps the most serious effect of varying velocity is the refraction (bending) of the sound beam, which may cause a beam that is normally horizontal to be deflected to the sea bottom, where it will be reflected upward, then down again, and so on. Along any straight line through the transducer, therefore, there are intervals of range where detection is impossible and others where it is possible. This and similar effects are most serious in low-frequency systems since they have the greatest nominal range.

Effects of varying velocity

Another effect of varying velocity is the general scattering of the sound beam as it passes through turbulent water, leading to rapid fluctuations of signal strength.

Transducers. The purpose of transducers (strictly, electroacoustic transducers) is to convert an electrical signal into an acoustic signal, or vice versa. Three types of transducer are employed in underwater applications: magnetostrictive, piezoelectric, and electrostrictive.

Magnetostrictive. In a magnetostrictive type, a magnetic field is applied to a piece of suitable magnetic material, causing the dimension of the piece to decrease along the axis parallel to the field. When the field is alternating, application of a steady magnetic field (polarization) is necessary to give the acoustic wave the same frequency as that of the electrical signal. The magnetostrictive effect also operates in reverse: received acoustic signals cause compression of the material, altering the magnetic field, which in turn produces an electromotive force in the electrical winding.

Piezoelectric. Piezoelectric transducers use crystals that change in dimension according to the applied electric field. If the field is alternating, the crystals vibrate and radiate an acoustic wave. Conversely, if the crystals are acted on by acoustic waves, they generate an electric field. Piezoelectric materials used include quartz, ammonium dihydrogen phosphate, tourmaline, and lithium sulfate.

Electrostrictive. Electrostrictive transducers are the most widely used of the three. Materials used include barium titanate and lead zirconate. The change of dimensions

depends on the magnitude but not the polarity of the applied electric field, and so polarization is needed as with magnetostrictive transducers. Electrostrictive transducers generally have impedances of a few hundred ohms.

Transducers usually are operated at resonance—*i.e.*, the frequency at which they vibrate naturally—to get reasonable efficiency, making their frequency bandwidth rather narrow. When wide-band operation is required, the transducer is operated in such a way that its natural resonances are outside the operating frequency band. Various methods are used to reduce the sharpness of resonance as well as the cost of transducers. In one method, slices of crystal or ceramic material are alternated with plates of metal; such "sandwich" transducers are common.

It is difficult to generalize about transducer design since there are so many different kinds. In size, transducers vary from tiny probes of perhaps one millimetre square to large low-frequency elements of perhaps 10 kilograms (22 pounds) in weight. Transducers (or, more correctly, transducer arrays) are usually made up of a number of elements to achieve particular directional effects and for a larger radiating surface that can transmit the desired power into the water without cavitation (formation of bubbles). For special applications, rigid transducer arrays have been developed measuring several metres in size and weighing almost a ton. For some stationary, sea-bottom mounted sonars the array may consist of numerous separate transducers linked by cable.

Transducer arrays

INHERENT LIMITATIONS OF SONAR

Noise in water. Acoustic noise in water sets the ultimate limit to the range of detection because it eventually obscures the wanted signal. Noise is defined as power received that is not part of the desired signal and that is not produced by the transmitter. Sources of noise can be inherent, can be caused by natural phenomena, or can be produced by humans or other animals.

Inherent noise. Inherent noise results from motion of molecules, caused by heat. The colder the water the less inherent noise is received. Basically, noise intensity from this source is independent of frequency although, like all noise, it is dependent on the bandwidth of the system.

Sea-state noise. The most important noise produced by natural phenomena is that resulting from wave action. This is called "sea-state noise." Its magnitude is dependent on the height of the waves. Essentially a low-frequency noise, its power diminishes rapidly as the frequency rises, becoming negligible in comparison with thermal noise around 50 to 150 kilohertz.

Animal and man-made noises. Noise-producing aquatic animals include several species of fish and shrimp, encountered frequently in warm waters. Man-made noise, particularly that generated by the ship carrying the sonar equipment, can be more serious. Although difficult to calculate in advance, it poses the main limitation to sonar performance.

Reverberation. When the sonar system must detect small targets, a random background due to "reverberation" can make detection difficult. Reverberation is the sum of all the numerous small echoes produced by reflections from sand and stone particles on the sea bottom, minute air bubbles, and other irregularities in the water. While background noise in the sea limits the maximum range of detection, reverberation may limit performance in all ranges. Since reverberation originates from the signal transmission, it has a power level at any time interval closely related to that of the signal. Consequently, to detect small objects in areas where reverberation level is high (due, for example, to a shallow and rough sea bottom), the beams must be as narrow as possible in the relevant dimension. When reverberation is the limiting factor in detection, increasing the transmitted power does not improve detection over most of the range. Consequently, many sonar systems operate at acoustic powers of fairly low peak level.

Directivity. Sonar is most efficient if the sound energy is confined to a narrow beam on transmission and the receiving transducer responds only to sound coming from a limited angle, or cone. If the transmitter is highly directional, then the sound intensity at any distant point, for a given power source, will be much higher than if the sound had been broadcast uniformly in all directions. Fewer unwanted objects will be illuminated by the transmitted beam. On reception, a high directivity means that the noise will be reduced. If the noise is isotropic (coming almost equally from all directions) the reduction is given by the directivity index (DI). Directivity index is defined for transmission as the ratio in decibels of the intensity at a particular point in the direction of maximum transmission to that which would result at the same point if all the transmitted power were spread uniformly in all directions. On reception, it is the ratio in decibels of the output power developed by a signal in the direction of maximum response to that developed by the same signal power if it were uniformly distributed over all directions. For ordinary transducers, the DI is the same whether the transducer is used for transmission or reception.

The directivity index

TECHNICAL ADVANCES

Improved sonar techniques are leading to (and often originating from) new applications in various fields of underwater operations. Since naval requirements obviously call for longer ranges of detection, higher powers and larger transducers are being provided, augmented by the sophisticated computer technique of "signal processing," whereby low-level signals can be extracted from a noise background.

Side-scan sonar. In nonmilitary applications, one technique of considerable importance is side-scan sonar. This sonar has a narrow (about 1°) beam in the horizontal plane, but relatively wide (10° or 20°) in the vertical plane, looking sideways from the ship or towed body on which it is mounted. As the ship proceeds along its route, a map of the acoustic scattering from the seabed is built up on recording paper. With experience, these records can be interpreted in geological terms, giving in effect a map of the hardness and roughness of the seabed rocks; it reveals geological faults, sand waves, and ridges. The general arrangement is shown in Figure 36.

Existing equipment for geological applications has a range of about 1,000 metres, but a model constructed by the British National Institute of Oceanography has a range of

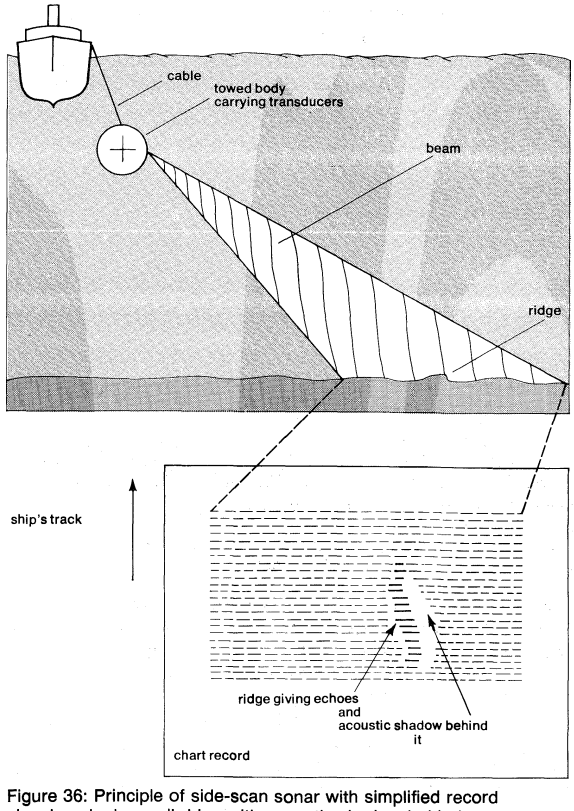


Figure 36: Principle of side-scan sonar with simplified record showing single small ridge with acoustic shadow behind.

approximately 15 kilometres working at a relatively low frequency of 6.6 kilohertz. The speed of search is high, but the results are seriously affected by refraction. Smaller, high-frequency, side-scan sonars, operating at 300 kilohertz (or even higher) with a range of about 200 metres, survey an area in finer detail. They can delineate such objects as underwater pipelines, oyster beds, and wrecks.

WPESS sonar. The growing trend toward higher resolution in sonars has led to the development of what is called within-pulse electronic-sector-scanning (WPESS) sonar. When very narrow sonar beams are used and it is desired to search a sector rapidly by swinging the beams around, mechanical beam swinging slows the rate of search, making the method unsatisfactory for fish finding and mine hunting when the ship is moving forward. In the WPESS method, a wide sector is covered by each transmitted pulse, and a narrow receiving beam is rapidly swung electronically across this sector. The beam is swung so rapidly that the whole sector width is examined for echoes before the acoustic pulse has traveled its own length through the water. In this way, the whole sector is examined on every transmitted pulse. Thus the search rate is very high, even if very narrow receiving beams are used to provide high resolution. The display is usually a cathode-ray tube giving some kind of plan view of the sector, as illustrated in Figure 37.

Applications of WPESS sonar

The use of WPESS sonar is beneficial in the study of fish behaviour, in observing the behaviour of underwater gear, and in monitoring and directing the movements of divers. For some of these applications the acoustic frequency is made high (*e.g.,* 500 kilohertz) so that high resolution (*e.g.,* 0.5° in angle, 7.5 centimetres in range) can be obtained using only small transducers. The maximum range is then relatively small, perhaps only about 60 metres on 500 kilohertz equipment, or about 200 metres on 300 kilohertz equipment. On the other hand, the system can be used equally well at lower frequencies. In early trials, 37 kilohertz was employed with a range of detection on small fish shoals of nearly 800 metres.

Because of the trend toward higher resolution, sonar can compete with optical methods for underwater viewing. Sonar makes it possible to view objects in muddy and turbid water where optical methods are ineffective. It is thus of great importance in police searches, in civil engineering (*e.g.,* river and harbour works), and in studying fish behaviour. In these applications it is essential to use a sonar that provides high angular (or lateral) resolution as well as high resolution in range. Range resolution can usually be increased by using a short pulse. Lateral resolution, however, raises difficulties.

From *Underwater Journal* (September 1970)

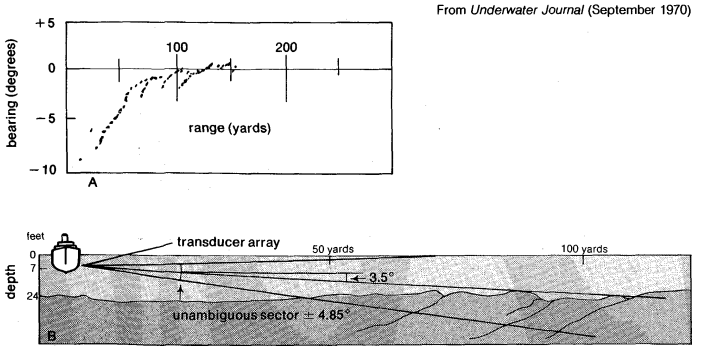


Figure 37: *Scanning sonar used with vertical "scan" in a shallow reservoir.*
(A) Bearing-versus-range display. (B) Geometry of bottom of reservoir deduced from A (vertical section).

Beamwidth considerations. Obtaining a beam of narrow angular width requires a transducer with dimensions equivalent to many wavelengths of the sonar carrier frequency. For example, a beamwidth of half a degree requires transducer dimensions of approximately 120 wavelengths. To keep the transducer small, a high frequency is necessary. Attenuation in water, however, increases rapidly with frequency. For example, at 500 kilohertz, it is in the region of 0.1 decibel per metre, while at 1 megahertz it is in the region of 0.3 decibel per metre. Consequently, the frequency cannot be increased indefinitely, and there is a limit to the reduction in size of the transducer. If, for example, the maximum range were to be approximately 50 metres, then 500 kilohertz would be a reasonable choice of frequency, the wavelength would be 0.3 centimetre, and the transducer length (l) for a 0.5° beam would be 36 centimetres.

Far-field beamwidth

The concept of the 0.5° beam is based on the idea that the measuring point is far removed from the transducer from which the beam boundaries are drawn (far-field condition). At close ranges, however, the concept of angular beamwidth is invalid; if the transducer is straight (*i.e.,* unfocused) the beam can hardly be narrower than the transducer length l, and the geometry becomes as shown in Figure 38. The far-field beamwidth θ is realized only after a range of about l^2/λ—*i.e.,* 40 metres from the transducer (λ is the wavelength in metres). This is almost the maximum range specified, so that "near-field" conditions exist almost throughout. The lateral resolution is therefore clearly measurable not in angle but in distance and is thus about 36 centimetres.

From B.K. Gazey and D.G. Tucker, *Applied Underwater Acoustics* (1966); Pergamon Press; reprinted with permission

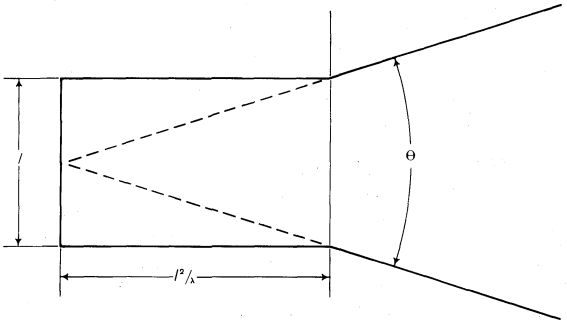


Figure 38: Geometry of beam near transducer (see text).

This situation in which the near-field conditions exist throughout almost the entire range is unsatisfactory. It can be overcome by using a system of signal processing, in which the receiving transducer is divided in half and the signals from the two half-transducers are multiplied together. Thus a target can give a signal output from the receiver only if it lies in the beams of both half-transducers. Moreover, the far-field beamwidth is one-half of that corresponding to the length of the transducer used normally and is realized at a relatively short distance from the transducer. Thus the near-field beam is extremely narrow, and lateral resolution of one-sixth of the transducer length (or less) is feasible. In the example given above, the overall length of the transducer would be 18 centimetres, and the lateral resolution is quoted as 3 centimetres, or 0.5°, whichever is the poorer. This provides an enormous advantage for the multiplicative system.

Influence of modern electronics. Older forms of electronics with thermionic tubes or discrete transistors have been expensive to produce, somewhat unreliable, and costly to maintain. Because of the revolution in electronics that brought forth microelectronics, or integrated-circuit technology, however, it is now possible for a system to be sophisticated and complex and, at the same time, relatively inexpensive and reliable. This change in electronics has influenced sonar design, and more refined and useful sonars have gone into commercial production.

Use of integrated circuits

The microelectronic units that first became available were of the digital type; *i.e.,* they represented circuits that had switching, or on-off, functions of the type used in digital computers. Several types of analog circuits later became available in microelectronic form, and it is possible to make even a complex analog system such as the WPESS sonar on this basis. (D.G.T./Ed.)

Instruments for monitoring industrial processes

Instruments that monitor variables in industrial processes originated with very simple devices that provide a measure of temperature, pressure, fluid level, and density. As instrumentation and processes developed, it became necessary

to measure increasingly complex chemical and physical properties until today there is an almost endless variety of instruments capable of monitoring process variables and leading to automatic operation. These automation systems increasingly tend to involve the digital computer as an integral component. In almost all instances the monitoring instrumentation is based upon some principle and technique thoroughly developed and previously used in a laboratory.

Monitoring instrumentation is widely used in the production and process industries. Process industries are those that transform raw materials into various chemical products. The primary variables monitored in the process industries are chemical, physical, and environmental properties. The basic chemical property monitored is composition. The physical properties monitored are turbidity, viscosity, density, and flow rate. The environmental properties monitored are temperature, pressure, and humidity.

Instrumentation in production industries

Production industries manufacture a physical product, such as machined parts, plastic items, and electronic components. These industries make wide use of modern instrumentation, although not to the same extent as the chemical process industries. Production applications tend to be more limited, with fewer variables monitored and fewer instruments used. There are two major reasons for this difference. First, the number of variables in production industries is less than in the process industries. Second, because of the sheer number and variety of objects produced, instrumentation is inherently more difficult to apply than in the process industries, which handle large volumes of materials such as liquids and gases having many properties in common.

Nevertheless, the production industries employ instrumentation to monitor such variables as physical dimensions, weight, optical properties, electrical properties, and radiation characteristics. Many of the instrumentation systems in the production industries are highly individual, used perhaps for only one specific item.

MONITORING CHEMICAL PROPERTIES

Chemical and petroleum processing plants play a large role in industry throughout the world. Since their processes involve the continuous flow of large volumes of materials through different steps, they are most amenable to monitoring and control by instrumentation. Indeed, such methods have been used since the beginning of the 20th century. For many years, however, only single-function instruments that measured temperature, pressure, or flow, and actuated simple control mechanisms, were used. By 1945 it had become clear that further progress depended upon development of sensors that would continuously analyze the composition of a given stream of chemicals.

Although automatic analytical instruments used in industry embody principles of laboratory instruments, the industrial devices look quite different. A production instrument must operate continuously over long periods, often in a harsh environment; hence, the instrument must be much more rugged than the laboratory model. Of the many laboratory principles employed in the process analytical instruments, the three in widest use today are refractometry, absorption spectroscopy, and gas chromatography.

Refractometers. Refraction is a term applied to the bending of a light beam as it passes from one material to another; for example, from air to water. Essentially, it is a measure of the difference in the velocity of light in the two materials. This measurement is usually obtained by making use of Snell's law, which gives the refractive index as the sine of the incident angle (the angle at which the light strikes the material) divided by the sine of the refracted angle (the angle at which the light leaves the material).

The refractive index, like molecular weight, density, and boiling point, is a characteristic physical property of matter and has long been used analytically in the laboratory. Since each chemical material has its own specific refractive index, the variation found in mixtures containing different proportions of two compounds becomes indicative of the composition. As a measure of chemical composition the refractive index can generally be applied with assurance only to mixtures of two components (called binary mixtures); measurement becomes ambiguous when more components are present. As an exception, there are instances in which a pseudobinary mixture occurs; the components separate into two distinct groups with each group behaving as one component.

Binary and pseudobinary mixtures

Measurement of the refractive index is simple and can be done to a high degree of sensitivity. It is possible to determine a change of one unit in the sixth decimal place of the refractive index with great accuracy. This was one of the first laboratory methods used by the developers of process monitors. The first such monitor appeared in 1952. To provide high sensitivity and especially to compensate for temperature changes (to which the refractive index is very sensitive), these first instruments operated on the differential cell principle, in which the difference in index between the sample in one part of the cell and a reference material in the other is measured. Early applications involved controlling fractionation of a complex hydrocarbon mixture by choosing the sampling point so that it behaved as a pseudobinary mixture. By 1965 most of these instruments had been replaced by the more powerful chromatographic analyzer, which completely analyzes the process mixture, permitting the use of computers and control techniques that earlier were impossible.

The critical angle refractometer

The critical angle refractometer appeared in 1962. The refractive index is measured by finding the angle at which a light beam striking the point of contact between a glass surface and the solution (or sample) is totally reflected. It is known as the critical angle, since rays at lesser angles permeate the liquid; at greater angles, they are totally reflected. The mechanism for measuring the critical angle is quite simple. A circular beam of light is transmitted to a prism face in contact with a solution, which divides it into a totally reflected light portion and a dark portion, depending on the critical angle value of the solution. As the solution composition changes, so does the dark portion. The position of the dividing line between light and dark in the returned beam is maintained by an automatic motor system in the control electronics and so becomes a measure of critical angle.

Although the critical angle refractometer is less sensitive than the differential instrument (by a factor of 100), the former requires only a surface reflection: no transmission of light through the solution is necessary. As a consequence, dark solutions and suspensions, as well as clear solutions, can be measured directly in the process line. There is no need to use a separate sampling system for the instrument.

The characteristics of the critical angle refractometer make it uniquely suited for the food industry. Sugar solutions are ideal candidates for measurement by refractometers because they are essentially a mixture of two pure substances, sugar and water. In the manufacture of many carbonated beverages a refractometer is used to control preparation of the sugar-water base to within 0.05 percent of the desired concentration.

Infrared analyzers. That part of the electromagnetic spectrum lying between the visible and microwave regions is broadly considered the infrared. The region of greatest practical importance for infrared instrumentation occurs between 2 and 15 micrometres (0.00008 and 0.0006 inch) in wavelength, where the strong absorption occurs at specific characteristic wavelengths for most organic and inorganic substances. These spectra are unique to each substance, because the frequency of the absorption bands is determined by the position and configurational relationships of the vibrating atoms in a molecule. Since the amount of energy absorbed by a material at any of its characteristic frequencies is a measure of its concentration in a mixture, a very sensitive and powerful technique is provided for the analysis of mixtures. Infrared instrumentation and these methods are widely used in analytical laboratories.

For further discussion of infrared spectometers, see the article ANALYSIS AND MEASUREMENT: *Spectroscopy.*

Chromatographic analyzer. The process chromatograph, which appeared about 1955, was the first analytical instrument capable of a complete compositional analysis.

Uses of the analyzer range from the simple monitoring of one component in a mixture to providing a computerized output analysis for 50–100 components in a process stream.

Gas chromatography is a procedure by which a volatile mixture is separated into its components by being injected into a moving, inert gas phase passing over a stationary absorbent bed in a tubing. The bed may be solid absorbent particles, or those particles coated with an involatile liquid, or the inner wall surfaces of a narrow capillary tube (0.25 millimetre in diameter) coated with the liquid. Separation of the sample components occurs because of the difference in their absorptivity on the surfaces of the absorbent bed. The separated components pass from the tubing into a detector, where their relative concentration is measured. For further discussion of the principles involved see ANALYSIS AND MEASUREMENT: *Chromatography.*

The process chromatograph

Fundamentally, a process chromatograph resembles its laboratory counterpart more closely than do most plant instruments. The differences result from requirements imposed by the plant environment and by automatic operation. The process instrument usually is divided into two distinct units—an analyzer unit located quite near the process stream and a programmer unit that can be placed where needed. The analyzer must obtain its own sample, analyze it, and process the detector information into some useful data. Depending on the intended use, these data are then delivered to either a recorder, a controller, or a computer. The instrument must repeat this cycle, sometimes as often as once each minute, for many months, without interruption but with a high degree of reliability and fast response.

Process chromatographs are used in many chemical plant operations; they were first successfully applied to the analysis of light hydrocarbons in petroleum refineries. The use of these instruments has expanded to many other areas, such as air pollution monitoring and control of combustion processes.

Measurement of pH. The pH is a measure of the degree of acidity or alkalinity of a solution. Almost every industrially important product has a characteristic pH value. It must be measured and controlled if it affects the efficiency of the process or product uniformity. This property may also be measured for corrosion inhibition or to control the pH value of discharge from an industrial plant into rivers and sewers.

Glass electrode for pH measurement

The basic sensor used for measurement of pH, a glass electrode or conductor, contains a thin glass membrane that makes it sensitive to the concentration of hydrogen ions (electrically charged particles, H^+) in a solution. Since the early part of the 20th century the glass electrode has been accepted as the standard for pH measurements. The glass electrode can be built to withstand the high temperatures and pressures of many plant environments. An acid dissolved in water produces many positively charged hydrogen ions (H^+) in the solution, whereas a base produces many negatively charged hydroxide ions (OH^-). When the number of hydrogen ions equals the number of hydroxide ions, the solution is neutral, like pure water, and its pH is seven. If there is an excess of hydrogen ions the solution is acidic (the pH is less than seven); if there is an excess of hydroxide ions the solution is basic (the pH is greater than seven). In order to measure pH, the glass electrode, which is connected to a reference electrode, is immersed in the sample solution, whereupon an electric voltage proportional to the hydrogen ion concentration results. This voltage can be displayed directly on a scale calibrated to read a pH value.

MONITORING PHYSICAL PROPERTIES

Turbidity. The analytical devices for the determination of turbidity make use of visible light—the portion of the electromagnetic spectrum that lies between 400 and 700 nanometres. The principles of analysis for the visible region of the spectrum are much the same as those for the infrared and ultraviolet regions except for the wavelength of light used.

The turbidimeter

In a typical turbidimeter light from a tungsten source is passed through a filter that selects light of a desired wavelength band for the analysis. The selected light is then passed through the process material, which absorbs it in proportion to its turbidity. The light that is not absorbed but transmitted is detected by a phototube in the detector compartment and electrically amplified. The amplified signal drives both a direct-reading meter on the instrument panel and a recorder. The measurement of both colour and turbidity may be accomplished by the same instrument: the colour is measured by absorption at a specific wavelength, and turbidity is measured by the absorption of light at all wavelengths.

Density. Several principles are employed in instruments that measure density (the ratio of mass to volume). Gas densities are generally determined either by measuring the buoyant effect of the gas or by comparing the rotation of a turbine wheel in a standard gas with the rotation in the gas to be measured. The densities of liquids can be measured by an automatic hydrometer (see below) or by measurement of the extent to which the liquid absorbs gamma radiation; since the pressure of a column of liquid depends upon the density of the liquid, density may also be determined by measuring the pressure exerted by a given height of liquid.

The hydrometer

The most widely used instrument for both gas and liquid density determination is the hydrometer. The operation of a hydrometer is based on Archimedes' law, which states that a body in a fluid is buoyed up by a force equal to the weight of the volume of fluid it displaces. By measuring the buoyant force acting upon an object of known volume immersed in a gas or liquid, one can determine the weight of that volume of the gas or liquid and thus its density.

The gas density balance is an example of the application of this principle. This device consists of a light dumbbell, the two balls of which are hollow, suspended in an atmosphere of the gas being measured. One ball is punctured and therefore admits the gas; it is thus free of buoyancy effects. The other ball is sealed and will tend to rise and fall to different degrees as it is buoyed by gases of differing densities. The force necessary to keep the dumbbell suspension from tilting yields a measurement of the force of buoyancy; thus, the density of the gas may be determined.

An example of the application of the hydrometer technique to liquid density measurement employs a different method, although the general principles involved are the same. In this method a probe is inserted into the liquid to be measured. Its tip contains a ball-plummet, which is suspended at an exact position under the pole of an electromagnet. The measurement of the electromagnetic force necessary to suspend the ball in the liquid reflects the force of buoyancy and thus the density of the liquid. Hydrometers of this type have been used in the formulation of liquid synthetic latex to terminate the reaction at a suitable value of density. Another typical use is in the pumping station of an oil pipeline in order to measure the density of the oil. This instrument is particularly useful in observing the passage of the interface dividing two batches of oil in transit in the pipeline.

Viscosity. The viscosity of a fluid—the degree to which it resists flow under applied force—can be measured continuously by utilizing any one of several physical effects: the damping of mechanical oscillations by the fluid, the drag of the fluid on a rotating cylinder, the drag of a float against a constant flow of liquid, or rate of flow in a confined area.

Of these methods, the measurement of the damping of mechanical oscillations by the fluid is most widely used. A small blade of steel, inserted into the process stream, is made to oscillate by an impulse signal. The oscillation is affected by the viscosity of the liquid in the same way as the ringing of a bell in a liquid: the more viscous the liquid, the faster the ring dies away. The pulse is sent to the probe many times per second from an electronic oscillator through a coaxial cable. The rate of decay of the oscillation is measured, and from this the damping effect of the fluid can be determined. This damping effect is a function of both the viscosity and the density of the fluid, so that a knowledge of the density allows a measurement of the viscosity.

This method of measurement has many applications in

the petroleum and chemical fields. These include automatic blending of lubrication oil and quality control of soaps, detergents, and margarines. Applications in other industries include quality control in the production of starches, glues, adhesives, and paper and textile coatings.

Flow rate. In the chemical industry flow rate measurement is essential in controlling all phases of processing and in determining the material balance for processing units. Once manufactured, the transmission of materials through pipelines between distant places calls for an accurate measurement of flow rate. A multiplicity of techniques is used in this measurement. Flow rate may be determined by measuring the change in pressure caused by either a constriction in a pipe or the insertion of a disk with an orifice into the flow stream. Measuring the impact pressure upon a probe inserted into the process stream will yield the flow rate, as will measuring the change in pressure resulting from a change in the direction of the stream. It is also possible to derive the flow rate by measuring the change in the velocity of sound as it passes through the material.

The turbine flowmeter and the orifice meter

One of the most widely used methods is the turbine flowmeter. A turbine rotor is allowed to rotate freely in the moving fluid, and its rotation causes a sudden distortion in the field of a small, powerful magnet located in a sensor unit outside the pipe. This distortion generates an alternating-current voltage that is transmitted to a small computer. The computer analyzes this information and calculates and displays the flow rate.

Another common flow rate measuring device is the orifice meter. A plate with a circular orifice at the centre is inserted into the process stream, causing the fluid as it passes through the orifice to increase in velocity and correspondingly decrease in pressure. A differential-pressure measuring device measures the fluid pressure just before and just beyond the orifice. Knowledge of this differential pressure allows calculation of flow rate. This type of flowmeter is the most widely used because it is simple and it has been long established in plant processes.

These devices measure the volume-flow rate. This knowledge is useful in monitoring, for instance, the blending of two fluids the densities of which are known, such as gasoline and methylcyclopentadienyl manganese tricarbonyl. In other cases, such as that in which a large quantity of raw material is being transmitted by pipeline and sold by weight, determination of the mass-flow rate is vital. This may be found by adding to a volume flowmeter a device that measures the density of the material and then calculating mass flow from these two measurements.

There are also flowmeters that directly measure the mass-flow rate. One of these utilizes two turbines in the flow stream, the first of which, driven at a constant speed, acts as an impeller and imparts a certain velocity to the fluid, depending on the fluid's mass. The second turbine, located downstream, is adjusted to slow the flow to its original rate; in doing so it receives a torque, or turning force, proportional to the force of the flow (angular momentum). The turbine deflects a spring at an angle proportional to the torque exerted upon it by the fluid. The result is a very accurate and direct measure of the mass flow.

MEASURING ENVIRONMENTAL PROPERTIES

Temperature and pressure. The accurate measurement of temperature is important in many chemical processes to avoid harm to materials and equipment resulting from temperatures outside a specified range. Its determination is also necessary in situations in which it is a needed variable in the computation of properties such as pressure, viscosity, or density. Methods of sensing temperature depend upon the measurement of the changes caused by temperature variations. Many devices, such as the familiar glass-stem thermometer, measure the change in volume of a substance, such as mercury, caused by a change in temperature (see above).

Importance of pressure measurement in industry

Pressure, like temperature, is a variable that must be measured accurately in industry, particularly in the chemical industry. Determination of pressure is vital, for instance, in the control of hydrogenation (addition of hydrogen) and distillation in petroleum processing. Again, like temperature, pressure is a variable needed in the calculation of other properties. Pressure-measuring devices vary with the range over which they are meant to be used. In the vacuum range, gas pressure is detected by measuring the current generated due to ionization of the gas or by measuring the thermal conductivity of the rarified gas. Pressures in this region are also calculated by compressing a known volume of the gas until it reaches a fixed pressure. When the new volume is measured, the original pressure can be computed by use of Boyle's law, which states that the product of the original volume and original pressure is equal to the product of the new volume and new pressure.

In the atmospheric pressure range and above, elastic pressure elements are widely used; they measure the expansion caused by pressure. While some devices measure the expansion of a diaphragm or a bellows, the most commonly used industrial sensor is the so-called Bourdon tube, consisting of a tube in the form of a 250° arc. The process pressure is connected to the fixed socket end of the tube, while the tip end is sealed and connected via a series of links and gears to a pointer. Because of the difference between its inside and outside radii, the Bourdon tube tends to straighten when pressure is applied. The resulting motion of the sealed tip is a function of this pressure, and thus the position of the pointer yields a measure of the process pressure.

Bourdon tube

A device that has many applications in this pressure range is the strain gauge, which is based upon the fact that metallic conductors subjected to strain exhibit a corresponding change in electrical resistance. There are many types of strain gauge, but all are constructed so that the process pressure causes a strain, and thus a change in electrical resistance, which is measured for a visual display.

Moisture. Because of the industrial and scientific importance of measuring water, and because of water's almost universal presence, analysis for water content is a widespread necessity. The presence of water in a process stream may be either a necessary or an unwanted factor. A definite concentration may be required for the production of desired properties, while, on the other hand, water can cause corrosion and other undesirable side effects in many processes. Moreover, measurements of water as a by-product of a reaction indicate the extent of reaction.

Water possesses relatively unusual properties. Many instruments with widespread applications, such as those based on infrared absorption, mass spectrometry, refractive index, and vapour chromatography, can be used for the monitoring of water. Certain methods, however, have been developed for use exclusively in the detection and measurement of water.

Electrolytic measurement of water

The electrolytic approach (involving electrolysis, or the production of chemical changes by an electric current) involves an electrolytic cell, containing a water-absorbing substance such as phosphorus pentoxide, through which a gas sample is passed at a constant velocity. Conditions are adjusted to ensure fairly complete absorption; the electrolytic decomposition of the water in the sample and the current resulting from electrolysis give a measure of the amount of water electrolyzed. With this quantity and that of the volume of gas flowing through the cell, the concentration of water in the sample can be calculated.

Applications of water-measuring devices of this type include the monitoring of air, ethylene, acetylene, and carbon dioxide in refining and allied petrochemical processing and the monitoring of natural gas pipelines for moisture content in order to prevent the formation of explosive hydrates. Instruments for different uses may be portable or designed for fixed installation and severe environmental exposure.

Water vapour in gases is analyzed by observing the change in frequency of specially coated oscillating quartz crystals (the piezoelectric effect). Water vapour is first adsorbed and then desorbed, causing a change in mass of the coating, which affects the frequency of vibration. The method is extremely sensitive and can detect the presence of as little as 0.1 part per 1,000,000 of water.

Measurement of moisture in materials moving on a track or web can best be done using the selective absorption of an infrared analyzer. The moisture analyzer is mounted in a rugged transversing system over the web to scan the

moving material on it. By measuring an average of the absorption of near-infrared energy by the moving sample, the analyzer provides a continuous, non-contacting measurement of moisture. This system can be used in a wide range of applications. It is capable of measuring webs of material traveling at high speed; conveyed particulate and granular material can be measured at moderate to high speeds. Since the sensing head is mounted several centimetres from the surface being measured, nothing is introduced that interferes with the process. Typical applications include the measurement of water in paper, cellophane, film, instant coffee, tea, soap, detergents, chemicals, rubbers, sand, clay, tobacco, and fibres.

Water content in sheet materials such as paper can be monitored by measuring capacitance—*i.e.*, the ability of the material to store an electrical charge when it is inserted between the plates of a capacitor. The method is sensitive to small quantities because water introduces high capacitance into the sheeting material's low capacitance; thus, the presence of water in a sample causes its capacitance to increase sharply. Devices that operate on the principle of electrical conductivity measurement have similar applications, particularly in the manufacture of food and such products as tobacco and cottonseed meal.

PRODUCTION-LINE MONITORING

Mechanical properties. The need for measurement and control of the dimensions of objects in production industries is obvious. Extremely small differences in product size can be critical for machined items, and occasionally accuracies in the range of 0.000025 millimetre must be maintained for such products as optical devices. Measurements in this range are difficult or impossible to achieve through the use of traditional devices, such as the micrometer; for this reason, highly accurate new instrumentation has been developed.

The capacitive sensor, used in a wide range of automatic inspection machines, can discriminate down to 0.00013 millimetre. This instrument uses the two plates of an electrical capacitor to gauge the size of the measured part. The capacitor forms part of an oscillating circuit, and changes in the spacing of the plates cause a change in the output frequency of the oscillator. The changes in dimensions can be derived by measuring this change in frequency.

Another extremely accurate method is the differential-transformer sensor. It depends upon the fact that the electromagnetic coupling between two coils (the electrical effects of two coils on each other due to their proximity) is influenced by the presence of a magnetic core common to both coils. If this core is moved while an alternating current is passed in one coil, extremely sensitive coupling can be obtained with the second coil, coupling so sensitive that changes in core movement of 0.00005 millimetre may be detected.

The electro-optical gauging system

The dimensions of rapidly moving items on a production line can be measured with an electro-optical gauging system. The system is essentially a sampling-type tracking instrument that gauges the dimensions of an object by sensing the optical discontinuity between the object and its background. By continually monitoring objects, the instrument senses the relative positions of each side of an object during a scan cycle. It then produces voltage signals that directly indicate the size of the object. A deviation from a preset value is indicated on a displacement meter. This instrument has found applications in measuring the formation of wire, the insulation of cable, and the extrusion of plastic and metal rods. The system measures objects as small as 0.05 millimetre or as large as 15 metres with an accuracy of 0.2 percent.

Nucleonic gauging

In production situations in which thickness of the product is important (*e.g.*, in the production of sheets of material), nucleonic gauging is often used. In this method nuclear radiation is beamed through the product to be measured, and an automatic device measures the amount of radiation transmitted. For materials of constant density, this transmission gives an indication of the thickness. The thickness of a coating on a base material may also be measured by this method through the use of two gauges: the first examines the uncoated material and acts as a standard, while the second measures the coated material. The difference between the two gives a measure of the thickness of the coating. These techniques are applied especially to the manufacture of metal sheet products and the coating of roll sheet products with plastics and other metals.

Figure 39 shows a computer-controlled rolling mill. The computer compares the actual input thickness with the desired and actual output thickness of the strip, compares them to stored data, and generates correction signals to reset the mill rolls in the event of out-of-tolerance conditions.

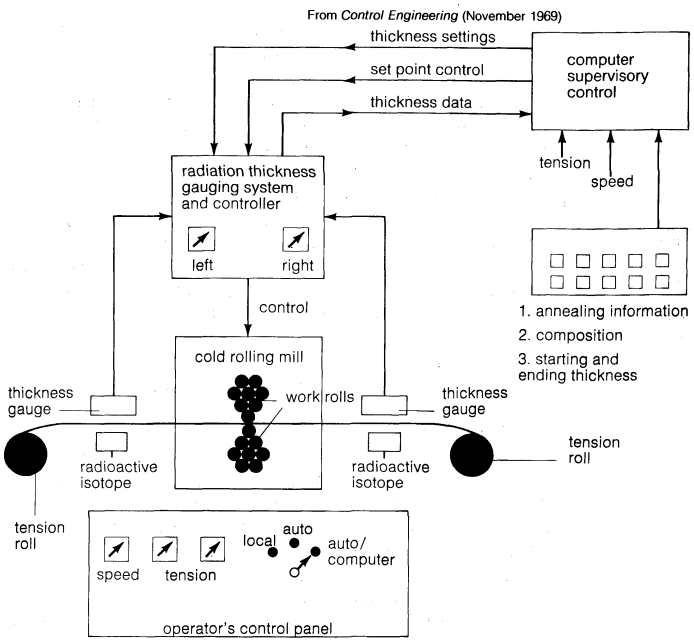


Figure 39: Computer-controlled rolling mill.

Weight measurement is a problem that often occurs in industry. Although the principle of the lever as a means of weighing is attributed to Archimedes (3rd century BC), it was not until the 15th century that something new was added to these fundamentals by Leonardo da Vinci. The basic principle of the lever platform scale, the multiple lever scale, and an indicating scale are all attributed to the genius of Leonardo. Since the 17th century no fundamental changes in the principle of lever scales or spring balances have occurred. It was not until the 19th century, with the advent of the hydraulic-load cell, and, more recently, the discovery and development of the strain gauge, that there have been many significant changes in weighing principles and devices.

Electronic instrumentation is used rather than traditional scales or balances in situations in which speed is an important factor (weighing in roughly $^1/_{10}$ second can be accomplished). Electronic instrumentation also offers the benefits of a measurement output that can be easily displayed, stored, or utilized in some form of automatic inspection or control. There are three principal methods in use. These are the load cell, in which the weight of the object causes a strain in a supporting member, which is then measured; the pressductor, in which the stress in the supporting member is measured magnetically; and the force balance, in which the weight of the object is supported by an electromagnetically generated equal and opposite force.

The load cell and pressductor

The load cell is a highly developed device working on the strain principle. A column is compressed under the applied weight and the resulting strain is sensed by a wire resistance strain gauge. This gauge is a fine wire, which, when subjected to strain, undergoes a change in resistance. A measurement of this change yields a measure of the weight causing the strain.

The pressductor is a block of magnetic laminations cemented together with four holes drilled through them. A wire carrying alternating current is threaded through two of these holes, and a pickup wire connected to a voltage meter is threaded through the other two. No voltage is

detected by this pickup wire in the absence of stress, but stress caused by weight on the supporting column causes a voltage to be developed that is proportional to that weight.

Both the load cell and the pressductor are used in measuring large weights—from one kilogram to many tons. If the very rapid weighing of lighter objects is desired, the force balance is generally used. The object to be weighed is so placed as to deflect a moving coil by its weight. This coil is a conductor carrying a current in a steady magnetic field, and its deflection from rest position is detected and corrected by a balancing current. The current needed is proportional to the deflecting weight, which can thus be determined.

In many manufacturing operations and in process control systems, weight measurement is the most direct and accurate method of control. In the handling of liquids and solids, weight control is relatively independent of process characteristics, physical location, time, atmospheric conditions, and temperature. Units of weight are easily converted to monetary values, thus facilitating economic control. In handling difficult materials, such as chlorine, liquefied petroleum gases, and similar substances, weighing provides a safe and convenient method of control without contact with the product.

Optical properties. Optical properties of products, such as their colour or the degree to which they reflect or transmit light, are important variables that must be controlled in industrial production. Quality control of products such as textiles or optical devices depends upon these measurements. The measurement of colour can also be used in the automatic mechanical sorting of items.

Visible colour depends upon the three primary colours: red, yellow, and blue. The basic method for the determination of the colour of an object makes use of three photocells. One-third of the light reflected from the object is diverted to each cell, each equipped with a filter to receive only red, blue, or yellow light, or some equivalent, specific wavelengths related to total colour. Each cell transmits a voltage proportional to the light received, and the ratio of these three voltages is automatically measured. Since each colour category has its own characteristic proportion of these three colours, total colour can be ascertained from this voltage ratio.

Reflected or transmitted light

Measurement of reflected or transmitted light is made if the lightness, darkness, or degree of transparency of a product is important. The basic method consists of shining light onto or through the measured object. The reflected or transmitted light is collected by one or more photocells set to receive all wavelengths of light. The ratio of the energy received by the photocells to the known energy of the light source is a measure of the reflectiveness or transmissivity of the object. Knowledge of this property is useful, for example, in the glass and optical industries, in which the transparency of the object is important, and in production situations such as that of the paper industry, in which defects show up as dark or light spots.

Radiation characteristics. Instruments that measure radiation characteristics are used in the production industries for two major purposes: for the determination of temperature, which is done by instruments that measure infrared emission, and for the detection of product defects, such as faulty wiring, poor internal adhesive bonding, and stress defects, which are found through use of instruments that measure infrared emission, X-ray transmissivity, or sound waves. In most instances these instruments are employed as continuous monitors during high-speed production. Without such instrumentation this detection is impossible or involves destruction of the product.

The radiation pyrometer

A versatile device used in production industries to determine and monitor temperatures up to 3,900° C is the radiation pyrometer. The infrared radiation emitted by the process is focused upon a detector, which emits a voltage proportional to the intensity of the radiation. This voltage can be converted to a measure of the temperature of the measured object, because temperature is a function of the total radiation emitted. The radiation pyrometer is employed in the detection and control of furnace temperature and in various other industrial situations, especially in the metal processing industry.

The radiation pyrometer and similar infrared detectors are also used in production testing for defects, which are indicated by undesirable production of internal heat. An example occurs in the electronics production industries, in which the radiation pyrometer is used in the evaluation of electrical performance through the detection of electrical overstress. This overstress causes excessive power dissipation and thermal interaction, which results in heating of the elements involved. The infrared detector senses this heating as an increase in infrared radiation and thus can be used to pinpoint the defective elements.

A simple thermocouple can be used to continuously measure the temperature of moving strip products such as steel. In galvanizing, the control of this temperature is important just before the sheet moves into a bath to be coated with zinc. The sensor employs a thermocouple embedded in a magnetic block that grips the surface of the moving sheet metal in a free-sliding manner.

In some industries X rays are used to detect product defects. Standard X-ray techniques detect internal configurations, and in situations in which examination of a thin internal section of a multilayered sample is needed, axial transverse laminography can be used. In this method the sample and film are synchronously rotated during X-ray examination, with the result that one plane remains in sharp focus. This plane can be varied to allow inspection of any section of the sample.

Production testing also may use sound waves rather than electromagnetic radiation. Ultrasonic energy (sound waves at frequencies above the limits of human hearing) is beamed into the material, and the reflection, or echo, is detected. Defects are indicated by the appearance of certain characteristic echo patterns; the method is used, for instance, in the detection of poor adhesive bonding or residual stress.

Computerized systems. Computers are used at every stage in the production industries. Though the major area of applicability is in production testing and inspection, they are increasingly used to control virtually all phases of production. An example of computer use in production testing is found in an automatic test system for electronic and electrical equipment and components. The computer automatically carries out 16 different tests on the unit, including frequency, voltage, and resistance tests. The computer controls the sequence of these tests and evaluates and records the results.

Computerized inspection allows manipulation of situations that might otherwise occur too quickly for control. A computer, for instance, is used to inspect sheet metal traveling at 21 metres per second. The presence of holes or cracked edges is detected by electronic sensors, and the thickness is measured with an X-ray gauge. This information is fed into the computer and stored. The computer automatically measures the total length of sheet metal that passes and commands the system to stop and cut the sheet at a preset length. It then prints out a summary of all the measured characteristics of the sheet; this summary is attached to the coil of sheet metal and serves as a quality-control document throughout the manufacturing process. The computer also provides for the operator a visual display of each of the parameters being measured, with a running total of all the measured characteristics. If a defect such as excessive thickness is detected, the operator orders the computer to rewind the sheet to the spot at which the fault began, cut out the flawed portion, rejoin the ends, and recommence operation.

Production control by computer

In the sheet metal industry, as in many others, computers are used to control many phases of production, not just production testing. Computers in sheet metal processing track the product through the process, operate the production machinery, control the width and thickness of the metal, regulate furnace temperatures, test the product, and, finally, route it for storage or other processing.

Other significant instruments

BIOMEDICAL INSTRUMENTATION

Throughout history physicians and physiologists have sought information about the workings of the body through

the use of their own sense organs. To delve into processes too subtle or too rapid for the human senses, these scientists today utilize some form of transducer to extend their senses and to convert the physiological phenomenon into an electrical signal for measurement.

Examples of biomedical instrumentation are many and varied. Some instruments, such as the electrocardiograph and the electroencephalograph (see below), detect electrical signals generated by the body itself. Others, such as the X-ray machine, operate by beaming radiation into the body. Still others use a variety of transducers to measure physiological properties such as lung function and blood-flow rate. If measurements cannot feasibly be made on the body, samples are analyzed remotely. The variety of such instrumentation is immense, so only a few examples can be discussed here.

Blood-flow rate. The measurement of blood-flow velocity provides important information about cardiovascular function. A number of instruments are used for this measurement.

The thermistor velocity sensor operates by means of a small heat-sensitive thermistor sensor inserted into a catheter tip and positioned in the bloodstream. (A thermistor is an electrical resistor made of a material whose resistance varies sharply in a known manner with temperature.) The rate at which heat is lost from this tip measures the flow rate, because the thermistor will cool faster if the flow velocity is greater. The temperature of the thermistor is measured, and power is supplied so as to maintain it at some constant temperature. Measurement of this power gives an indication of the flow velocity at the point of measurement.

The electromagnetic flowmeter

The most widely used device for measuring the rate of pulsatile (*i.e.,* caused by the pumping action of the heart) blood flow is the electromagnetic flowmeter. The principle behind its operation is that blood, or any other conductive material, generates an electrical voltage when moving through a magnetic field, and this voltage is proportional to the flow rate. There are two major methods of electromagnetic flow determination. In one, the artery or vein is surgically exposed, and an electromagnet is fastened around it. The voltage generated by the passage of the blood through the centre of the electromagnet is measured and the flow rate ascertained. In another method, a catheter probe is inserted into the bloodstream. The probe tip contains an axially wound coil electromagnet, along with two electrodes, to measure the voltage produced as the blood flows around the probe. The first method is used primarily in research (on dead tissue) or in surgical situations, whereas the second is almost exclusively employed in diagnosis.

Heart and brain condition. Two of the most important phenomena monitored by biomedical instruments are heart and brain condition. While a number of instruments and operating principles exist in this area, the two important measuring instruments operate by detecting electric potentials generated within the body.

Experimentation has shown that the surface of the exposed brain generates potential differences of between 300 and 600 microvolts, though the attenuating layer of fluid, bone, skin, and hair reduces this to a maximum of 200 microvolts. These potentials fluctuate rhythmically, and, though not fully understood, may appear in certain patterns that are signs of abnormality. The electroencephalograph measures these patterns through the use of electrodes placed in pairs on the surface of the skull. Each pair of electrodes sends a signal to one recording channel of the electroencephalograph; this signal consists of the voltage difference between the pair. The rhythmic fluctuation of this potential difference is shown as peaks and troughs on a line graph by the recording channel. Any number of electrode pairs and recording channels may be used, but eight-channel instruments are the most common. Electroencephalograms are commonly used in diagnosis, and expert interpretation can find brain disturbances and pin down their exact site long before they are reflected in the patient's behaviour or clinical status. They are also part of the standard equipment of operating theatres. A lessening of the blood supply to the brain, for instance, will show up as a change in the electroencephalogram of an anesthetized patient long before any other outward signs are easily noticeable.

The electrocardiograph

The electrocardiograph operates in a somewhat similar way to measure the electrical impulses emanating from the heart. These impulses originate at the heart's pacemaker (the body part that regulates the heart's rhythmic activity) and move as a wave through the heart, initiating its muscular action. Many qualities of this wave, such as the pattern of initiation and passage through the heart and the general direction of its travel, can be important indicators of heart damage and disease. The electrocardiograph measures this electrical activity through the use of single electrodes placed at various points around the chest. These measure the emitted voltage at each point, and this signal is displayed on either an oscilloscope screen or a line graph. Analysis of the fluctuations in voltage is thus possible, and comparison of the readings from the different locations on the chest yields the general direction, or electrical axis, of the wave.

Along with its many uses as a diagnostic instrument and as a monitor during surgery, the electrocardiograph has a special utility in intensive treatment of coronary patients. The patient is continuously monitored by the electrocardiograph, which feeds its output into a small computer that has been programmed to recognize any dangerous patterns of electrical activity and to sound an alarm if these should appear. Heart function is also measured by a variety of flow rate and pressure measuring devices, some of which are put into catheter probes and inserted through the blood vessels into the heart itself.

Lung function. Since the 1940s a number of physiological tests have been developed for qualitative and quantitative evaluation of pulmonary functions. Perhaps the most widely used testing instrument is the spirometer, which measures lung capacity by measurement of the height to which the exhaled breath raises a bell turned upside down in water. The impedance pneumograph measures the volume of air inhaled by measuring the degree to which an electric current is impeded in its passage through the chest; the volume of air inhaled can be calculated from the difference in resistance.

The spirometer

Another method for pulmonary function study makes use of radioactive gases. The subject inhales an atmosphere in which all the gases are normal except the nitrogen, some of which is replaced by a radioisotope of the rare gas xenon. Radiation detectors on the chest reveal in which parts of the lung this atmosphere either is failing to penetrate or is present in abnormal volumes. A pattern of the lung areas can be obtained with radioactive sensitive film or on the fluorescent screen of a special scanning instrument.

Laboratory analysis. As knowledge in medical science increases, the need grows for instrumentation in laboratory analysis. In wide use is a device capable of providing 40 to 60 analyses an hour; it consists of a long fine tube through which are drawn samples of such substances as blood serum or urine, separated by quantities of neutral solution (distilled water in most cases). The samples and stretches of water between them are separated by air bubbles. A proportioning pump, by way of a manifold, introduces the air bubbles into the sample streams and advances precise quantities of all solutions and diluents into the system. For specific analyses, reactions are initiated in the samples with reagents that cause a noticeable colour change. These are monitored in a colorimeter, which measures the degree of absorption of this colour by the sample; this measurement is displayed as a peak on a moving graph paper. The degree of absorption is related to the composition of the sample, and the concentration of many constituents, such as albumen and uric acid, can be determined. This type of instrument is especially versatile if a small computer is used to control and record its functions. From each patient a blood sample is sent to a multichannel analyzer, which provides information to a small computer. The computer not only sets up calibration curves and checks periodically for errors but also provides a tabulation of each patient's results, indicating those surpassing safe levels. About 1,100 tests can be run each hour with such computerized instrumentation.

The colorimeter

Few instrumental methods are available to analyze compounds of high molecular weight in the human body, such as proteins in blood plasma and muscle. The white cells (leukocytes) in blood continuously synthesize about 2,000 proteins, which originate by a complex series of biochemical processes using genes (deoxyribonucleic acid; DNA) as a template. These proteins can now be separated and studied by a two-dimensional electrophoresis technique (see Figure 40). In this technique the proteins extracted from blood are first separated according to their isoelectric point by focusing them through a tube filled with a gel. The gel is then extruded and becomes the sample, which is separated in the second dimension by using a voltage applied across a rectangular gel slab. This causes the proteins to migrate down the gel slab and separate according to their molecular weight.

By courtesy of the Institute of Clinical Biochemistry, University of Oslo

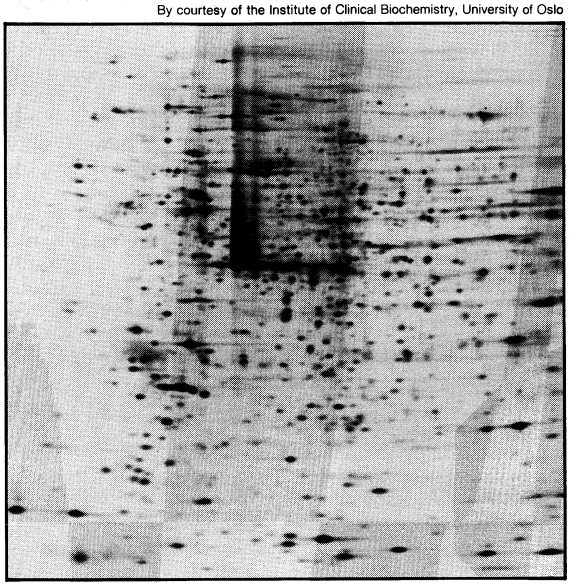


Figure 40: Two-dimensional electrophoresis separation of proteins from human leukocytes. Each spot corresponds to a separated protein.

Computer use. The use of computers in the biomedical field is continually growing, most obviously in processing the output of biomedical measuring devices. Computers can record and store the electrical analogue of a physiological event and play it back later, at a different speed if desired. They can also be programmed to recognize certain patterns in the data and to signal this recognition.

Computer-assisted measuring instruments

An example of a computer-measuring-instrument combination using the special abilities of the computer is a system to scan and analyze cell images. The computer is used in conjunction with a vidicon scanner and a high-resolution microscope. A stained blood smear is illuminated with monochromatic light and scanned with a vidicon. The information from some 4,000 points is stored in the computer, which is programmed to analyze it into the five categories of white blood cells.

Computerized axial tomography (CAT) combines the technologies of X-ray, scintillation counting, computer processing, and cathode-ray tube (CRT) display. The result is a picture of a transverse section of a body part that looks like an anatomic section. It is a reconstructed image of the area, formed by the computer's ability to calculate small differences in attenuation of a narrow beam of X rays sweeping through the tissues. These differences cannot be seen by ordinary X-ray techniques. The images formed can show a transverse section of the brain, for example, in which eight slices—each 1.3 centimetres thick—will clearly define the condition of the structure. This noninvasive technique can show lesions of the brain, such as tumours, hematomas, cysts, and metastatic diseases. Diagnosis of tumours is about 99 percent accurate. Abnormalities of the liver, pancreas, bladder, and related structures can also be evaluated with CAT.

Another noninvasive instrument, which produces images similar to CAT, is based on nuclear magnetic resonance (NMR). NMR has been used by analytical chemists since the 1950s as a tool to determine molecular structures because it produces a spectrum that uniquely defines protons within atomic nuclei in a molecule. In the human body, proton densities vary from tissue to tissue, and images can be formed that can reveal cancerous tissue and outline structures within the body. The NMR image is formed by superimposing a linear magnetic field gradient on the uniform magnetic field gradient applied to the object being studied. The NMR signal is a one-dimensional projection of a three-dimensional object. By taking a series of these projections at different gradient orientations, a two- or even three-dimensional image can be produced and displayed by a computer and CRT, just as in the CAT technique. Images based on proton densities in the body provide medical information quite different from that of the CAT, which is primarily based on electron densities.

A third noninvasive method of imaging interior portions of the body is based on ultrasonics. In this technique a beam of high-frequency sound energy between one and 15 megahertz is pulsed into the body and its echoes—which occur when the beam strikes a tissue or fluid interface—are converted to electrical impulses and displayed on a CRT. The outline of structures is thereby produced. Images can be produced to follow the course of pregnancies, to show organs, and to outline the presence of tumours. A distinct advantage of this instrumentation is that it is inexpensive, portable, and simple to operate.

DEVICES FOR ATMOSPHERIC MEASUREMENT

Meteorological instrumentation. The three most commonly measured atmospheric properties are temperature, humidity, and atmospheric pressure. Routine measurement of these properties involves use of such instruments as mercury or alcohol thermometers, wet and dry bulb hygrometers, and mercury barometers. Modern instrumentation must often be used, however, in such situations as remote weather sensing or automatic monitoring of indoor atmospheric variables. An example would be the instruments most often used to monitor these three variables in a radiosonde, or weather balloon. A thermistor measures temperature by the change it causes in its electrical resistance. A hair hygrometer measures humidity by the change it causes in the length of a fibre. Atmospheric pressure is measured by a type of aneroid barometer, which detects the distension that the pressure causes on a metal diaphragm.

Modern instrumentation is also used in meteorology to measure a variety of other atmospheric properties. Automatic systems monitor wind speed and direction, ozone concentration, and total rain falling through the air, in addition to a number of other special variables. Wind direction is most often monitored by the electronic sensing of the orientation of a wind vane; wind speed is monitored by measurement of the pressure it causes, by determination of the rate at which it cools a thermistor, or by sensing the speed of a rotating device used to measure wind speed (cup anemometer). Spectrophotometric techniques (*i.e.*, involving measurement of bands in the light spectrum) are used to monitor ozone concentration, and radar reflection is studied to determine the total rain in the air over large areas, leading to the detection of showers, cold and warm fronts, and cyclones.

Monitoring air purity. Safety engineering often requires monitoring the composition of the atmosphere to determine the concentration of oxygen, carbon dioxide, carbon monoxide, and other gases.

Oxygen deficiency, for instance, can be encountered in tunnels, ships' holds, storage tanks, and inert-gas-purged rooms. One method of quick detection of subnormal oxygen levels makes use of a fuel-cell detector. This cell consumes oxygen from the atmosphere surrounding the detector and generates a proportional electrical current.

Detecting oxygen deficiency

The accumulation of explosive or toxic flammable gases in the air can be measured with an instrument that continuously detects these substances, sounds an alarm, and automatically actuates ventilating devices. The explosion meter functions by drawing a sample of the air to be

monitored over a wire heated sufficiently to burn the gases on its surface. The heat of this combustion changes the resistance of the wire, this change being sensed to indicate electrically the amount of explosive gases.

A portable instrument analyzes for traces of many gases in air, such as missile fuel components, ammonia, chlorine, phosgene, hydrogen chloride, and nitrogen dioxide. The instrument incorporates a device called an alpha-particle emitter that ionizes the incoming sample and creates an ion current. A chemical reagent converts the gas to be detected into small aerosol particles, which in turn reduce the ion current in proportion to the number of particles present. This simple device can detect concentrations in the range of parts per million or lower.

Carbon monoxide is the most insidious of toxic gases. It gives no warning of its presence and occurs more frequently than all other dangerous gases. Carbon monoxide can produce serious physical effects at 100 parts per million concentration. The atmosphere in many industrial operations is monitored for this gas with an instrument that functions by the selective absorption principle. The monitored air is drawn through a detector cell containing a catalyst in which a series of thermocouples has been embedded. Absorption of the carbon monoxide liberates heat, which is indicated by the thermocouples.

Pollution control. People are becoming acutely aware of the degree to which they are contaminating the environment. Hundreds, if not thousands, of by-products considered to be pollutants are found in the air, water, and soil and in plants and animals. Awareness of the nature and extent of pollution is largely based on modern analysis instrumentation, which has made detection and study of pollutants possible. Pesticide and mercury pollution, industrial and automobile emissions, and many other contaminants require the increasing speed and sensitivity of modern instruments, whether in detection by laboratory analysis or by automatic pollutant monitoring.

POLLUTION DETECTION AND MONITORING DEVICES

Detection. The detection of environmental pollutants involves a multitude of techniques, partly because of the large number of pollutants. In the air alone there are more than 40 separate categories of substances considered to be actually or potentially harmful, with many of these categories containing more than one substance. Examples of techniques employed in the detection of some of the major pollutants, such as mercury and pesticides, illustrate the types of problems encountered.

Detection of mercury

Mercury in many forms is known to occur in trace amounts throughout the environment. Deposits of mercury in inland waters constitute an especially great hazard through the production of the more toxic methylmercury by microorganisms in the bottom mud. Methylmercury and other such organic mercury compounds attack the human central nervous system and remain in the body for long periods of time. Widespread incidents of methylmercury poisoning have been reported only since the mid-20th century. In Japan more than 168 illnesses and 52 deaths were reported in two separate incidents in the 1950s from consumption of mercury-contaminated fish. Since then several hundred persons in a number of countries have died or suffered serious illness from eating seed grain that had been treated with mercury compounds.

Highly sensitive and accurate instruments are needed to measure the concentration of mercury, since the presence of only a few parts per billion can be significant, and a concentration of only 200 parts per billion in blood causes noticeable symptoms. Instrumentation is needed not only in the simple detection of mercury but also in tracing its natural cycles through food webs and chains so that it can be determined, for instance, what concentration of mercury in lake water will produce a dangerous mercury concentration in the person who eats fish from that lake. A sensitive instrument used in this measurement is the ultraviolet spectrophotometer. This instrument passes ultraviolet radiation through a sample in which all the mercury has been converted to the elemental vapour form, and the amount of radiation transmitted is then measured. The frequency of radiation chosen, 2537 angstroms, is one for which mercury vapour has an intense absorption. The amount of 2537-angstrom radiation transmitted gives an indication of the concentration of mercury in the sample. Concentrations of mercury below the parts-per-billion range can be measured with this instrument.

Detection of pesticides

Pesticide contamination is another problem that is measured by modern instrumentation. Like mercury, pesticides are found virtually everywhere in the environment. Since they were developed specifically to kill living organisms, it is hardly surprising that many of them are harmful. Some of the basic problems in detection are similar to the problems in mercury detection: the pollutants must be identified in very small concentrations, and their presence in various materials must be sensed so as to gain information about their distribution and cycle patterns. An additional factor with pesticides, however, is their variety; approximately 300 organic chemicals are marketed in more than 10,000 different formulations. Moreover, these pesticides in the environment often change into other substances through chemical reaction. Some of these secondary pollutants, like the polychlorinated biphenyls resulting from DDT (dichlorodiphenyltrichloroethane) reactions, can be dangerous. Thus, detection is often a matter of analyzing not only for a specific pesticide but also for a variety of them plus their by-products.

Given these instrumental requirements for high accuracy and multicomponent analytical ability, the gas/liquid chromatograph has come to be widely used. This instrument operates by passing the sample through a separation column (see above *Chromatographic analyzer*). As the sample passes through the column, individual molecules are absorbed and then released at different times from the column surface. If a column of the proper type is employed, the components of the sample emerge from the column completely separated from each other, with the most strongly absorbed emerging last. This emerging stream then passes through a detector, which detects each component by means of its thermal conductivity or reaction to ionization. The relative response is detected and usually presented on a recorder. This instrument can be used in the simultaneous measurement of many components in a sample and can detect concentrations in the parts-per-trillion range.

Monitoring. The study and control of pollution often necessitates the use of detection instrumentation for continuous monitoring of pollutant levels in rapidly changing systems. Air-pollutant levels, for instance, may undergo significant change in a matter of hours, and harmful changes at one point of a water pipeline can occur in less than one second. Since the older methods of laboratory analysis take many hours or days, modern instrumentation is desirable not only for its speed but also for its ability to be integrated into a system of automatic control. While continuous monitoring can be useful in virtually any pollution situation, its applications today are primarily in the areas of air and water pollution.

Air pollution

Atmospheric monitoring systematically collects and evaluates aerometric (involving the density and weight of gases, such as air) and related data to provide information on pollutant concentrations in ambient air, emissions from polluting sources, and meteorological conditions. Information derived from monitoring is applied in various pollution-control activities: assessing pollution effects on human beings and their environment; studying pollutant interactions and patterns; establishing ambient air-quality standards; developing control strategies and regulations; evaluating the effectiveness of these control strategies when adopted; activating emergency procedures to prevent air-pollution episodes or to reduce their severity; and guiding efforts to minimize the effect of air pollution by applying mathematical modeling and planning systems.

One of the special problems encountered in air-pollution monitoring is the necessity of virtually simultaneous monitoring over a large area. Indeed, many environmental scientists feel that a global monitoring system is needed. To accomplish this with fixed sampling stations would require vast amounts of money, labour, and equipment. The instrumental alternative involves remote sensing. One method is monitoring by aircraft equipped with a cor-

relation spectrometer, which compares the spectrum of incoming light with a reference spectrum of the pollutant being monitored. Only the light that matches the reference spectrum is detected, the amount of this light being a measure of pollutant concentration. Such systems are used to monitor nitrogen and sulfur oxides. More ambitious systems involving satellites carrying the monitoring instrument are also being used.

Water pollution

The objectives of water-pollution monitoring are much the same as those of air-pollution monitoring. Many of the same instruments used to monitor air pollutants are employed for compositional analysis. In addition, other qualities of water besides composition must be monitored. Thus, water-pollution control involves the use of instruments that monitor such qualities as turbidity, colour, pH, and presence of microorganisms. A special requirement in this area is often extreme speed of operation. This need occurs especially in water-treatment plants and in the control of industrial waste water, problems in which large volumes of rapidly moving water in transit lines must be monitored. In these situations the monitoring must be done on-line (*i.e.*, as the water moves through the pipe) because of the large time lag involved in sample removal and remote analysis. An example of on-line analysis is the ultraviolet photometric method of monitoring trace amounts of oil in water. A continuous sampling system divides the water sample into two parts, one of which is filtered to remove undissolved oil. The filtered and unfiltered streams are then run through separate cells of an ultraviolet analyzer in which ultraviolet radiation is beamed through each stream. Since the difference in their absorption characteristics is solely the result of the presence of oil, oil content can be continuously monitored.

LABORATORY INSTRUMENTATION

From earliest times instruments have been employed in the scientific laboratory. Early instruments were usually quite simple devices, such as thermometers, rulers, optical devices, and meters. As instrumentation became more complex and capable of greater scope, laboratories were developed for specific purposes; products produced elsewhere, either by research groups or in manufacturing, were measured and analyzed with great accuracy. Such analytical instrumentation laboratories, now firmly established as necessary to world technology, take many forms and can best be categorized by function.

Quality control of processes. Manufacturing processes generally involve a number of laboratories associated with the various steps in production. These laboratories are located at the plant facilities in close proximity to the origin of their samples. In large industrial complexes there may also be a centrally located test laboratory.

The tests and analyses, which are made with very simple instruments, are intended to control process variables and the quality of the product. Typical examples of such tests include measurement of liquid density using a hydrometer, determination of the particle size of a granular product using sieves, and measurement of viscosity by timed flow of a liquid through a narrow-bore tubing.

Since many of the quality control tests are done on products that are common among many industries, the problem of standardization of these tests arose long ago. In many countries standards have been established and are regularly brought up to date. In the United States, for example, the American Society for Testing and Materials (ASTM) publishes a multivolume set of specifications of the instrumentation and methods for 4,300 standard tests applied to almost every material used in industry. These tests cover not only the chemical but also the physical properties of materials, such as paper, ferrous castings, petroleum products, fuels, paints, plastics, electrical components, radioisotopes, and pure chemicals. An example of an ASTM physical test is the ultrasonic inspection of metal tubing for longitudinal discontinuities. Quality control laboratories use these documented testing procedures for much of their work.

Analytical laboratories in industry. Large industrial companies manufacturing chemical, petroleum, or consumable products must have a capable analytical laboratory to supply accurate information on all phases of company activities, from early research on new products to final process control. While some of the older chemical-type analyses are still used, the greater part of the analytical work today is accomplished by instrumental methods. These instruments include spectrometers of various types—infrared, atomic absorption, X-ray, mass, electron, spark emission, Mössbauer, and nuclear magnetic resonance. Other instruments based on the principles of gas chromatography, micromeritics, physical chemistry, magnetic susceptibility, and electron microscopy are also widely used.

Identifying an unknown compound

For qualitative identification of compounds, a combination of results from several complementary instruments is advantageous in rapidly obtaining an answer. Laboratories usually obtain spectra for a given unknown compound on three different instruments: a mass spectrometer, an infrared spectrometer, and a nuclear magnetic resonance spectrometer. Since each instrument functions best for pure compounds, the sample is subjected to a separation and purification step (usually by chromatographic means) before a spectrum is obtained. Each spectrum reveals important clues to the identity of the unknown compound. These instruments are operated by trained personnel, and the combined spectra are interpreted by a specialist, though some functions are now being performed by computer systems.

Chromatography

As a method of separating the individual components of a complex mixture, gas chromatography is unequalled. Although the principle of gas–liquid partition chromatography was clearly stated in 1941, the field really opened in 1952 when the simple instrumentation needed to separate and analyze a mixture of fatty acids was described. Since then intensive work with this technique has been carried out in laboratories all over the world, and many instruments, with varying degrees of complexity, have been developed. The present importance of gas chromatography reflects not only the comparative ease of using the technique but also its speed in analyzing very complex mixtures containing a wide range of compounds and concentrations. When the full capability of the method is applied to complex mixtures with more than 100 components, digital electronics can be used to help obtain the analysis. A digital computer built into these devices can accept the chromatogram from a mixture and type out a complete analysis of each component. A complex mixture can be analyzed in minutes, an accomplishment that prior to the discovery of this technique bordered on the impossible, regardless of time. Analyses can be made with gas chromatographs for almost any compound possessing some volatility at 500° C. Even a large number of metals have been analyzed by first converting them to volatile compound form and obtaining a chromatogram of the compounds. For analysis of nonvolatile compounds a form of chromatography in which the carrier gas is replaced by a carrier liquid is used. Liquid chromatography is extremely useful in the medical field; a complete analysis of an APC (aspirin, phenacetin, and caffeine) tablet can be made in less than five minutes. Chromatography can measure trace concentrations of compounds present in the parts-per-million range; for this reason it is useful for measuring pollutants in the environment.

Mass spectrometry

The different masses (isotopes) of an element were first separated in 1910 and the first mass spectrograph was built in 1919. Mass spectroscopy has a long history, but only since 1950 has its powerful ability to determine molecular structure and to identify molecules been developed. This development is due to the increasing capabilities of the instrumentation available. Mass spectrometers are an important analytical tool in most major laboratories.

A mass spectrum is produced when a gas, liquid, or solid is vaporized and injected into an ionizing source (usually an electron beam) under high vacuum. This breaks the molecules into characteristic ion fragments, which are then separated and recorded as the mass spectrum. The first mass spectrometers used a magnetic field to separate the positively charged fragment ions, and present instrumentation is often based on magnetic separation.

A mass spectrum contains a great amount of information

about a molecule. From the molecular weight it indicates and from the masses of fragment ions it reveals, the structure of the original molecule can usually be deduced. The instrument can be universally applied to compound identification. Since gas chromatography has separating power and mass spectrometry has identification power, the two instruments are commonly combined into the gas chromatograph/mass spectrometer (GC/MS) system (see Figure 41). This provides a uniquely capable instrument for the analysis of complex mixtures.

GC/MS computer system

By courtesy of Hewlett-Packard Instruments

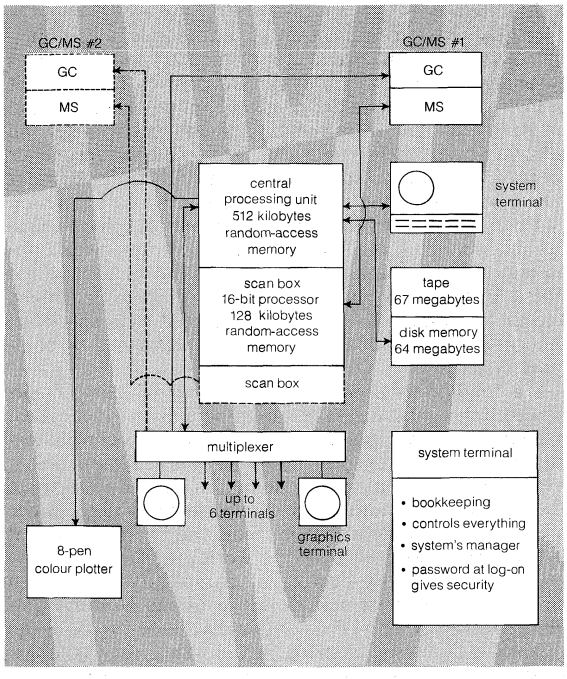


Figure 41: Computerized gas chromatograph/mass spectrometer (GC/MS) system.

A GC/MS instrument is such a powerful analytical tool that it is an indispensible instrument in both research and analytical laboratories. So much useful information is produced and so many data analysis techniques are employed that a computer is a necessary part of the GC/MS system. During the 1970s great advances were made in the development of the system. After injection of a sample of a complex mixture containing 200 compounds into the gas chromatograph, for example, an operator sits in front of a CRT display module to analyze the data produced by the computer. As the more than 2,000 mass spectra associated with the compounds are produced, the computer compares each mass spectrum with more than 80,000 mass spectra of reference compounds and identifies it within seconds. Totally new compounds are also identified by the reconstruction of their molecular structures from their mass spectral patterns. Several operators can work on data or samples simultaneously on the computerized GC/MS, which has revolutionized analysis in fields such as biomedicine, forensics, and environmental pollution.

Infrared spectroscopy

As early as 1905 it was recognized that the selective absorption of infrared radiation arises from the mutual vibrations of the atoms in a molecule (*i.e.,* resulting from the molecules jumping from one energy level to another). The first practical infrared instrumentation was built in 1940 and demonstrated its utility with a number of startling applications. Similar instruments were manufactured, and a revolutionary period of growth for chemical analysis by infrared spectroscopy began. An infrared spectroscopy section is an important part of present-day analytical laboratories.

The atoms forming a molecule can vibrate only in certain fundamental modes characteristic of that molecular configuration. When the frequency of incident radiant energy occurs at the same vibrational frequency as the vibrating atoms, the molecule absorbs that radiation. An absorption spectrum in the infrared region of 2 to 15 micrometres in wavelength provides a characteristic fingerprint pattern of a compound. This spectrum can be used for qualitative identification and quantitative analysis of a substance. The qualitative aspects have predominated; the characteristic absorption bands of more than 130,000 compounds have been determined and cataloged as reference spectra. These are stored on magnetic tape. By use of a computer system, the spectrum of an unknown compound can be compared with 40,000 reference spectra on the magnetic tape in 40 seconds.

Nuclear magnetic resonance

The spectroscopy of nuclear magnetic resonance began in 1946. The energy of radio frequency waves is too small to vibrate, rotate, or interact with atoms or molecules. It is sufficient, however, to affect the magnetic moment (tendency to produce motion) and spin of the nucleus of atoms, so that a resonance absorption of radio frequency radiation by the nucleus occurs when atoms are placed in a magnetic field and irradiated with the proper frequency radio waves. A spectrum is produced by observing resonances for a compound as the magnetic field or radio frequency is scanned.

In analytical chemistry, NMR spectroscopy detects the shape and structure of molecules. Molecules containing hydrogen atoms have the strongest effect, and most instruments are designed to produce NMR spectra of hydrogen atoms. Both qualitative and quantitative information is provided, since each different type of bond of the hydrogen atom gives its own unique resonance. NMR became of vital importance in scientific laboratories as the first commercial spectrometer, appearing in 1953, was followed by a number of other instruments. Some of these are quite complex for research, while others are quite simple for routine analytical work.

In the next larger size range to the molecule are the clusters of molecules called colloids and then the group of particles with dimensions reaching from microscopic resolution to visibility by the eye, particles of 0.1 to 1,000 micrometres in size. These particles are related to natural phenomena of common experience. In the atmosphere they influence the colour of the sky and sunset. Because of their basic role in industrial processes and their importance in air pollution, particles influence much of modern life. They are used, for example, as catalysts in processes that are completely altered by a slight change in their properties. The tastes of coffee, chocolate, and peanut butter, for example, are influenced strongly by the fineness of the ground particles in them.

Micromeritics

Micromeritics is the technology of fine particles. Instruments for measuring the physical characteristics of particles are of prime importance; the physical properties that determine the characteristics of finely divided matter are the primary ones of particle size, surface area, and the sizes of the pores that extend downward from the surfaces.

Enormous surface area can be involved in a small mass of powdered material. A one-centimetre cube of finely divided powder may have a surface area as large as 10 million square centimetres. This surface area is measured by an automatic analyzer that determines the quantity of gas necessary to cover the area with a single layer of molecules of nitrogen gas. The pore structure of a surface area can be measured by simply determining the amount of mercury that can be forced into these pores at various increasing pressures. The particle size of a powder can be determined in an automatic instrument by observing how fast the particles settle through a viscous fluid.

(F.W.K./Ed.)

SURVEYING INSTRUMENTS

The fundamental measurements made by the surveyor include distances, angles (both horizontal and vertical), altitudes, and directions. To carry out these measurements, several different types of instruments and supplementary equipment are used.

Distance-measuring equipment. Distances are measured by mechanical devices (chain or steel tape) or by electronic means.

The chain. Gunter's chain, the 300-year-old measuring instrument by which all survey measurement in the English-speaking countries and much of it elsewhere was long

done, has been superseded by the steel tape and electronic equipment. Gunter's chain is 66 feet long; 80 chains equal one mile, and 10 square chains equal an acre. The chain is subdivided into 100 links. A rod or perch is 25 links. Each link is a short section of wire connected to the next link by a loop. At each end of the chain is a brass handle. The 66-foot unit is still called a chain and is still in use in property descriptions and in the public land systems.

The steel tape. A graduated steel ribbon, or a flat wire, with handles at each end, is the basic modern means of measuring distance. Its length is taken as the straight-line distance between the two end marks at 68° F, because the tape's actual length varies with different types of support, different tensions, and different temperatures. In the field, if extreme accuracy is desired, the type of support, the tension, the temperature, and the difference in height of the two end marks are recorded and the measured length corrected accordingly.

In most routine surveys the tape is held by hand at the ends, kept high enough to clear ground objects, made horizontal by estimation, and placed in the direction of measurement by eye or with an instrument. The desired tension is estimated, and the positions of the graduations are brought to the ground marks by plumb bobs. Sometimes the air temperatures are recorded. For higher accuracy, tripods or other supports are employed, or the tape is fully supported on smooth surfaces. The slopes and temperatures of the tape are determined. Tension is regulated with a spring balance.

Eliminating temperature problems

When especially accurate measurements are necessary, particularly for baselines in a triangulation system, a steel tape must be used at night or when the sky is overcast; otherwise radiant heat would make it impossible to determine tape temperature. Tapes made of Invar (trademark) avoid this difficulty. Invar is a steel alloy with an extremely low coefficient of thermal expansion (about 0.00000045 foot per foot per degree Fahrenheit). With such a tape, a temperature difference of 10° F from the design temperature would result in an error of only about 1/20 of an inch in 1,000 feet. Unfortunately, Invar tapes are too easily damaged for general use. Before the introduction of Invar, iron bars immersed in melting ice were used for daytime measurements.

Electronic instruments. Sir Robert Watson-Watt, a Scottish physicist working in England, introduced an electronic pulse-type instrument in 1935. Called hiran, it measures distances over water between land survey stations obscured from each other by the Earth's curvature. An airplane flying high enough to be visible from both stations crosses over the line between them while sending out a series of pulses. These are automatically relayed back to the plane from each station. The minimum sum of distances represented by the timed returns is chosen, and the nearby sums are corrected to be averaged with it. The result is corrected for the altitude of the plane and refraction of the atmosphere, and the final value is taken as the distance between the two stations.

The phase-shift type of system was first developed in 1948 and gives very accurate results. The process is comparable to conventional radio transmission, in which the carrier frequency is modulated by the audio frequencies that the receiver makes audible.

In phase-shift devices, the carrier frequency is either light (generated by a laser or an electric light beam) or an ultra-high-frequency radio beam. The light beam requires a clear line of sight; the radio beam can penetrate fog, haze, heavy rain, dust, sandstorms, and some foliage. Both types have a transmitter-receiver at one survey station. At the remote station, the light type uses a set of corner mirrors; the high-frequency type utilizes a retransmitter (requiring an operator) identical to the transmitter-receiver at the original station. A corner mirror has the shape of the inside of a corner of a cube. It returns light toward the source from whatever angle it is received, within reasonable limits. A retransmitter must be aimed at the transmitter-receiver.

In both types of instrument, the distance is determined basically by the length of time it takes the radio or light beam to travel to the target and back (see RADAR). The elapsed time is determined by the shift in phase of the modulating signal during its travels. Electronic circuitry detects this phase shift and converts it very accurately into the exact fraction of a small unit distance. By using several frequencies for the modulator signal the total distance can be computed.

Shift in phase of modulating signal

Angle-measuring equipment. The theodolite is the basic instrument for measuring horizontal and vertical angles. It consists essentially of a sighting tube in the form of a telescope, and horizontal and vertical scales set up so that horizontal and vertical angles of the object sighted in the telescope can be read from the scales.

Theodolites and transits. There are two general types of theodolites: direction theodolites (also called simply theodolites) and repeating theodolites (also called American transits). In the direction theodolite, the graduated horizontal circle remains fixed while the telescope is pointed at several objects in turn and the direction to each is read on the horizontal scale. In the repeating theodolite, the same angle is measured a number of times on the horizontal scale, which is rotated by as much as the arc of each measurement. The successive values of each measurement are accumulated on the horizontal circle, the final reading representing the total arc of all the measurements taken. The desired value is then obtained by dividing the final reading by the number of repetitions. The repeating theodolite, or transit, is used in the United States, Canada, and, to some extent, in Central and South America.

The theodolite is basically a more accurate instrument than the transit, measuring angles to 5 seconds of arc or less; the transit only measures to about 20 seconds of arc. The transit's repeating characteristic, however, permits the angle to be measured several times and an average reading obtained in about the same time that it takes to measure once with a theodolite, and the end result is equal accuracy. When angles are established by one setting from one fixed station, as in placing marks for construction, the theodolite is more accurate.

Modern transits and theodolites nearly all have stadia hairs (see Figure 42). When the vertical angle is observed, along with the distance on the rod (staff) intercepted by two of the horizontal stadia hairs, the surveyor can determine approximately the horizontal distance and the difference in height between a point in the telescope and the point on the rod or staff (see below) where the cross hair falls.

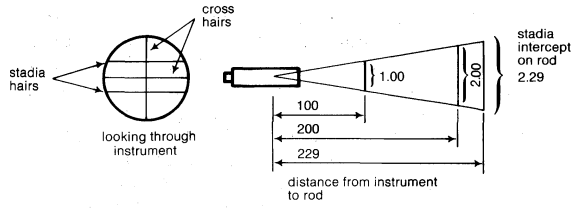


Figure 42: Principles involved in the use of telescope stadia hairs and a staff to determine distance. In the example illustrated, the distance to the rod is equal to 100 times the stadia intercept on the rod (see text).

Vertical circle attachments. Except in astronomical observations, vertical angles seldom need be measured with the accuracy required for horizontal angles. In all instruments, the vertical circle is attached to the end of the horizontal axle and turns with the telescope. In American transits, it is read with a vernier attached to a standard and thus depends on the accuracy of the leveling of the instrument. In modern European instruments, the vertical circle is read with an optical device governed by a pendulum and therefore is independent of any inaccuracies in the hand leveling of the instrument. The accuracy of the setting of the reading device in both instruments can be tested by reading the same vertical angle with the telescope both direct and reversed.

Testing accuracy of the setting

Plane-table alidades. Modern plane-table alidades consist of a telescopic sight of about 13–16 power equipped with stadia hairs and mounted on a metal graduated straightedge called a blade, with which survey data can be plotted on the supporting drawing board. The telescope, on horizontal bearings, does not reverse, and it measures

angles of elevation and depression up to as much as 20° to 30°, as indicated on a vertical circle. A prismatic eyepiece makes it possible to observe when looking down into the telescope to avoid the interference of the plane table. Also mounted on the blade are a circular spirit level for leveling the plane table and a trough compass that operates when the alidade is aimed nearly along the magnetic meridian.

Plane-table surveying has been nearly eclipsed by photogrammetry. It is still used, however, for mapping small areas, for filling in where the lack of detail prevents the proper orientation of aerial photography, or where detail is concealed on the photograph.

Measurement of direction. There are two major techniques for measuring direction—by means of compasses and by celestial observations.

Compasses. Two needle-type compasses are used: box compasses and circular compasses. Box compasses must be used with angle-measuring instruments, as the compasses operate only when the line of sight points toward magnetic north. The bearing desired must then be measured from this direction.

Circular compasses can be mounted on theodolites and are often permanently mounted on transits. When provided with open sights, they can be used alone, either mounted on light tripods or held in the hand.

Magnetic declination, called variation by mariners, is the angle from the direction of true north to the direction of the needle. This angle varies from place to place, and at any place it changes slowly through large, little-understood swings. This is called the secular change. The diurnal change is a more or less regular swing as follows: quiescent at its mean position from 9 PM to 3 AM, then easterly to 2 to 5 minutes at 8 or 9 AM, and then westerly to 2 to 5 minutes at 1 or 2 PM, thence to mean at 9 PM. There is also an annual swing in the order of one minute, varying according to locality.

Secular and diurnal changes

Celestial observations. Astronomic north, the direction of the celestial North Pole, is determined by observations of the Sun or stars with transits or theodolites. The most accurate method of determining direction is to observe the North Star (Polaris) in latitudes where it is visible and not too high to give accurate results. In southern latitudes there are six bright stars about 30° from the Pole, well enough distributed so that one is usually above the horizon. These are observed in the same way as Polaris.

As the first step, the horizontal direction of the star is measured from a mark on the ground. The precise time is recorded at the moment the instrument is pointed at the star. With these values and the latitude and longitude of the instrument, the bearing of Polaris or other star with relation to an Earth station can be computed. If any of these data are unavailable, further observations are required.

A second method is sometimes used. The altitude, direction, and time of observation are noted for a star in the east and one in the west at about the same elevation. By referring to astronomical tables the resulting bearings of the mark are obtained. These are averaged, and thus observational errors in the vertical angles are reduced.

Vertical measuring equipment. Theodolites and transits measure vertical (as well as horizontal) angles. Surveying levels and barometric-pressure devices measure vertical levels, or elevations, exclusively.

Surveying levels. There are three types of surveying levels: (1) the wye, or the dumpy, level, (2) the tilting level, and (3) the self-leveling level. The first type consists of a telescopic sight, like that of a transit but usually of slightly higher magnification, to which a long spirit level is attached and adjusted so that the bubble centres when the line of sight is horizontal. The whole is mounted on a vertical spindle in a leveling head so that it can be turned in any direction. The telescope of the wye level can be rotated and changed end for end in the wyes (Y-shaped supports) for adjustment.

The tilting level is similar, but the telescope with the main bubble attached can be separately tilted up or down by means of a micrometer screw, giving it greater accuracy. The spindle is first set nearly vertical with a circular spirit level, and the main bubble is centred with the micrometer screw.

The self-leveling level is similar to the tilting level except that it has no micrometer screw. Instead, it contains an optical part that is operated by a pendulum in such a way that the line of sight is automatically made horizontal. Some types are extremely accurate, and the pendulum eliminates the operation of centring the main bubble, which is especially time-consuming.

For very accurate work, tilting levels are equipped with coincidence-reading bubbles. By an optical device the elongated bubble appears to be split into two sections, with a central portion apparently removed. As the telescope is leveled, the sections of the bubble are brought closer together until their ends coincide; then the bubble is precisely centred. Often tilting and self-leveling levels have three horizontal cross hairs so that three readings of the staff (rod) can be taken and averaged. Optical micrometers also yield greater accuracy in rod readings; they consist of thick glass disks with parallel, optically flat surfaces usually mounted in front of the telescope. When the micrometers are rotated on a horizontal axis, they move the line of sight up or down and parallel to itself. A graduated drum indicates the distance moved from their normal position, so that when the line of sight is placed exactly on a rod graduation, they read the very small additional quantity to be applied.

Coincidence-reading bubbles

The level rod (staff). The rod is used to measure the vertical distance between the point on which it stands and the line of sight. It is made of wood with graduations and large, clear numbers usually printed on it. It is divided either into centimetres or half-centimetres or into hundredths of a foot or yard. Between divisions it is alternately black and white, with the exact values occurring where the colour changes. For routine leveling, the American rod is made in two sections. The rear section slides upward until, when fully extended, the face reads continuously up to 3.7 or 4 metres. It has a target that may be set where desired.

For accurate leveling the graduations are painted on a steel or Invar ribbon fastened at the bottom and held taut by a spring at the top. A special thermometer is held with its bulb against the ribbon, and a circular spirit level is attached to aid in keeping the rod vertical.

Aneroid barometers. Vertical measurements can be approximately determined by finding the difference in barometric pressure at the two elevations. Aneroid barometers and hypsometers measure such differences. Aneroid barometers are devices in which changes in atmospheric pressure cause a needle to move over a scale. Instruments of this type designed for surveying are called altimeters. A type that records time along with pressures is usually placed at a known elevation (the base), and the second instrument is taken to the points where measurements of elevation are desired. When the second instrument is read, the time is recorded; thus the simultaneous reading of the base instrument can be selected. The difference of the two readings must be corrected for the unit weight of the air, which is estimated from the barometric pressure, temperature, and humidity.

More accurate results, independent of the unit weight of the air, can be obtained by the two-base method. Recording aneroids are placed at two bases, preferably one higher and one lower than the elevations to be determined. In effect they weigh the air. Each field reading is adjusted in proportion to the relative height above and below the two bases, and so the sum of the two heights equals the known difference in height between the bases. Within a radius of 16 kilometres, this method gives elevations within about 0.6 metre.

Two-base method

In Australia, some effort has been made to use three bases over a large area and to compensate according to the barometric pressures shown on a weather map.

Hypsometers. The hypsometer is a device that measures barometric pressure by determining the temperature required to boil distilled water. By means of suitable tables, this temperature can be converted to barometric pressure. The bulb of an accurate thermometer is immersed in a very small container of water that is brought to a boil. The results can be handled as described for the aneroid barometer, but usually the boiling point is again determined at the starting point, and, if there is a difference in the

measured altitude there, the field elevations are correlated in proportion to elapsed time. (P.K./Ed.)

INSTRUMENTS FOR SPACE EXPLORATION

During the 1970s many successful uses of instrumentation in space were developed. These were centred on miniaturization of existing instruments and their adaptation for special uses and remote control, rather than the development of entirely new instruments. The accomplishments of the 1970s were embodied in the Viking missions to Mars, the Pioneer Venus missions, and the Voyager missions involving flybys of the giant outer planets and their satellites. In all these space experiments, computers played a key role in control of the instruments and transmission of their data back to Earth. They also were used to help process much of the data and to enhance photographs taken by the planetary probes.

The Viking missions. Two identical Viking spacecraft, launched 30 days apart in 1975, successfully landed on Mars after traveling through space for almost a year (see Figure 43). Each spacecraft carried an impressive array of instrumentation primarily concerned with the detection of life, although other conditions were studied as well. Each had an orbiter section that delivered a lander module to a selected point for descent to the surface and then served as a relay station to control the lander's functions and transmit data back to Earth. Radio transmission over the 322-million-kilometre distance required 40 minutes round-trip. The orbiter carried television cameras, an infrared thermal mapper, and an infrared spectrometer for water-vapour mapping. By the end of the orbiter mission in August 1980, almost 55,000 pictures of the planet had been taken.

The landers operated for several years and, although no evidence of biological life was detected, a huge amount of information on soil chemistry and meteorological conditions was obtained. Three types of instrumentation systems were mounted on the lander: an X-ray fluorescence spectrometer to analyze surface materials, a module to perform biological experiments, and a molecular analysis module to identify atmospheric constituents and organic molecules in the Martian soil. The lander had a long mechanical arm to scoop up soil from the surface and deliver a sample to the instruments. Surface dust was introduced to the X-ray fluorescence spectrometer. The spectra produced were matched by computer to known spectra for elements of atomic number greater than 11.

Analysis of Martian soil

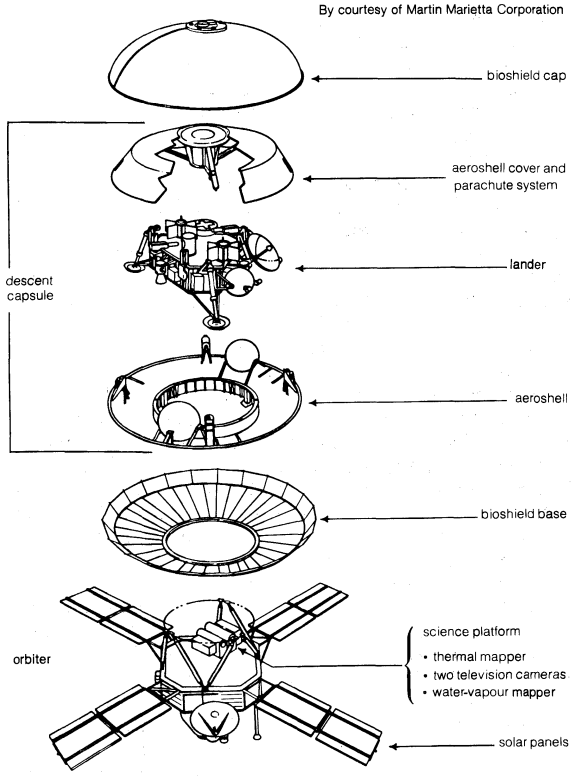


Figure 43: Sections of the Viking spacecraft sent to Mars.

The biological module performed three experiments on portions of a common soil sample collected from the Martian surface. The pyrolytic release experiment measured the assimilation of a mixture of radioactive carbon dioxide and carbon monoxide from the atmosphere by living matter after incubation under the Martian atmosphere. The degree of assimilation could be determined by analyzing the gases and pyrolysis products of organic matter that might be present with a gas chromatographic procedure. The labeled release experiment detected assimilation of organic material and its conversion to carbon dioxide by living matter. Martian soil was moistened with a nutrient mixture of simple, radioactively labeled organic compounds, and the headspace gas over the mixture was monitored for an increase of radioactivity with time. The gas exchange experiment was similar; a gas chromatographic analysis was conducted on the headspace gas.

Detection of carbon compounds

The module containing the gas chromatograph/mass spectrometer (GC/MS) unit was designed not only to analyze for atmospheric gases but primarily to detect organic carbon compounds—a necessary condition for the presence of life. The organic compounds were investigated by pyrolyzing a sample of soil to decompose it into compounds easily identifiable by the GC/MS technique. A total of 14 organic analyses were performed on four soil samples and 57 atmospheric samples. The GC/MS analyses showed the soil contained no organic molecules, despite detection sensitivities in the parts-per-billion range. Having successfully completed their mission, the GC/MS instruments were turned off during March–April 1977.

The Pioneer missions. In 1978 two Pioneer Venus spacecraft were launched, requiring six months to reach the planet. The spacecraft, one consisting of an orbiter and the other essentially comprising four probes that landed on the surface, carried numerous instruments. The orbiter used a Bennett-type mass spectrometer to directly measure the ions present in the Venusian ionosphere. Electrons and their temperature were measured electronically. A quadrupole mass spectrometer operating for 243 Earth days measured the neutral gas composition of the upper atmosphere. An ultraviolet spectrometer produced valuable information about the sulfur oxide content of the atmosphere and its photochemical changes. A radiometer mapped thermal emission data, providing large-scale meteorological information.

Of the four probes, the main probe entered the Venusian atmosphere in December 1979 and descended slowly to the surface (part of the way by parachute), producing data for nearly one hour below an altitude of 62 kilometres. A magnetic mass spectrometer measured the composition of gases relative to carbon monoxide, the dominant gas, producing 51 mass spectra before impact. A gas chromatograph provided an alternate method to measure atmospheric gas composition. A solar flux radiometer measured solar radiation as a function of altitude to increase knowledge of the extraordinary surface temperature of Venus.

The Voyager missions. Launched in 1977, each of the two Voyager spacecraft carried sophisticated infrared and ultraviolet spectrometers to explore the outer reaches of the solar system. The infrared spectrometer used was a Fourier transform infrared (FTIR) interferometer type, which extracts the infrared spectrum from an interferogram rather than producing it from an optical scan. It operates from thermal emission of the sample. In this application, the solar bodies being studied are the samples whose emissions are gathered by a telescope.

RESEARCH IN INSTRUMENTATION

In the evolution of instruments for use in research, spectroscopy is increasing in importance; the instruments tend to be of greater complexity and capability, with more automatic operations and more assistance from computers. Among the newer methods of spectroscopy, the emphasis is on the electron and the ion. Photochemical dissociation spectroscopy, a type of chemical spectroscopy, is the product of pulsed high-power lasers, molecular beams, and

time-of-flight mass spectrometry. The ion microanalyzer, which combines the functions of a mass spectrometer and an ion emission microscope, provides a visual display of the complete composition of a minute section of a solid surface.

Ion-mobility spectrometry

Ion-mobility spectrometry (IMS) originated in 1970 as plasma chromatography. The technique is based on ions reacting with trace molecules at atmospheric pressure. The capabilities of ion-mobility spectrometry are many: it has uncomplicated instrumentation giving femtogram (10^{-15} gram) sensitivities; it is capable of operating as a five-mode gas chromatographic detector; it can be used to study ion-molecule reactions; and it has an ability to selectively monitor ultra-trace components in air. Major advances have been made in development and deployment of IMS instrumentation in defensive military applications as a detector of chemical warfare agents, principally the deadly nerve gases. (F.W.K./Ed.)

SYSTEMS OF WEIGHTS AND MEASURES

The concept of weights and measures, though it has come in modern times to include temperature, luminosity, pressure, and electric current, once consisted of only four basic measurements: mass (weight), volume (liquid or grain measure), distance, and area. The last three are, of course, closely related.

Basic to the whole idea of weights and measures are the concepts of uniformity, units, and standards. Uniformity, the essence of any system of weights and measures, requires accurate, reliable standards of mass and length and agreed-on units. A unit is the name of a quantity, such as kilogram or pound; a standard is the physical embodiment of a unit, such as the platinum-iridium cylinder kept by the International Bureau of Weights and Measures at Paris as the standard kilogram.

Two types of systems are distinguished historically: an evolutionary system, such as the British Imperial, which grew more or less haphazardly out of custom, and a planned system, such as the present-day Système Internationale (SI), in universal use by the world's scientific community and by most nations.

Early units and standards

ANCIENT MEDITERRANEAN SYSTEMS

Body measurements probably provided the most convenient bases for early linear measurements; early weight units may have derived casually from the use of certain containers, or calculations of what a person or animal could lift or haul.

The historical progression of units has followed a generally westward direction, the units of the ancient empires of the Near and Middle East finding their way, mostly as a result of trade, to the Greek and then the Roman empires, thence to Gaul and Britain via Roman conquest.

The Egyptians. Although there is evidence that many early civilizations devised standards of measurement and some tools for measuring, the Egyptian cubit is generally recognized as having been the most ubiquitous standard of linear measurement in the very ancient world. Devised about 3000 BC, it was based on the length of the arm from the elbow to the extended fingertips and was standardized by a royal master cubit of black granite, against which all the cubit sticks in use in Egypt were measured at regular intervals.

The cubit

The royal cubit (524 millimetres, or 20.62 inches) was subdivided in an extraordinarily complicated way. The basic subunit was the digit, doubtlessly a finger's breadth, of which there were 28 in the royal cubit. Four digits equaled a palm, five a hand. Twelve digits, or three palms, equaled a small span. Fourteen digits, or one-half a cubit, equaled a large span. Sixteen digits, or four palms, made one *t'ser.* Twenty-four digits, or six palms, were a small cubit.

The digit was in turn subdivided. The 14th digit on a cubit stick was marked off into 16 equal parts. The next digit was divided into 15 parts, and so on, to the 28th digit, which was divided into two equal parts. Thus, measurement could be made to digit fractions with any denominator from 2 through 16. The smallest division, $1/16$ of a digit, was equal to $1/448$ part of a royal cubit.

The accuracy of the cubit stick is attested by the dimensions of the Great Pyramid of Giza; although thousands were employed in building it, its sides vary no more than 0.05 percent from the mean length of 230.364 metres (9,069.45 inches).

The Egyptians developed methods and instruments for measuring land at a very early date. The annual flood of the Nile created a need for benchmarks and surveying techniques so that property boundaries could be readily reestablished when the water receded.

The Egyptian weight system appears to have been founded on a unit called the *kite,* with a decimal ratio, 10 *kites* equaling 1 *deben,* and 10 *debens* equaling 1 *sep.* Over the long span of Egyptian history, the weight of the *kite* varied from period to period, ranging all the way from 4.5 to 29.9 grams (0.16 to 1.05 ounce). About 3,400 different weights have been recovered from ancient Egypt, some in basic geometric shapes, others in human and animal forms.

Egyptian liquid measures, from large to small, were *ro, hin, hekat, khar,* and cubic cubit (0.14 cubic metre [37 U.S. gallons]).

The Babylonians. The earliest of all known weights is possibly the Babylonian mina, which in one surviving form weighed about 640 grams (23 ounces) and in another about 978 grams (34 ounces). Archaeologists have also found weights of five minas, in the shape of a duck, and a 30-mina weight in the form of a swan. The shekel, familiar from the Bible as a standard Hebrew coin and weight, was originally Babylonian. Most of the Babylonian weights and measures, carried in commerce throughout the Middle East, were gradually adopted by other countries. The basic Babylonian unit of length was the *kus* (about 530 millimetres, or 20.9 inches), also called the Babylonian cubit. The Babylonian *shusi,* defined as $1/30$ *kus,* was equal to 17.5 millimetres (0.69 inch). The Babylonian foot was $2/3$ *kus.*

The Babylonian liquid measure, *ka,* was the volume of a cube of one handbreadth (99 to 102 millimetres, or 3.9 to 4 inches). The cube, however, had to contain a weight of one great mina of water. The *ka* was a subdivision of two other units; 300 *ka* equaled 60 *gin* and one *gur.* The *gur* represented a volume of almost 303 litres (80 U.S. gallons).

The Hittites, Assyrians, Phoenicians, and Hebrews derived their systems generally from the Babylonians and Egyptians. Hebrew standards were based on the relationship between the mina, the talent (the basic unit), and the shekel. The sacred mina was equal to 60 shekels, and the sacred talent to 3,000 shekels, or 50 sacred minas. The Talmudic mina equaled 25 shekels; the Talmudic talent equaled 1,500 shekels, or 60 Talmudic minas.

The volumes of the several Hebrew standards of liquid measure are not definitely known; the *bat* may have contained about 37 litres (nearly 10 U.S. gallons); if so, the *log* equaled slightly more than 0.5 litre (0.13 U.S. gallon), and the *hin* slightly more than 6 litres (1.6 U.S. gallons). The Hebrew system was notable for the close relationship between dry and liquid volumetric measures; the liquid kor was the same size as the dry homer, and the liquid *bat* corresponded to the dry *'efa.*

Greeks and Romans. In the 1st millennium BC, commercial domination of the Mediterranean passed into the hands of the Greeks first and then the Romans. A basic Greek unit of length was the finger (19.3 millimetres, or 0.76 inch); 16 fingers equaled 30 centimetres (one foot), and 24 fingers equaled one Olympic cubit. The coincidence with the Egyptian 24 digits equaling one small cubit suggests what is altogether probable on the basis of the commercial history of the era, that the Greeks derived their measures partly from the Egyptians and partly

Use of hands and feet in early Greek measurement

from the Babylonians, probably via the Phoenicians. The Greeks apparently used linear standards to establish their primary liquid measure, the *metrētēs,* equivalent to 39.4 litres (10.4 U.S. gallons). A basic Greek unit of weight was the talent (equal to 25.8 kilograms, or 56.9 pounds), obviously borrowed from Eastern neighbours.

The Romans, adapting the Greek system, subdivided the foot into 12 ounces, or inches (unciae), using the same word and the same subdivision, unciae, for 1/12 pound (libra), which equaled 327.45 grams. The Romans made five feet equal to one pace, or double step, a thousand of which made up the Roman mile (*mille passus*). The *sextarius* (0.53 litre, or 0.14 U.S. gallon) was the basic Roman unit of volume; it had several subdivisions and multiples, of which the largest, the amphora, borrowed from the Greeks, was 48 *sextarii* (25.5 litres, or 6.7 U.S. gallons).

ANCIENT CHINESE WEIGHTS AND MEASURES

Completely separated from the Mediterranean-European history of metrology is that of ancient China; yet the Chinese system exhibits all the principal characteristics of the Western. It employed parts of the body as a source of units—for example, the distance from the pulse to the base of the thumb. It was fundamentally chaotic in that there was no relationship between different types of units, such as those of length and those of volume. Finally, it was rich in variations. The *mou,* a unit of land measure, fluctuated from region to region from 0.08 to 0.13 hectare (0.2 to 0.3 acre). Variations were not limited to the geographic; a unit of length with the same name might be of one length for a carpenter, another for a mason, and still another for a tailor.

Shih huang-ti, who became the first emperor of China in 221 BC, is celebrated for, among other things, his unification of the regulations fixing the basic units. The basic weight, the *shih,* or *tan,* was fixed at about 60 kilograms (132 pounds); the two basic measurements, the *chih* and the *chang,* were set at about 25 centimetres (9.8 inches) and 3 metres (9.8 feet), respectively. A noteworthy characteristic of the Chinese system, and one that represented a substantial advantage over the Mediterranean systems, was its predilection for a decimal notation, as demonstrated by foot rulers dating back as far as the 6th century BC. Measuring instruments, too, were of a high order.

A unique characteristic of the Chinese system was its inclusion of an acoustical dimension. A standard vessel used for measuring grain and wine was defined not only as to weight but also as to pitch when struck; given a uniform shape and fixed weight, only a vessel of the proper volume would give the proper pitch. Thus the same word in old Chinese means "wine bowl," "grain measure," and "bell." Measures based on the length of a pitch pipe and its subdivision in terms of millet grains supplanted the old measurements based on the human body. The change brought a substantial increase in accuracy.

MEDIEVAL WEIGHTS AND MEASURES

Medieval Europe inherited the Roman system, with its Greek, Babylonian, and Egyptian roots. It soon proliferated through daily use and language variations into a great number of national and regional variants, with elements borrowed from Scandinavia and from the Arabs and original contributions growing out of the needs of medieval commerce.

Charlemagne's attempts to achieve standardization

A determined effort by Charlemagne to impose uniformity at the beginning of the 9th century was in vain; differing usages hardened. The great trade fairs, such as the Champagne Fairs of the 12th and 13th centuries, enforced rigid uniformity on merchants of all nationalities within the fair grounds and had some effect on standardizing differences among regions, but the variations remained. A good example is the ell, the universal measure for wool cloth, the great trading staple of the Middle Ages. The ell of Champagne, two feet six inches, measured against an iron standard in the hands of the Keeper of the Fair, was accepted by Ypres, Ghent, Arras, and the other great cloth-manufacturing cities of northwest Europe, even though their bolts varied in length. In several other parts of Europe the ell itself varied, however.

The basic Roman unit of weight, the libra, acquired a Germanic name in parts of northern Europe but retained its Roman identity in the English abbreviation of pound as lb. Similarly the Roman mile survived, while the pace on which it was based vanished; it ceased to be a thousand of anything and instead became varying numbers of feet and yards, measures inherited from earlier northern Europe.

Medieval liquid measure was generally based on the *pinte,* or pint, which was approximately equal to the modern English quart; the quart was a medieval unit of dry measure, very close to its modern English equivalent in volume.

The English and U.S. customary weights and measures systems

ENGLISH UNITS AND STANDARDS

Out of the welter of medieval weights and measures emerged several national systems, reformed and reorganized from time to time; ultimately nearly all these were replaced by metric. In Britain and its American colonies, however, the ancient system survived.

By the time of Magna Carta (1215) abuses of weights and measures were so common that a clause was inserted in the charter to correct those on grain and wine. A few years later a royal ordinance entitled "Assize of Weights and Measures" defined a broad list of units and standards so successfully that it remained in force for nearly 600 years. A standard yard, "the Iron Yard of our Lord the King," was prescribed for the realm, divided into the traditional 3 feet, each of 12 inches, "neither more nor less." The perch (later the rod) was defined as 5½ yards, or 16½ feet. The inch was subdivided into three barley corns.

The furlong (a "furrow long") was eventually standardized as an eighth of a mile; the acre, probably from an old Anglo-Saxon word, as an area 4 rods wide by 40 long.

Troy and avoirdupois weights

The influence of the Champagne Fairs may be seen in the separate English pounds for troy weight, probably from Troyes, one of the principal fair cities, and avoirdupois, the term used at the fairs for goods that had to be weighed—sugar, salt, alum, dyes, grain. The troy pound, for weighing gold and silver bullion and apothecaries' drugs, contained only 12 troy ounces.

A multiple of the English pound was the stone, which added a fresh element of confusion to the system by equaling neither 12 nor 16 but 14 pounds. The sacks of raw wool, which were medieval England's principal export, weighed 26 stones, or 364 pounds; huge standards, weighing 91 pounds, or one-fourth a sack, were employed in wool weighing. The sets of standards, which were sent out from London to the provincial towns, were usually of bronze or brass. Discrepancies somehow crept into the system, and in 1496, following a Parliamentary inquiry, new standards were made and sent out, a procedure repeated in 1588, under Elizabeth I.

No revision of law was found necessary for 200 years after Elizabeth's time, but several refinements and redefinitions were added. Edmund Gunter, a 17th-century mathematician, conceived the idea of taking the acre's breadth (four perches, or 22 yards), calling it a chain, and dividing it into 100 links. In 1701 the corn bushel in dry measure was defined as "any round measure with a plain and even bottom, being 18½ inches wide throughout and 8 inches deep." Similarly, in 1707, the wine gallon was defined as a round measure with an even bottom and containing 231 cubic inches; however, the ale gallon was retained at 282 cubic inches. There was also a corn gallon and an older, slightly smaller wine gallon.

A new Weights and Measures Act of 1824 sought to clear away some of the medieval tangle. A single gallon was decreed, defined as the volume occupied by "10 imperial pounds weight of distilled water weighed in air against brass weights with the water and the air at a temperature of 62 degrees of Fahrenheit's thermometer and with the barometer at 30 inches."

The same definition was reiterated in the Act of 1878, which redefined the yard: "the straight line or distance between the centres of two gold plugs or pins in the bronze bar . . . measured when the bar is at the temperature of sixty-two degrees of Fahrenheit's thermometer, and when

it is supported by bronze rollers placed under it in such a manner as best to avoid flexure of the bar."

By an act of Parliament in 1963, all the English weights and measures were redefined in terms of the metric system, with a national changeover beginning two years later.

THE U.S. CUSTOMARY SYSTEM

In his first message to Congress in 1790 George Washington drew attention to the need for "uniformity in currency, weights and measures." Currency was settled in a decimal form, but the vast inertia of the English weights and measures system permeating industry and commerce and involving containers, measures, tools, and machines, as well as popular psychology, prevented the same approach, though it was advocated by Thomas Jefferson, from succeeding. In these very years the metric system was coming into being in France, and in 1821 Secretary of State John Quincy Adams, in a famous report to Congress, called the metric system "worthy of acceptance . . . beyond a question." Yet Adams admitted the impossibility of winning acceptance for it in the United States, until a future time "when the example of its benefits, long and practically enjoyed, shall acquire that ascendancy over the opinions of other nations which gives motion to the springs and direction to the wheels of the power."

Differences between the English and U.S. systems

Instead of adopting metric, the United States tried to bring its system into closer harmony with the English, from which various deviations had developed; for example, the United States still used "Queen Anne's gallon" of 231 cubic inches, which the British had discarded in 1824. Construction of standards was undertaken by the Office of Standard Weights and Measures, under the Treasury Department. The standard for the yard was one imported from London some years earlier, which guaranteed a close identity between the American and English yard; yet Queen Anne's gallon was retained. The avoirdupois pound, at 7,000 grains, exactly corresponded with the British, as did the troy pound at 5,760 grains; however, the U.S. bushel, at 2,150.42 cubic inches, again deviated from the British. The U.S. bushel was derived from the "Winchester bushel," a surviving standard dating to the fifteenth century, which had been replaced in the British Act of 1824. It might be said that the U.S. gallon and bushel, smaller by about 17 percent and 3 percent, respectively, than the British, remain more truly medieval than their British counterparts.

But at least the standards were fixed; from the mid-19th century, new states, as they were admitted to the Union, were presented with sets of standards. Late in the century pressure grew to enlarge the role of the Office of Standard Weights and Measures, which, by Act of Congress effective July 1, 1901, became the National Bureau of Standards (since 1988 the National Institute of Standards and Technology), part of the Commerce Department. Its functions, as defined by the Act of 1901, included, besides the construction of physical standards and cooperation in establishment of standard practices, such activities as: developing methods for testing materials and structures; carrying out research in engineering, physical science, and mathematics; and compilation and publication of general scientific and technical data. One of the first acts of the bureau was to sponsor a National Conference on Weights and Measures to coordinate standards among the states; one of the main functions of the annual conference became the updating of a model state law on weights and measures, which resulted in virtual uniformity in legislation.

Apart from this action, however, the U.S. government remained unique among major nations in refraining from exercising control at the national level. One noteworthy exception was the Metric Act of 1866, which permitted use of the metric system in the United States.

The metric system of weights and measures

DEVELOPMENT AND ESTABLISHMENT OF THE METRIC SYSTEM

One of the most significant results of the French Revolution was the establishment of the metric system of weights and measures.

European scientists had for many years discussed the desirability of a new, rational, and uniform system to replace the national and regional variants that made scientific communication difficult. The first proposal to closely approximate what eventually became the metric system was made as early as 1670. Gabriel Mouton, the vicar of St. Paul's Church in Lyon, suggested a length measure based on the arc of one minute of longitude, to be subdivided decimally. Mouton's proposal contained three of the major characteristics of the metric system: decimalization, rational prefixes, and the Earth's measurement as basis for a definition. Mouton's proposal was discussed, amended, criticized, and advocated for 120 years before the fall of the Bastille and the creation of the National Assembly made it a political possibility. In April of 1790 one of the foremost members of the assembly, Talleyrand, introduced the subject and launched a debate that resulted in a directive to the French Academy of Sciences to prepare a report. After several months' study, the academy recommended that the length of the meridian passing through Paris be determined from the North Pole to the equator, that one ten-millionth of this distance be termed the metre and form the basis of a new decimal linear system, and, further, that a new unit of weight should be derived from the weight of a cubic metre of water. A list of prefixes for decimal multiples and submultiples was proposed. The National Assembly endorsed the report and directed that the necessary meridional measurements be taken.

The work of Gabriel Mouton

On June 19, 1791, a committee of 12 mathematicians, geodesists, and physicists met with Louis XVI, who gave his formal approval. The next day the king attempted to escape from France, was arrested, returned to Paris and imprisoned; a year later, from his cell, he issued the proclamation that directed two engineers, Jean Delambre and Pierre Méchain, to perform the operations necessary to determining the length of the metre. The intervening time had been spent by the scientists and engineers in preliminary research; Delambre and Méchain now set to work to measure the distance on the meridian from Barcelona, Spain, to Dunkirk in northern France. The survey proved arduous; civil and foreign war so hampered the operation that it was not completed for six years. While Delambre and Méchain were struggling in the field, administrative details were being worked out in Paris. In 1793 a provisional metre was constructed from geodetic data already available. In 1795 the firm decision was taken to enact adoption of the metric system for France. The new law defined the length, mass, and capacity standards and listed the prefixes for multiples and submultiples. With the formal presentation to the assembly of the standard metre, as determined by Delambre and Méchain, the metric system became a fact in June 1799. The motto adopted for the new system was "For all people, for all time."

Establishment of the metric system

The standard metre was the Delambre–Méchain survey-derived "one ten-millionth part of a meridional quadrant of the earth." The gram, the basic unit of mass, was made equal to the mass of a cubic centimetre of pure water at the temperature of its maximum density (4° C). A platinum cylinder known as the Kilogram of the Archives was fabricated and declared the standard for 1,000 grams.

The litre was defined as the volume equivalent to the volume of a cube, each side of which had a length of one decimetre, or 10 centimetres.

The are was defined as the measure of area equal to a square 10 metres on a side. In practice the multiple hectare, 100 ares, became the principal unit of land measure.

The stere was defined as the unit of volume, equal to one cubic metre.

Names for multiples and submultiples of all units were made uniform, based on Greek prefixes.

The metric system's conquest of Europe was facilitated by the military successes of the French Revolution and Napoleon, but it required a long period of time to overcome the inertia of customary systems. Even in France Napoleon found it expedient to issue a decree permitting use of the old medieval system. Nonetheless, in the competition between the two systems existing side by side, the advantages of the metric proved decisive; in 1840 it was established as the legal monopoly in France, and from

Table 5: Elemental and Derived SI Units and Symbols

quantity	SI units		
	unit	formula	symbol
Elemental units			
Length	metre	—	m
Mass	kilogram	—	kg
Time	second	—	s
Electric current	ampere	—	A
Temperature	kelvin	—	K
Luminous intensity	candela	—	cd
Plane angle	radian	—	rad
Solid angle	steradian	—	sr
Derived units			
Acceleration	metre/second squared	m/s^2	
Area	square metre	m^2	
Capacitance	farad	$A \cdot s/V$	F
Charge	coulomb	$A \cdot s$	C
Density	kilogram/cubic metre	kg/m^3	
Electric field strength	volt/metre	V/m	
Energy	joule	$N \cdot m$	J
Force	newton	$kg \cdot m/s^2$	N
Frequency	hertz	s^{-1}	Hz
Illumination	lux	lm/m^2	lx
Inductance	henry	$V \cdot s/A$	H
Kinematic viscosity	square metre/second	m^2/s	
Luminance	candela/square metre	cd/m^2	
Luminous flux	lumen	$cd \cdot sr$	lm
Magnetic field strength	ampere/metre	A/m	
Magnetic flux	weber	$V \cdot s$	Wb
Magnetic flux density	tesla	Wb/m^2	T
Power	watt	J/s	W
Pressure	newton/square metre	N/m^2	
Resistance	ohm	V/A	Ω
Stress	newton/square metre	N/m^2	
Velocity	metre/second	m/s	
Viscosity	newton-second/square metre	$N \cdot s/m^2$	
Voltage	volt	W/A	V
Volume	cubic metre	m^3	

that point forward its progress through the world has been steady, though it is worth observing that in many cases metric was adopted during the course of a political upheaval, just as in its original French beginning. Notable examples are Latin America, the Soviet Union, and China. In Japan the adoption of metric came about following the peaceful but far-reaching political changes associated with the Meiji Restoration of 1868.

In Britain, the Commonwealth nations that still retain the British Imperial System, and the United States the progress of metric has been discernible. The United States became a signatory to the Metric Convention of 1875 and received copies of the International Prototype Metre and the International Prototype Kilogram in 1890. Three years later the Office of Weights and Measures announced that the prototype metre and kilogram would be regarded as fundamental standards from which the customary units, the yard and the pound, would be derived.

Increasing adoption of the metric system in the 20th century

Throughout the 20th century, use of the metric system in various segments of commerce and industry increased spontaneously in Britain and the United States; it became almost universally employed in the scientific and medical professions. The chemical industry, the electric power industry, photography, optometry, and the electronics industry all adopted metric at least in part. The U.S. Army adopted metric for ordnance. Legislative proposals to adopt metric generally began to be made in the U.S. Congress and British Parliament. In 1968 Congress passed legislation calling for a program of investigation, research, and survey to determine the impact on the United States of increasing worldwide use of the metric system. The program concluded with a report to Congress in July 1971 that recommended: " . . . on the basis of evidence marshalled in the U.S. metric study, this report [D.V. Simone, "A Metric America, A Decision whose Time has Come," *National Bureau of Standards Special Publication 345*] recommends that the United States change to the International Metric System through a coordinated national program over a period of ten years, at the end of which the nation will be predominantly metric." Parliament went further and took steps to begin a long-range program of changeover.

In the meantime the metric system itself was replaced by the new Système Internationale (SI), or International System, which is fundamentally an expansion of 18th-century metric to incorporate scientific and technological developments of the 20th century.

THE INTERNATIONAL SYSTEM OF UNITS

Just as the original conception of the metric system had grown out of the problems scientists encountered in dealing with the medieval system, so the new International System grew out of the problems a vastly enlarged scientific community faced in the proliferation of subsystems improvised to serve particular disciplines. At the same time, it had long been known that the original 18th-century standards were not accurate to the degree de-

Table 6: Common Equivalents and Conversion Factors for U.S. Customary and SI Systems

approximate common equivalents		conversions accurate within 10 parts per million	
1 inch	= 25 millimetres	inches × 25.4*	= millimetres
1 foot	= 0.3 metre	feet × 0.3048*	= metres
1 yard	= 0.9 metre	yards × 0.9144*	= metres
1 mile	= 1.6 kilometres	miles × 1.60934	= kilometres
1 square inch	= 6.5 square centimetres	square inches × 6.4516*	= square centimetres
1 square foot	= 0.09 square metre	square feet × 0.0929030	= square metres
1 square yard	= 0.8 square metre	square yards × 0.836127	= square metres
1 acre	= 0.4 hectare†	acres × 0.404686	= hectares
1 cubic inch	= 16 cubic centimetres	cubic inches × 16.3871	= cubic centimetres
1 cubic foot	= 0.03 cubic metre	cubic feet × 0.0283168	= cubic metres
1 cubic yard	= 0.8 cubic metre	cubic yards × 0.764555	= cubic metres
1 quart (liq)	= 1 litre†	quarts (liq) × 0.946353	= litres
1 gallon	= 0.004 cubic metre	gallons × 0.00378541	= cubic metres
1 ounce (avdp)	= 28 grams	ounces (avdp) × 28.3495	= grams
1 pound (avdp)	= 0.45 kilogram	pounds (avdp) × 0.453592	= kilograms
1 horsepower	= 0.75 kilowatt	horsepower × 0.745700	= kilowatts
1 millimetre	= 0.04 inch	millimetres × 0.0393701	= inches
1 metre	= 3.3 feet	metres × 3.28084	= feet
1 metre	= 1.1 yards	metres × 1.09361	= yards
1 kilometre	= 0.6 mile (statute)	kilometres × 0.621371	= miles (statute)
1 square centimetre	= 0.16 square inch	square centimetres × 0.155000	= square inches
1 square metre	= 11 square feet	square metres × 10.7639	= square feet
1 square metre	= 1.2 square yards	square metres × 1.19599	= square yards
1 hectare†	= 2.5 acres	hectares × 2.47105	= acres
1 cubic centimetre	= 0.06 cubic inch	cubic centimetres × 0.0610237	= cubic inches
1 cubic metre	= 35 cubic feet	cubic metres × 35.3147	= cubic feet
1 cubic metre	= 1.3 cubic yards	cubic metres × 1.30795	= cubic yards
1 litre†	= 1 quart (liq)	litres × 1.05669	= quarts (liq)
1 cubic metre	= 264 gallons	cubic metres × 264.172	= gallons
1 gram	= 0.035 ounce (avdp)	grams × 0.0352740	= ounces (avdp)
1 kilogram	= 2.2 pounds (avdp)	kilograms × 2.20462	= pounds (avdp)
1 kilowatt	= 1.3 horsepower	kilowatts × 1.34102	= horsepower

*Exact. †Common term not used in SI.
Source: National Bureau of Standards Wall Chart.

Table 7: Prefixes and Their Symbols in the SI System

multiples and submultiples	prefixes	symbols
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k
10^{2}	hecto	h
10	deka	da
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

manded by 20th-century scientific operations; new definitions were required. After lengthy discussion the 11th General Conference on Weights and Measures, meeting in Paris, birthplace of the metric system, in October 1960, formulated a new International System of Units. Six base units were adopted and defined:

Base units of SI measurement

Length: metre. Since 1983 the metre has been defined as the distance traveled by light in a vacuum in 1/299,732,458 second.

Mass: kilogram. The standard for the unit of mass, the kilogram, is a cylinder of platinum-iridium alloy kept by the International Bureau of Weights and Measures, located in Sèvres, near Paris. A duplicate in the custody of the National Bureau of Standards serves as the mass standard for the United States. (This is the only base unit still defined by an artifact.)

Time: second. The second is defined as the duration of 9,192,631,770 cycles of the radiation associated with a specified transition, or change in energy level, of the cesium atom.

Electric current: ampere. The ampere is defined as the magnitude of the current that, when flowing through each of two long parallel wires separated by one metre in free space, results in a force between the two wires (due to their magnetic fields) of 2×10^{-7} newton (the newton is a unit of force equal to about 0.2 pound) for each metre of length.

Thermodynamic temperature: kelvin. The thermodynamic, or Kelvin, scale of temperature used in SI has its origin or zero point at absolute zero and has a fixed point at the triple point of water (the temperature and pressure at which ice, liquid water, and water vapour are in equilibrium), defined as 273.16 kelvins. The Celsius scale is derived from the Kelvin scale. The triple point is defined as 0.01° on the Celsius scale, which is approximately 32.02° on the Fahrenheit scale.

Light intensity: candela. The candela is defined as the luminous intensity of 1/600,000 of a square metre of a radiating cavity at the temperature of freezing platinum (2,042 K). (L.J.C./Ed.)

BIBLIOGRAPHY

Principles and instruments: The very extensive literature on measuring instruments is scattered throughout science and technology—in the sciences that are the sources of instrument principles, in the technologies that provide its techniques, and in association with the fields of application of measurements. A key to these sources of information is provided by JULIAN F. SMITH and WILLIAM G. BROMBACHER, *Guide to Instrumentation Literature* (1965). BRIAN ELLIS, *Basic Concepts of Measurement* (1966), is an analytical treatment of the philosophy and logic of measurement. CHESTER H. PAGE and PAUL VIGOUREUX (eds.), *The International System of Units (SI)* (1972), describes the system of units and standards that are the internationally accepted basis for communication of the results of measurement. DOUGLAS M. CONSIDINE (ed.), *Encyclopedia of Instrumentation and Control* (1971), treats comprehensively the broad field of the applications of measuring instruments in research, development, manufacturing, medicine, etc. (emphasis is on guidance for the instrument user). CHARLES S. DRAPER, WALTER MCKAY, and SIDNEY LEES, *Instrument Engineering*, 3 vol. (1952–55), is a detailed and sophisticated treatment of principles and methods for solution of problems in measurement and control, with emphasis on the needs of the developer and designer of instrument systems. See also LESLIE W. LEE, *Elementary Principles of Laboratory Instruments*, 2nd ed. (1970). BERNARD M. OLIVER and JOHN M. CAGE (eds.), *Electronic Measurements and Instrumentation* (1971), covers many of the instrument system functions that follow conversion of a measurand to an electrical signal. THOMAS N. WHITEHEAD, *The Design and Use of Instruments and Accurate Mechanism* (1934, reprinted 1954), develops fundamental principles for achieving accuracy in the operation of measuring instruments. An accumulation of important papers that deal with precise, fundamental measurements is the NATIONAL BUREAU OF STANDARDS, *Precision Measurement and Calibration:* vol. 1, *Statistical Concepts and Procedures;* vol. 2, *Temperature;* vol. 3, *Electricity: Low Frequency;* vol. 4, *Electricity: Radio Frequency;* vol. 5, *Frequency and Time;* vol. 6, *Heat;* vol. 7, *Radiometry and Photometry;* vol. 8, *Mechanics;* vol. 9, *Colorimetry;* vol. 10, *Ionizing Radiation;* and vol. 11, *Image Optics* (1968–). C.F. HIX and R.P. ALLEY, *Physical Laws and Effects* (1958), includes a compilation and description of the phenomena that provide the basis for conversion of measurands into instrument signals. See also BARRY E. JONES, *Instrumentation, Measurement, and Feedback* (1977), and *Instrument Science and Technology*, 2 vol. (1982–83); FRANCIS T. FARAGO, *Handbook of Dimensional Measurements*, 2nd ed. (1983); ISAAC ASIMOV, *The Measure of the Universe* (1983); and DIAGRAM GROUP, *Comparisons: Of Distance, Size, Area, Volume, Mass, Weight, Density, Energy, Temperature, Speed and Number Throughout the Universe* (1980).

The following treat the devices that convert the measurand to instrument signals: KURT S. LION, *Instrumentation in Scientific Research: Electrical Input Transducers* (1959); HARRY N. NORTON, *Handbook of Transducers for Electronic Measuring Systems* (1969); KENNETH ARTHUR, *Transducer Measurements* (1971); GLENN F. HARVEY (ed.), *ISA Transducer Compendium*, 2nd ed., 3 vol. (1969–72); and JOHN A. ALOCCA and ALLEN STUART, *Transducers: Theory and Application* (1984).

The following treat in detail specific instrument functional elements: RICHARD F. SHEA (ed.), *Amplifier Handbook* (1966); HERMANN SCHMID, *Electronic Analog/Digital Conversions* (1970); MYRON H. NICHOLS and LAWRENCE L. RAUCH, *Radio Telemetry*, 2nd ed. (1956); CESAR A. CACERES (ed.), *Biomedical Telemetry* (1965); MISCHA SCHWARTZ, *Information Transmission, Modulation, and Noise* (1959); CLAUDE E. SHANNON and WARREN WEAVER, *The Mathematical Theory of Communication* (1949); WILLIAM R. BENNETT and JAMES R. DAVEY, *Data Transmission* (1965); C.B. PEAR, JR. (ed.), *Magnetic Recording in Science and Industry* (1967); HAROLD R. LUXENBERG and RUDOLPH L. KUEHN (eds.), *Display Systems Engineering* (1968); JOSEF CZECH, *Oszillografen-Messtechnik* (1959; rev. Eng. trans., *Oscilloscope Measuring Technique: Principles and Applications of Modern Cathode Ray Oscilloscopes*, 1965); JULIUS S. BENDAT and ALLAN G. PIERSOL, *Measurement and Analysis of Random Data* (1966); JOHN A. ALOCCA and ALLEN STUART, *Electronic Instrumentation* (1983); and G.N. PEGGS (ed.), *High Pressure Measurement Techniques* (1983). (J.Sn.)

Development and applications: F.W. KARASEK, "Notes on Analytical Instrumentation," set of 94 articles, appearing in *Research/Development* (monthly) between 1969 and 1978, that describe a wide range of instrumentation in use in science and industry; BÉLA G. LIPTÁK, *Instrument Engineers' Handbook: Process Measurement*, rev. ed. (1982), a text describing the details of instrument design and application for each of the major classes of automatic process instruments from flowmeters to analytical devices; DOUGLAS A. SKOOG and DONALD M. WEST, *Principles of Instrumental Analysis*, 2nd ed. (1980), a clearly written textbook describing instruments and their uses; ROBERT M. SILVERSTEIN and G. CLAYTON BASSLER, *Spectrometric Identification of Organic Compounds*, 4th ed. (1981), covering the fundamental aspects of spectral interpretation when combining the three spectroscopies of mass, nuclear magnetic resonance, and infrared to identify a compound; and M.H. CARR, *The Surface of Mars* (1981), a beautifully illustrated description of the Viking mission to Mars and its results. Refer also to *Scientific American* (monthly), which offers articles on various aspects of instrumentation; *Science* (weekly), a magazine written in technical style; and AMERICAN SOCIETY FOR TESTING AND MATERIALS, *Annual Book of ASTM Standards*, a work containing complete details of the instrumentation and tests that are accepted standards of industry in the United States. Specific applications are treated in *Innovations in Industrial Level Measurement* (1983), a monograph based on a conference held in London in October 1982; SIMO H. LAURILA, *Electronic Surveying in Practice* (1983), describing advanced methods; J.M. DEALY, *Rheometers for Molten Plastics* (1982), a practical guide to the measurements; and ARNE HENDEN and R.H. KAITCHUCK, *Astronomical Photometry* (1982), concentrating on optical-photoelectric photometry. (F.W.K.)

Thermometers: The most complete references for thermometers and thermometry in general, covering all phases of the subject and giving extensive bibliographies, is the series *Temperature: Its Measurement and Control in Science and Industry,* 4 vol. (1941–72). Among textbooks dealing in whole or in part with some part of thermometry, two that give the most complete coverage are H.D. BAKER, E.A. RYDER, and N.H. BAKER, *Temperature Measurement in Engineering,* 2 vol. (1953–61); and M. TERNY, *La Mesure des températures au laboratoire et dans l'industrie* (1962). A review emphasizing measurements of reasonably high precision is R.E. BEDFORD, T.M. DAUPHINEE, and H. PRESTON-THOMAS, "Temperature Measurement," in *Tools and Techniques in Physical Metallurgy,* ed. by F. WEINBERG, vol. 1, ch. 1 (1970). Cryogenic thermometry is ably reviewed in C.A. SWENSON, "Low Temperature Thermometry, 1 to 30 K," *Critical Reviews in Solid State Sciences,* 1:99–136 (1970). Two excellent accounts of the history of thermometers are H.C. BOLTON, *Evolution of the Thermometer: 1592–1743* (1900); and W.E.K. MIDDLETON, *A History of the Thermometer and its Uses in Meteorology* (1966). Two extensive bibliographies of research papers dealing with all phases of thermometry published from 1953 through 1962 are C. HALPERN and R.J. MOFFAT, "Bibliography of Temperature Measurement: January 1953 to June 1960," *National Bureau of Standards Monograph 27* (1961); and C. HALPERN, "Bibliography of Temperature Measurement: July 1960 to December 1962," *ibid.,* suppl. 1 (1963). The complete text of the IPTS-68 appears in "The International Practical Temperature Scale of 1968," *Metrologia,* 5:35–44 (1969). Later monographs include ROBERT P. BENEDICT, *Fundamentals of Temperature, Pressure, and Flow Measurements,* 2nd ed. (1977); B.F. BILLING and T.J. QUINN (eds.), *Temperature Measurement: Papers from the European Conference on Temperature Measurement Held at the National Physical Laboratory, Teddington, 9–11 April 1975* (1975); and T.J. QUINN, *Temperature* (1983). (R.E.Be.)

Microscopes: S. BRADBURY, *Evolution of the Microscope* (1967), is a general account of the history of the microscope from early days to the present. Works focusing on optical microscopes include B.D. PAYNE, *Microscope Design and Construction,* 2nd ed. (1957), a survey of the basic principles of the mechanical design of the instrument; L.C. MARTIN, *The Theory of the Microscope* (1966), an advanced mathematical treatment of the physical theory of the microscope; A.L.E. BARRON, *Using the Microscope,* 3rd ed. rev. (1965), a discussion of various types of instruments and their application; G.H. NEEDHAM, *Practical Use of the Microscope* (1965), a comprehensive treatment of the practical aspects of the use of the microscope and its accessories; W.H. BURRELLS, *Industrial Microscopy in Practice* (1961), a discussion of the industrial applications of various microscopes; G.W. WHITE, *Introduction to Microscopy* (1966), a general treatment of the instrument and its uses; V.E. COSSLETT, *Modern Microscopy* (1966), a survey of the various types of microscopes and their limitations. Special types of microscopes are treated in the following: K.F.A. ROSS, *Phase Contrast and Interference Microscopy for Cell Biologists* (1967), a simple discussion of phase and interference microscopes and their biological applications; M. FRANCON, *Progress in Microscopy* (1961), an advanced treatment of phase contrast and interference microscopy; W. KRUG, J. REINITZ, and G. SCHULZ, *Contributions to Interference Microscopy* (Eng. trans. 1964), a technical treatise on the interference microscope in all its forms; A.F. HALLIMOND, *The Manual of the Polarizing Microscope,* 3rd ed. with appendixes (1970), a treatment of the microscope emphasizing its application in mineralogy and crystallography; P. GAY, *An Introduction to Crystal Optics* (1967), a work on crystallography dealing with the application of the polarizing microscope in the identification and characterization of minerals; S. TOLANSKY, *Interference Microscopy for the Biologist* (1968), a discussion of the applications of interferometry in studies of surface contours of biological materials; V.E. COSSLETT and W.C. NIXON, *X-ray Microscopy* (1960), a detailed treatise on the X-ray microscope and its potential; P. GRIVET *et al., Electron Optics,* rev. ed. (Eng. trans. 1965), a detailed treatment of the theoretical background of electron optical devices; M.E. HAINE and V.E. COSSLETT, *The Electron Microscope* (1961), a comprehensive discussion of the design of electron microscopes; C.E. HALL, *Introduction to Electron Microscopy,* 2nd ed. (1966), a general text on the theory and applications of electron microscopy; P.R. THORNTON, *Scanning Electron Microscopy* (1968), a thorough treatment of the theory of the scanning electron microscope and its applications in material science; E.W. MULLER and T.T. TSONG, *Field Ion Microscopy* (1969), a work on the theory, practice, and applications of field-ion microscopes; T. MULVEY, "The History of the Electron Microscope," in S. BRADBURY and G.L.E. TURNER (eds.), *Historical Aspects of Microscopy* (1967); and G.A. MEEK, *Practical Electron Microscopy for Biologists* (1970), a good practical account of the instrument and its operation. See also AARON E. KLEIN, *The Complete Beginner's Guide to Microscopes and Telescopes* (1980); MARCEL V. LACQUIN and M. LANGERON, *Handbook of Microscopy* (1983); MICHAEL SPENCER, *Fundamentals of Light Microscopy* (1982); K.E. CARR and P.G. TONER, *Cell Structure: An Introduction to Biomedical Electron Microscopy* (1983); and GERARD L'E. TURNER, *Collecting Microscopes* (1981). (S.Br./Ed.)

Telescopes: Techniques of observation with the help of telescopes are detailed in P. CLAY SHERROD and THOMAS L. KOED, *A Complete Manual of Amateur Astronomy: Tools and Techniques for Astronomical Observations* (1981), including a discussion of telescope setup; ROBERT T. DIXON, *Dynamic Astronomy,* 5th ed. (1989), a comprehensive, well-illustrated text; and JAY M. PASACHOFF, *Astronomy, from the Earth to the Universe,* 3rd ed. (1987). A highly readable and profusely illustrated discussion of radio telescopes is given in GERRIT L. VERSCHUUR, *The Invisible Universe Revealed: The Story of Radio Astronomy* (1987). More-technical accounts are offered in JOHN D. KRAUS *et al., Radio Astronomy,* 2nd ed. (1986); and W.N. CHRISTIANSEN and J.A. HÖGBOM, *Radiotelescopes,* 2nd ed. (1985). BERNARD LOVELL, *The Story of Jodrell Bank* (1968), presents a dramatic story of the construction of the 76-metre Jodrell Bank Radio Telescope. Other modern types of telescopes are described in D.S. FINLEY *et al.,* "Design of the Extreme Ultraviolet Explorer Long-Wavelength Grazing Incidence Telescope Optics," *Applied Optics* 27(8):1476–80 (April 15, 1988); ADELAIDE HEWITT (ed.), *Optical and Infrared Telescopes for the 1990s,* 2 vol. (1980); and JOSEPH F. BAUGHER, *The Space-Age Solar System* (1988). Numerous articles on various types of telescopes are found in *Sky and Telescope* (monthly). Historical surveys tracing the development of the telescope from the earliest versions to those constructed during the late 1980s include ARTHUR BERRY, *A Short History of Astronomy* (1898, reprinted 1961); CHARLES ALBERT REICHEN, *A History of Astronomy* (1963); HENRY C. KING, *The History of the Telescope* (1955, reprinted 1979); and RICHARD LEARNER, *Astronomy Through the Telescope* (1981). (B.L.Kl./K.I.K.)

Sonar: F.V. HUNT, *Electroacoustics: The Analysis of Transduction, and its Historical Background* (1954), covers the early history of sonar extremely well. An elementary textbook is D.G. TUCKER, *Underwater Observation Using Sonar* (1966); more-advanced textbooks include: D.G. TUCKER and B.K. GAZEY, *Applied Underwater Acoustics* (1966); J.W. HORTON, *Fundamentals of Sonar* (1957); and R.J. URICK, *Principles of Underwater Sound for Engineers,* 3rd ed. (1983). Sophisticated sonar techniques are described in a simple way in D.G. TUCKER, *Sonar in Fisheries: A Forward Look* (1967). (D.G.T./Ed.)

Systems of weights and measures: For an extensive listing of references, see J.T. MILEK and V. ANTOINE (comps.), *Bibliography of the Metric System* (1968). Some of the more important works are A.E. BERRIMAN, *Historical Metrology: A New Analysis of the Archaeological and Historical Evidence Relating to Weights and Measures* (1964, orig. pub. in 1953), one of the few works on this subject; H.W. CHISHOLM, *On the Science of Weighing and Measuring and Standards of Measure and Weight* (1877), an interesting early treatment by a former Warden of the Standards in London; L.J. CHISHOLM, *Units of Weight and Measure,* (1967, reprinted 1974), a brief history of the metric system, with discussion and definitions of U.S. and metric units, especially valuable for its extensive tables of conversion equivalents; R.C. COCHRANE, *Measures for Progress* (1966), a history of the U.S. National Bureau of Standards emphasizing the extent of NBS involvement in 20th-century scientific progress; W. HALLOCK and H.T. WADE, *Outlines of the Evolution of Weights and Measures and the Metric System* (1906), an excellent account of early metric history; L.D. JACKSON, *Modern Metrology* (1882), a detailed history of weights and measures units, with valuable tables of comparative equivalents; LEWIS V. JUDSON, *Weights and Measures Standards of the United States,* updated ed. (1976), an excellent history of U.S. physical standards; A.E. KENNELLY, *Vestiges of Pre-Metric Weights and Measures* (1928), a brief metric history and discussion of pre-metric units; B. KISCH, *Scales and Weights: A Historical Outline* (1965), an excellent treatment of this subject; E. NICHOLSON, *Men and Measures* (1912), a somewhat rambling but readable history of weights and measures in England; RALPH W. SMITH, *The Federal Basis for Weights and Measures,* (1958), a detailed and accurate history of the complete development of U.S. weights and measures; and C. WARREN, *The Early Weights and Measures of Mankind* (1913), a general historical treatment. See also INSTRUMENT SOCIETY OF AMERICA, *Landmarks in Metrology—1983* (1983); and *The World Measurement Guide: Editorial Information Compiled by THE ECONOMIST,* 4th rev. ed. (1982). (L.J.C.)

Measurement Theory

Measurement theory studies the problems in associating numbers with objects and empirical phenomena. It specifies what relations among objects are faithfully represented by assigning numerical quantities to those objects. It also explains the ways in which different measures of the same quantity are related to one another and considers the problems of error in the measurement process.

This article is divided into the following sections:

History and nature of measurement theory 792
Axiomatic basis of measurement 793
  Axioms of order 793
  Axioms of extension 793
  Axioms of difference 793
  Axioms of conjointness 794
  Axioms of geometry 794
General principles and problems 794
  Dimensions and units of measurement 794
  Measurement via nonmeasurement theories 795
  The problem of error 795
Measurement in special disciplines 796
  Measurement in physics 796
  Measurement in the social sciences and psychology 797
Bibliography 797

HISTORY AND NATURE OF MEASUREMENT THEORY

History of magnitudes theory

An early and relatively abstract theory of magnitudes was developed in the 4th century BC by two mathematicians, Eudoxus of Cnidus, probably a member of Plato's Academy, and Theaetetus, author of the book on the regular solids and one of those on irrational numbers in Euclid's *Elements,* in which their theory of magnitudes was given its fully developed presentation. Theirs was a highly sophisticated theory of measurement. Historical evidence suggests that it was developed for purely theoretical purposes, to account for incommensurable magnitudes—pairs of numbers, such as 1 and the square root of 2, for which no third number exists of which both are even (or integral) multiples—rather than to solve any practical problems, such as those of surveying. Problems of measurement, of both a theoretical and a practical sort, continued to be discussed throughout ancient times and during the Middle Ages.

In modern times, Euclid's view of the theory of magnitudes was still relevant as late as the 18th century; it appeared, for example, in Isaac Newton's *Arithmetica Universalis* (2nd ed., 1728). The modern, axiomatic treatment of the theory—*i.e.,* its logical development from fundamental concepts and axioms—can be dated from a famous paper by Hermann von Helmholtz, a German physicist and physiologist, entitled "Zählen und Messen erkenntnistheoretisch betrachtet" (1887; *Counting and Measuring*). Considerable impetus was given by a classical work by Ludwig Otto Hölder, a German mathematician, entitled "Die Axiome der Quantität und die Lehre vom Mass" (1901; "The Axioms of Quantity and the Theory of Measure"), in which he provided two sets of axioms, one for the measurement (or, as he said, "the quantity") of additive attributes, such as mass, which can be built up by addition of parts to form a total bulk, and another set for nonadditive attributes involving the measurement of intervals, as in temperature or longitude. The tradition of Helmholtz and Hölder has been of importance primarily in the physical sciences.

The consideration of measurement in the social sciences, especially in economics, also has had a long history. As early as the 18th century, Jeremy Bentham, a British Utilitarian moralist, in the treatise *The Principles of Morals and Legislation* (1789), attempted to provide a theory for the measurement of utility, or value. In the 19th century there was an effort to establish that much of economic theory could be developed solely on the basis of ordinal preferences among the economic goods being studied, without any more demanding kind of numerical measurement. The work of Vilfredo Pareto, an Italian economist and sociologist, entitled *Manuale di economia politica* (1906; "Handbook of Political Economics"), was especially important in this connection. In the mid-20th century the study of measurement in economics and in psychology was stimulated by the theory of utility set forth in *Theory of Games and Economic Behavior* (1944) by the mathematician John von Neumann, a scientist of unusual breadth, and the economist Oskar Morgenstern. It makes use of the formation of so-called risky alternatives (see below *Economics and psychology*).

The three basic problems of measurement theory

An examination of specific examples of axiomatic measurement can lead to a general view of measurement theory. The first problem for any such theory is to justify the assignment of numbers to objects or phenomena—to pass from empirical procedures and operations to a numerical representation of these procedures. From an axiomatic standpoint, the problem—known as the representation problem—is first to characterize the formal or abstract properties of these procedures and observations and then to show mathematically that these axioms permit the construction of a numerical assignment in which familiar abstract relations and operations, such as "is greater than or equal to" (symbolized $\geq$) and "plus" (+), correspond structurally to the empirical (or concrete) relations and operations.

The second fundamental problem is that of the uniqueness of the representation—*i.e.,* how close it is to being the only possible representation of its type. The representation of mass, for example, is unique in every respect except the choice of unit; *e.g.,* the numerical representation of the mass of a given object will differ depending on the choice of the pound or the gram as the unit. Ordinary measurements of temperature are unique in everything except the choice of both unit and origin—the Celsius and Fahrenheit scales differ not only in the size of unit but also in the zero point. A U.S. psychophysicist, S. Smith Stevens, was among the first to place great emphasis on the uniqueness of measurement in relation to the problem of units and on the consequences that this uniqueness has for handling data.

A subject closely related to the uniqueness of physical measurement is dimensional analysis, which, in a broad sense, attempts to explain why the various physical measures exhibit simple relations in the fundamental equations of classical physics. If length, time, and mass are taken to be the fundamental dimensions of mechanics, for example, then all other quantities, such as force or momentum, can be represented simply as products of powers of these dimensions—a fact that has strong implications for the forms of physical laws (see below *Universal, system, and material constants*).

In 1981 a general classification of measurement structures was proposed that is based on properties of the mathematical group of transformations that relate one representation to another or, what is equivalent, to the group of automorphisms of the structure (*i.e.,* isomorphisms of the structure with itself). This classification has proved fruitful. Structures having representations involving all of the real numbers turn out to have only a limited number of possible scale types: the several cases discussed by Stevens plus a few others. In turn, with this known, it has proved possible to describe quite fully whole classes of structures that exhibit these scale types. A catalog of the

possible qualitative structures that give rise to unidimensional measurement is rapidly being developed.

A third central problem for the theory of measurement is error. In spite of the early development of precise astronomical measurement, no systematic theory of observational error appears to have been developed until the 18th century, first in an article by Thomas Simpson, an English mathematician, in 1757 and then in the fundamental work of two French mathematicians, Joseph-Louis, comte de Lagrange, who worked on the theory of numbers and celestial mechanics, and Pierre-Simon, marquis de Laplace, also an astronomer, famous for his fundamental work, *Traité de méchanique céleste* (1798–1827; *Celestial Mechanics*). Today, any significant physical measurement is routinely reported with some indication of the probable error, and theories are tested by confirming them within the errors of measurement. In the classical tests of Albert Einstein's general theory of relativity, for example, the discrepancies between predictions and observations fell mainly within the estimated errors of measurement (for further discussion of errors of measurement, see below *The problem of error*).

AXIOMATIC BASIS OF MEASUREMENT

Axiom systems for measurement differ among themselves in the amount and type of structure that characterize them. All include an ordering relation, but that alone is not enough to formulate many scientific laws because its numerical representation is not closely prescribed. Additional structure and an increased uniqueness of the representation arise either by having an empirical operation of addition (extensive measurement), by having entities that have several independent components (difference and conjoint measurement), or by other primitives that lead to a geometric representation.

Axioms for binary relations

Axioms of order. Whenever a measurement is made, it is made in such a way that the order induced on the objects by the assigned measure is the same as that obtained by the basic empirical operation in question. Such measures are said to be order preserving. Of course, every numerical inequality involved in measurement is transitive (*i.e.,* such that $x \geq y$ and $y \geq z$ implies that $x \geq z$) and connected (*i.e.,* either $x \geq y$ or $y \geq x$). It is therefore necessary that the empirical ordering—symbolized $\gtrsim$ (with $\sim$ instead of $-$ to suggest indistinguishability)—be transitive and connected in order for a representation of it to be possible by means of $\geq$. Relations exhibiting these two properties are generally called weak orders, in which "weak" simply means that indifference (symbolized $\sim$)—in which $x \sim y \equiv x \gtrsim y$ and $y \gtrsim x$ (using $\equiv$ to mean "if and only if")—is not necessarily equality. Not every weak order, however, has an order-preserving numerical representation; the lexicographic order in the plane $(x,y) \gtrsim (x',y') \equiv x > x'$ or $x = x'$ and $y \gtrsim y'$, is a counterexample. A second, independent property of numbers, which must in turn be reflected in the empirical ordering, is that rational numbers—numbers that can be expressed as the quotient of two integers—are order-dense (*i.e.,* they are such that between any two distinct real numbers lies a rational number) and countable (*i.e.,* they can be placed in one-to-one correspondence with the integers).

In 1895 it was shown by Georg Cantor, a German mathematician, that, for any empirical ordering, the existence of a numerical representation that preserves its order is equivalent to the ordering's being a weak one that includes in its domain a countable, order-dense subset. Any two such representations are so related that one can be mapped upon the other in a strictly increasing fashion.

Though various ordinal categorizations are widely used (*e.g.,* brightness of stars, magnitude of earthquakes, hardness of minerals), the fact that equivalent ordinal representations are seldom in simple proportion to one another makes them unsatisfactory in the statement of many scientific laws. To increase the uniqueness of the representation, some structure in addition to order must be preserved by the representation.

Axioms of extension. For many physical attributes—including mass, length, time duration, and probability—the objects or events exhibiting the attribute may be combined, or concatenated, to form new objects or events that also exhibit the attribute. Both alone and combined, for example, objects have mass. Denoting the concatenation of a and b by $a \circ b$, the assignment (designated φ) of a given set of numbers to the objects is called an additive representation of the empirical ordering $\gtrsim$ and the concatenation o, provided that it is not only order preserving but also additive in the sense that $\varphi(a \circ b) = \varphi(a) + \varphi(b)$. Attributes that have an operation for which an additive representation exists are called extensive.

Order and concatenation

Theories that are intended to account for the existence of such numerical representations must state empirical laws about ordering and concatenation separately, as well as how they interrelate. The key property of an empirical ordering is that it is a weak order, and that of a concatenation is that it is insensitive to the order of combination—concatenation is weakly commutative (*i.e.,* its order can be reversed), $a \circ b \sim b \circ a$, and weakly associative (*i.e.,* its terms can be regrouped), $(a \circ b) \circ c \sim a \circ (b \circ c)$—in which $\sim$ indicates an equivalence relation. The key property relating them, weak monotonicity, is that the empirical ordering $a \gtrsim b$ holds if and only if the ordering of its concatenation with any object c, $a \circ c \gtrsim b \circ c$, also holds. As in ordinal measurement, a further and more subtle property, called Archimedean, is needed. This property is formally analogous to the numerical Archimedean property that if $x > y > 0$, then for some integer n, $ny \geq x$ (meaning y is not "infinitesimally" small relative to x). With these stipulations, the numerical representation can be constructed.

The basic idea of the construction is simple and, in somewhat modified form, is widely used to carry out approximate measurement. For example, one measures the length of an object a (say, a house stud) to an accuracy corresponding to some (small) object u (say, the quarter-inch marked on a tape measure) as follows. One counts the number, m, of copies of u that strung together bracket a in the sense that $(m+1)u \geq a \geq mu$. To improve accuracy, a smaller unit is usually selected, but that is equivalent to stringing together n copies of a and finding the number $m(n)$ of copies of u that approximates na. The limit toward which $m(n)/n$ tends as n approaches infinity exists and is defined to be $\varphi(a)$, which is the measurement sought; it can be shown to be order preserving and additive. In contrast to ordinal measurement, such an additive representation is rare. It is determined except for a positive multiplicative constant or coefficient setting the scale; in other words, only the choice of the unit is arbitrary. As a result, a representation of this sort is referred to as a ratio scale.

Sequences such as na, $n = 1, 2, \ldots$, being ubiquitous in measurement practice, are called standard sequences. Practical examples are standard sets of weights in multiples of a gram and the metre rule, subdivided into millimetres.

For some purposes, including the development of other kinds of measurement—*e.g.,* difference and probability measurement—it is necessary to generalize the theory to cover concatenation operations (o) that are defined only for some pairs of objects and not for others. Probability, for example, is not additive over all pairs of events, but only over those that are disjoint.

In the late 1970s and early 1980s there was developed a general axiomatic theory of concatenation structures that do not have additive representations. In these structures neither commutativity, associativity, nor positivity is assumed, but only weak order, monotonicity, Archimedeaness (of two different types), and some degree of solvability. Those concatenation structures that have representations at least as rich as a ratio scale divide into five major types that have been rather fully characterized.

Axioms for quaternary relations

Axioms of difference. Length, but not mass, may be treated also in terms of intervals (a,b) on a line. Here the empirical ordering $\gtrsim$ is a quaternary rather than a binary relation since it involves the comparison of two pairs, and so four points, such as the end points of two intervals; and concatenation is defined only for adjacent intervals; *i.e.,*

$$(a,b) \circ (b,c) = (a,c).$$

In addition to weak ordering, solvability, and Archimedean

properties, various other axioms, depending on the exact representation to be constructed, must be satisfied. An important representation is that of absolute difference, in which an empirical ordering holds if and only if the absolute values (symbolized $| \ldots |$) of such differences are ordered; *i.e.*,

$$(a,b) \succsim (c,d) \equiv |\varphi(a) - \varphi(b)| \geq |\varphi(c) - \varphi(d)| ,$$

for which key axioms are that any interval is at least as large as the null interval; *i.e.*, $(a,b) \succsim (a,a)$, and that an interval is unordered, or the same when measured "backwards"; *i.e.*, $(a,b) \sim (b,a)$.

Such representations are called interval scales because they are unique in every respect except for the possibilities of multiplication by a positive constant and addition of a constant; *i.e.*, both the unit and zero of the measure are arbitrary, and the equality of intervals is preserved through all transformations generated by altering these two scale factors.

Axioms of conjointness. A major hindrance to the development of fundamental measurement outside physics, other than probability, has been the failure to uncover suitable empirical concatenation operations for attributes such as utility, loudness, intelligence, and hunger. An alternative approach, implicit in much derived measurement in classical physics, was made completely explicit in the behavioral sciences in the 1960s. It rests on the fact that many—indeed, most—attributes are defined for entities having at least two independent components, each of which affects the attribute. Kinetic energy and momentum, for example, are both affected by mass and velocity; mass is varied by changing either density or volume; preference among gambles is manipulated both by outcomes and by the probabilities of their occurring; and loudness of tones depends on frequency as well as intensity. In conjoint-measurement theories, the way in which each component affects the attribute in question is studied by discovering which changes must be made in one component to compensate for changes in the other. When measures for the components already exist, the exchange relationship between the two components is often presented graphically in the form of so-called equal-attribute (or indifference) curves.

Additive and polynomial representations

The simplest representation is additive—or, by making an exponential transformation, multiplicative (which is the way it is usually represented in physics)—in the sense that numerical functions φ_1 and φ_2 of the two components exist such that an empirical inequality holds if and only if the sums of these functions of its elements exhibit a numerical inequality in the same direction; *i.e.*,

$$(a_1,a_2) \succsim (b_1,b_2) \equiv \varphi_1(a_1) + \varphi_2(a_2) \geq \varphi_1(b_1) + \varphi_2(b_2).$$

Axioms sufficient for such a numerical representation consist of weak ordering, a form of solvability, an Archimedean property, and what amount to analogues of weak monotonicity. The simplest of these, independence, states that an inequality of one component with a common value—*e.g.*, $(a_1,x) \succsim (b_1,x)$—is not altered by substituting any other common value y for x. When this property is appropriately generalized to three or more components, no more stipulations are needed; for only two components, however, it must be supplemented by another property, called the Thomsen condition.

General nonadditive representations also have been studied. The first to be worked out were certain simple polynomials on several components. For three components there are four such simple polynomials: additive $(\varphi_1 + \varphi_2 + \varphi_3)$, multiplicative $(\varphi_1\varphi_2\varphi_3)$, or a combination of both, either $(\varphi_1\varphi_2 + \varphi_3)$ or $\varphi_1(\varphi_2 + \varphi_3)$. (In these expressions the subscripted indices 1, 2, 3 can, of course, appear in any of their permutations.)

Axioms of geometry. Historically, numerical representation and uniqueness theorems for geometry have been at least as important as those for extensive measurement and more important than the other kinds of measurement considered above. Since the discovery of analytic geometry in the 17th century by René Descartes, an important early modern philosopher, the representation of a geometric point by a pair of numbers (the coordinates), in the case of a plane, or a triple of numbers, in the case of space, has been of fundamental importance not only in geometry but also in much of the physical sciences.

The view that the formal adequacy of axioms for geometry is established by proving a representation theorem became explicit in the latter part of the 19th century. The most important single work may have been the classic *Grundlagen der Geometrie* (1899; *The Foundations of Geometry*), by David Hilbert. Today the distinctive feature of geometric measurement is that the representation is made in terms of pairs or triples or, in some applications in the social sciences, n-tuples of numbers, rather than in terms of simple numbers, as in extensive, difference, or concatenation measurement.

Much of the literature on the foundations of geometry has been devoted to showing the different sorts of qualitative concepts that can be taken as basic or primitive in stating the axioms of geometry. One of the simplest sets of axioms, given by Alfred Tarski, a Polish-U.S. mathematician and logician, employs the concept of equidistance and that of betweenness for three points that lie on a line. Many of the properties of the quaternary relation of absolute difference discussed above (see *Axioms of difference*) are also properties of equidistance viewed as a quaternary relation among four points, viz., $(a,b) \sim (c,d)$ just when the distance between a and b is equivalent to that between c and d.

GENERAL PRINCIPLES AND PROBLEMS

Dimensions and units of measurement. Each measurable attribute constitutes a dimension; inasmuch as these are not all independent, some can be expressed as functions of others—a fact that underlies both the method of dimensional analysis and the existence of sets of units.

Dimensions and constants

Dimensions and their algebra. Many attributes are extensively measurable. Some dimensions, however, such as density, are not extensive, and in such cases a law can be stipulated that allows the derived, nonextensive measure to be expressed as a product of powers of two extensive measures. This possibility arises from two empirical facts. First, mass varies with both substance and volume, and the ordering that it induces over substance–volume pairs is such that a multiplicative conjoint representation exists. Thus, the conjoint measure of substance can be expressed as a product of its measure of mass and the reciprocal of its measure of volume. Both of these attributes, of course, have independent extensive measures. Second, a qualitative law, known as a law of distribution, relates concatenations of volume to those of mass via the conjoint ordering. From this law it is possible to prove that the conjoint and extensive measures are power functions of each other; indeed, in this case the conjoint measure of substance is, for an appropriate choice of exponent, simply the ratio of the extensive measure of mass to that of volume; *i.e.*, density. In cases such as the dependency of kinetic energy and momentum on mass and velocity, the conjoint components are both extensive and the law relating them via the conjoint ordering also is called a law of distribution; interesting cases of such laws appear to involve quantities that enter laws of conservation, such as that of momentum and energy. In any event, all nonextensive measures of classical physics are expressible as products of powers of extensive ones. Some such model is the usual starting point of dimensional analysis. In the 1980s it was shown that any operation, associative or not, that has a ratio scale representation is distributive in a suitably solvable, Archimedean conjoint structure if and only if the conjoint structure is additive. This means that the physical representation of conjoint scales as products of powers of extensive scales has been extended to a much wider class of concatenation structures so long as they have ratio scale representations, and this theory may permit a generalization of the classical structure of units beyond physics.

Universal, system, and material constants. Sometimes two apparently distinct attributes are not independent but covariant. In the classic example of inertial mass and gravitational mass, the covariation is described by a constant known as the universal gravitational constant.

This example should be contrasted with the covariation of two measures for specific systems. Mass and volume, for example, are exactly proportional for any homogeneous substance, or length and force (within limits) for a specific spring. The constants describing such local covariation, called material or system constants (depending on the context), constitute a form of derived measurement.

Many more complex physical laws may be described as stating combinations of values—configurations—of certain dimensions that can obtain in a particular class of physical system. Such laws include not only measurable dimensions of the system but also system and material constants characteristic of the particular system in question. A curious fact of physical theory is a principle first explicitly enunciated in 1914 by the U.S. analyst Edgar Buckingham, the so-called π-theorem, according to which, when the dimensional measures and constants of some physical law are grouped into one or more terms in which the dimensions all cancel out, some function of these dimensionless quantities (or π-arguments) must equal zero for any realizable configuration of the system. Such dimensional invariance was shown in the late 1970s to be nothing more nor less than an application of the concept of meaningfulness in the sense of a law's being invariant under the automorphisms of the structure. This concept is similar to the definition of a geometric object as something that is invariant under the automorphisms of the geometry. Current work shows how this concept of meaningfulness is related to the idea of a law's being expressed in terms of the primitives of the measurement structures involved.

Dimensional analysis. In dimensional analysis the investigator, in part, attempts to discover a particular physical law by assuming that it has the character of a π-theorem and that the relevant variables and constants are known. When they are known (an important proviso), a simple calculus permits him to discover all of the dimensionless terms. In many cases this yields considerable insight into the law describing the system; sometimes empirical observations, as in a wind tunnel, are used to obtain an empirical approximation to the unknown function of the π-arguments.

Units of measurement

Primary and derived units. One is free to choose the unit of measurement for each dimension, whereupon all other values are uniquely determined. Since some dimensions are related to others as products of powers, considerable simplicity is effected by choosing arbitrarily only the units of the dimensions in a maximum set of independent dimensions (base) and then letting the known dependencies determine all of the other units. Such a system of units is called coherent; those of the base are said to be primary, and all others are derived, or secondary. At the Tenth General Conference on Weights and Measures (1954), length, mass, time duration, temperature, and either charge or current were adopted as the base dimensions, with units: metre, kilogram, second, degree absolute (kelvin), and the coulomb or ampere, respectively. This system is abbreviated MKSA.

A unit is ideally defined in terms of some highly invariant and readily reproduced or observable natural phenomenon, such as the wavelength of a highly reproducible, monochromatic light source. Less ideal, although still used, are carefully made and maintained unique objects, such as the standard metre in Paris.

Measurement via nonmeasurement theories. Few actual measurements are carried out in a fundamental manner using only ordering and concatenation information because such procedures, when possible, are too slow and expensive, and in fact they are not always possible. It will be helpful to describe some of the alternative procedures employed in physics and psychology.

Indirect measurement

In physics. At least three kinds of indirect measurement occur in physics. First, permanent standard sequences are sometimes constructed—as for weight and length—and measurement then involves approximate matching against members of the sequence. Precision and standardization of conditions are important in making such measurements.

Second, when a law is found to hold between two or more variables, and system or material constants enter—such as the density and viscosity of a fluid—these constants are measured indirectly in terms of other measures by means of the law.

Third, various theories of specific physical phenomena may provide indirect means of measuring an attribute that may otherwise be difficult to measure; such theories may, in some cases, lead to convenient measuring devices. A displacement in a piece of equipment—a spring balance, for example, or a voltmeter—may be systematically related to some other attribute (force or potential). Another example is provided by the measurement of astronomical distances in terms of the shift of the object's spectral lines toward the red, a technique that is based on a theory of the existence of a systematic relation between the velocity of recession and the distance of a galaxy and one about the Doppler effect, which spreads out the waves from a receding light source and thus reduces their frequency.

In psychology. Since few, if any, psychological attributes are currently measured fundamentally, no approximate measurement using standard sequences is possible. Whether conjoint measurement will alter this situation is not yet clear.

Most formalized theory in psychology is stated in terms of physical characteristics of the stimuli and responses: various physical and temporal aspects of the stimuli, the physical nature and probability structure of the rewards provided for appropriate performance, and the time and nature of the response made. Frequently, some of the data are collapsed into relative frequencies, which are interpreted as estimates of conditional probabilities. Though, in principle, a theory could simply assert the dependence of response measures upon the various physical measures of the independent variables, this has not proved practicable; in most theories some sort of internal representation of the stimuli is postulated to exist in the subject. Decision processes are assumed to operate on these representations, which decisions then lead to the response. Such models include free parameters that often have natural interpretations as psychological measures. If these can be estimated from the observable data, as is frequently possible, they constitute a type of indirect psychological measurement not unlike the second class of physical ones. In many cases the parameters seem to play a role much like that of time constants in physical systems.

There is no satisfactory analogue to the third method enumerated above for physics because no sufficiently elaborate and accepted psychological theories yet exist. Many so-called operational measures of psychological concepts—hours of deprivation for hunger or skin resistance for anxiety—pretend to be of this type, but they are not. The absence of questions of calibration is but one indication of the inadequacy of such measures.

The problem of error. Inasmuch as measurement, in practice, is always fraught with error, the types and basic theory of error must be examined.

Types of error. The classification of errors used in the theory of observations has been of great importance, both historically and practically, especially in astronomy. In that theory, four kinds of errors of observation are ordinarily distinguished. There are instrumental errors, such as those of the telescope in astronomy, that have as their source inaccuracies in the instruments used for observation. There are personal errors, sometimes called errors due to the personal equation, that arise from the different practices and reactions of human observers. Nevil Maskelyne, the fifth astronomer royal of England, provided a famous example of such personal differences in observation when he discharged his assistant in 1796 because he had observed the transits of stars and planets about a half a second later than Maskelyne himself. Systematic errors are a third kind of error of observation. A typical example of this kind would be the measurement of atmospheric pressures corrected to sea level in terms of a faulty value for the height of the station. Such an error would introduce a systematic discrepancy in all observations of atmospheric pressure for that station. The fourth and most important kind of error of observation is that of random error, so called because its causes are not understood. This type of error reflects the fact that, when observers make repeated

Errors of observation

measurements, some variation in the results will occur no matter how accurate the instruments may be.

In principle, all four types of error need to be taken into account in refining observations. In practice, however, the attention of theorists has been directed mainly to random errors. Since the end of the 18th century, the application of probability theory in the estimation of random errors has been extensive.

Other kinds of error

Next to errors of observation, perhaps the most important kind of error is that due to sampling. If an investigator, for instance, wishes to know something about the television-viewing habits of the population, he ordinarily cannot plan to observe the viewing habits of each and every member of the population. He selects, instead, a sample of the population and from that sample infers the viewing habits of all. Errors of sampling, of course, are expected, but, the larger the sample, other things being equal, the smaller the sampling error will be. An elaborate statistical theory, both for constructing the kind of sample that should be drawn and for making inferences from the sample data to the characteristics of the population, is widely used (see STATISTICS).

Still another classification of errors involves the division into direct and indirect errors. If a surveyor, for example, measures the angles of a triangle, direct errors can arise from the measurement of the first two angles. These errors will induce errors of measurement indirectly into the value of the third angle, which he will compute instead of measuring directly. In advanced scientific practice, a large number of errors of measurement are indirect because elaborate theory and a wide variety of ancillary measurements enter into the measurements of significant quantities, such as the speed of light or the mass of the Earth.

Still other kinds of errors are concerned with interpolation and extrapolation. In using mathematical tables, for example, one often wants to read a table to one more decimal place than those tabulated, and it is common to make a linear interpolation between two adjacent figures in order to extend the table. Since the curve reflected in the figures, however, may in fact be parabolic or exponential or the like, such an assumption of linearity may generate error. Similar problems arise in extrapolating beyond the limits of a given set of values. In both cases it is possible to compute the approximate error introduced.

Theory of errors. The theory of errors has received a great deal of attention in the development of science and mathematics, especially in that of probability theory. Though, historically, the concept of standard error has been most widely used in reporting the results of measurement, in modern statistics the concept is of less importance. The computation of the standard error is based on the assumption that repeated observations of the same phenomenon—for example, the speed of light—are to be assigned equal weight. It is useful to explore a case of n observations of a phenomenon, $x_1, \ldots, x_n$. If $\bar{x}$ is the mean of these observations—*i.e.*, $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$—and s^2 is the sample variance, defined as the mean squared deviation modified by the factor $n/(n-1)$ in order to make it an unbiased estimator—*i.e.*,

Standard error and variance

$$s^2 = [(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2]/(n-1),$$

then the standard error of $\bar{x}$ is $s/\sqrt{n}$, and in most cases the "true" value of the observed variable lies in the interval between $\bar{x}$ plus $s/\sqrt{n}$ and $\bar{x}$ minus $s/\sqrt{n}$. A more detailed interpretation of the standard error and a critical scrutiny of the assumptions underlying its use are beyond the scope of this article (see STATISTICS; also see below *Bibliography*, the work by Isaac Todhunter). A few brief remarks will have to suffice here.

Under very general assumptions, it may be shown that the sum of independent random errors of observation has, as the number of terms becomes large, a normal (bell-shaped) distribution with mean μ and standard deviation σ, in which μ is the sum of the means of the individual sources of error, and σ^2 is the sum of their individual variances (standard deviations squared); *i.e.*, $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_q^2$, in which there are q sources of error. (For discussion of the normal distribution with population mean μ and variance σ, see STATISTICS.) This formulation, known as the central limit theorem, is one of the fundamental theorems of modern probability theory. Its importance lies in the fact that it is largely independent of the particular character or distribution of the individual errors. It was first stated by Laplace in 1812 and given its first rigorous proof in 1901 under fairly general assumptions by the Russian Aleksandr Lyapunov, known for his mathematical studies of stability and probability.

Probabilistic models

In recent years different probabilistic models have been developed for the study of human behaviour, especially behaviour that expresses a choice or preference among several alternatives. Such models are devised to measure the strength of preference so as to predict the probability of choice—for experience in everyday life and in the laboratory has shown that men and animals will vary their choice of one alternative over another in what are externally identical conditions. One simple and useful model is this: if $v(x)$ is defined as the utility or value of x—*i.e.*, the numerical measure of the strength of preference for x—and $p(x,y)$ as the probability of choosing x over y, then the model postulates that $p(x,y,) = v(x)/[v(x) + v(y)]$. The problem for the theory of measurement is to show what properties the observed probabilities $p(x,y)$ must have in order for such a utility function v to exist. If a child, for instance, chooses chocolate ice cream (c) over licorice (l) with probability two-thirds, licorice over a banana (b) with probability two-thirds, and b over c with probability two-thirds, then no such strength-of-preference function v can be found for the set $\{b,c,l\}$.

MEASUREMENT IN SPECIAL DISCIPLINES

Classical physics

Measurement in physics. The development of the central concepts of classical physics—such as the mechanical concepts of mass, force, momentum, and kinetic energy, and electromagnetical concepts such as those of current, voltage, resistance, and impedance—has been closely intertwined with the development of procedures for measuring quantitatively the properties of the phenomena associated with each of these interrelated concepts. The theory and practice of extensive measurement, as discussed above (see *Axiomatic basis of measurement*), and the theory of dimensions and units (see above *Dimensions and units of measurement*) have been developed mainly in the context of classical physics.

One of the most important attitudes toward measurement in classical physics was the firmly rooted belief that, with sufficient effort, errors of measurement could be eliminated in principle and that there would be no limitation to the precision of measurement and, consequently, to the accuracy to which theories might be pushed in testing their validity. Although classical physics continues to be of major importance in applied science, a much more restrained attitude toward measurement now is customary. In the applications of classical electromagnetic theory, for example, it is customary to impose a lower limit defining the smallest admissible length to be considered: a typical restriction is to keep all lengths greater than one-tenth of a millimetre in order to avoid any atomic or subatomic effects that would fall properly within quantum theory.

Modern physics

A central concept of Einstein's theory of relativity—one that has had a profound influence on modern physics—is its revision of the classical concept of simultaneity, which restricts it to immediately adjacent events. There is no well-defined concept of the simultaneous occurrence of two events at astronomical distances; this means that classical procedures for the measurement of simultaneity have been shown to be mistaken in their fundamental assumptions and that the concept of simultaneous distant events is as mistaken as that of absolute position in space.

At another level of analysis are the Lorentz transformations of special relativity, which formulate the relations between two moving space-time frames. In other words, just two stipulations—that the macroscopic measurement of such relativistic distances be conducted along inertial paths (those followed naturally by unaccelerated particles) and that these measurements be invariant for any inertial frame of reference—are sufficient to establish the fundamental principle that any two inertial frames of reference are related by Lorentz transformations.

The theory of measurement has been important in quantum mechanics for two distinct but related conceptual reasons. One deals with the fundamental discovery that it is not possible to measure simultaneously with arbitrary precision the position and momentum of subatomic particles such as electrons or photons. The best known way of expressing this fact is in terms of the Heisenberg uncertainty principle, named after German quantum physicist Werner Heisenberg, which states that the product of the standard error of measurement of position and momentum must exceed a certain positive constant (for a more detailed discussion, see MECHANICS: *Quantum mechanics*).

The strangeness of this discovery, from the standpoint of classical physics, has been one of the most discussed problems of modern physics. Efforts to account for the theoretical uncertainties in measurement or the theoretical limits on the precision of measurement have been a focus of conceptual discussions of the foundations of quantum mechanics. From the standpoint of classical statistical theory, perhaps what is most surprising about the Heisenberg uncertainty principle and related aspects of quantum mechanics is that one is not able to study the covariance of position and velocity measurements. The necessity of looking only at the marginal probability distributions of position and momentum and at the marginal distributions of other conjugate observables in quantum mechanics is the source of the sharpest break between the methodology of quantum mechanics and that of classical statistical theory as used in classical physics and in other domains of science.

The second aspect of measurement that has been fundamental in quantum mechanics is the interaction between the measurement instrument and the measured object. In classical physics it was assumed that the interference of the measuring instrument can in principle be reduced to zero; in quantum mechanics the complicated and detailed analysis of the situation, in which the opposite is a fundamental assumption, has proved to be difficult and elusive.

Measurement in the social sciences and psychology. The behaviour of human beings poses special problems for measurement theory.

Economics and psychology. An important reason that economics is the most advanced of the social sciences—advanced in predictive power, social usefulness, and formal theory—is that many important economic variables may be easily measured. Much of both economic theory and data is cast in terms of price and quantity. The problems associated with these concepts are not so much problems of measurement as they are problems of aggregation to provide representative measures for large economic units, such as an entire industry or a whole economy. Social indexes, such as cost of living or gross national product, involve subtle questions of equivalence, statistical sampling, and the like.

Measurements of price, quantity, and utility

In spite of their success in using the concepts of price and quantity, economists recognize that individuals often base their decisions on more subjective concepts. One approach is to suppose that such individuals have a numerical measure of utility, which they try to maximize. Early attempts to introduce utility into economics employed a representation that, in essence, assumed that the utility of a bundle of commodities is the sum of the utilities of the several commodities; but little attempt was made to axiomatize the representation, and, for various reasons, that hypothesis was discredited. In the 1940s certain scholars suggested that utility could be measured by taking into account the risky character of decision alternatives. Taking a and b as the alternatives and P as a probability, they symbolized by aPb the risky alternative (gamble) in which one receives either a or b, the former with probability P and the latter with probability $1 - P$. They then axiomatized orderings over such structures to show that an order-preserving numerical assignment exists with the expected-utility property: $u(aPb) = u(a)P + u(b)(1 - P)$. Under these conditions, decision making becomes equivalent to maximizing expected utility. This idea has received considerable development since the end of World War II. Its main objective has been the more careful formulation of what an alternative is in terms of various outcomes associated with different chance events and the generalization of the representation theorem to one in which, in addition to constructing the utility function, a (subjective) probability function over events is also constructed from the preference ordering of the alternatives in such a way that subjective expected utility is order preserving.

Although these ideas originated in economics and statistics, psychologists were quick to recognize the problem as one of individual preferences, and they have run numerous experiments to assess its descriptive adequacy. To oversimplify considerably, the major qualitative property known as the (extended) sure-thing principle is not valid. Roughly, this property asserts that if, no matter what possibility is realized, the outcomes of one gamble are at least as satisfactory as the corresponding ones of another, then the former gamble will be preferred or indifferent to the latter. This principle is analogous to independence in additive conjoint measurement and to weak monotonicity in extensive measurement. As a postulate of rational behaviour, however, it has not been seriously questioned; thus, interest in the theory remains.

Another type of psychological research attempts to treat the person as a null-measuring instrument—*i.e.*, as a sensor that records the point at which a difference becomes imperceptible—by having him establish equivalent intensity ratios in different sense modalities, such as sight and sound. These procedures of cross-modality matching yield approximate power relations among the usual physical measures of intensity. Tests to measure ability, intelligence, or achievement constitute one of the most significant and controversial applications of measurement in psychology. The basic theory of tests, which has been extensively developed, does not lie within the framework of fundamental theories of measurement but is mainly statistical in character. The lack of development of fundamental theory is one of the reasons for the many disputes about the nature of IQ and other such measures. (For a discussion of test methodology see PSYCHOLOGICAL TESTS AND MEASUREMENT.)

Sensation

Sociology and demography. The main object of sociology as a science is the study of social groups and institutions and the causes of change within them. Many traditional theories in sociology have been of a speculative sort, not easily subjected to quantitative test. An increasing emphasis on quantitative theories and concepts, however, has inspired a greater concern with problems of measurement. The one part of sociology that from its beginning has emphasized quantitative data and analysis is demography, which is concerned with the quantitative study of human populations and the causes of changes in population. Though demographers are interested mainly in the measurement of factors of population change—including births, deaths, and migrations—they are also concerned with questions of population composition in terms of age, sex, and other characteristics. As in the case of sociology in general, there have been few fundamental problems of measurement of a direct sort in demography; the problems have been primarily of a statistical sort—to infer, for example, future birth trends from intensive field studies of selected samples of a population—although the trend is toward greater accuracy. Perhaps the most successful application of measurement in demography has been in the construction of the mortality tables used by insurance companies and government agencies.

Recent efforts at quantification

BIBLIOGRAPHY. DAVID H. KRANTZ *et al., Foundations of Measurement,* vol. 1 (1971), a comprehensive treatise in which the whole axiomatic view, including many examples of representation and uniqueness theorems, is developed extensively; JOHANN PFANZAGL, *Theory of Measurement,* 2nd rev. ed. (1971), a briefer general treatise; FRED S. ROBERTS, *Measurement Theory with Applications to Decisionmaking, Utility, and the Special Sciences* (1979, reissued 1984), an elementary introduction to axiomatic measurement, notable for its many illustrative examples; LOUIS NARENS, *Abstract Measurement Theory* (1985), a mathematically sophisticated treatment, covering new as well as classical results; R. DUNCAN LUCE and PATRICK SUPPES, "Preference, Utility, and Subjective Probability," in R. DUNCAN LUCE, ROBERT R. BUSH, and EUGENE GALANTER (eds.), *Handbook of Mathematical Psychology,* vol. 3, pp. 249–410 (1965), an extensive review of preference models; PETER C. FISHBURN,

Utility Theory for Decision Making (1970, reprinted 1979), a concise mathematical summary of results in utility theory; BRUNO KISCH, *Scales and Weights: A Historical Outline* (1965), an excellent survey of the more practical aspects, with special emphasis on the history of the instruments used for scales and weights; DAVID HILBERT, *Foundations of Geometry,* 2nd ed. (1971; originally published in German, 10th ed., 1968), one of the first fully satisfactory axiomatizations of geometry from a formal standpoint; JOHN VON NEUMANN and OSKAR MORGENSTERN, *Theory of Games and Economic Behavior,* 3rd ed. (1953, reprinted 1985), the original treatise on the theory of games; S.S. STEVENS, "On the Theory of Scales of Measurement," *Science,* 103(2684):677–680 (1946), his original classification of scales; ISAAC TODHUNTER, *A History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace* (1865, reprinted 1965), including an excellent discussion of the early history of the concept of standard error; JOHN VON NEUMANN, *Mathematical Foundations of Quantum Mechanics* (1955; originally published in German, 1932), a classical treatise containing what remains one of the best discussions of the effects of instrumentation on measurement.

(R.D.L./P.Su.)

Mecca and Medina

Mecca and Medina are the most sacred cities of Islām. Muḥammad, the founder of Islām, was born in Mecca, and it is toward this religious centre that Muslims turn five times daily in prayer. All devout Muslims attempt a pilgrimage to Mecca at least once in their lifetime. Medina is celebrated as the place from which Muḥammad conquered all of Arabia after his flight from Mecca (AD 622), and a pilgrimage is made to his tomb in the city's chief mosque. Because of these sacred features, only Muslims are allowed to enter the cities.

This article is divided into the following sections:

Mecca 798
  Physical and human geography 798
    The landscape 798
      The city site
      Climate
      Plant and animal life
      The city layout
    The people 799
    The economy 799
      Industry
      Transportation
    Administrative and social conditions 799
      Government
      Public utilities
      Education
      Health
  History 799
Medina 799
  Physical and human geography 800
    The landscape 800
      The city site
      The city layout
    The people 800
    The economy 800
      Agriculture
      Industry
      Transportation
  History 800

MECCA

The holiest of Muslim cities, Mecca (in Arabic Makkah) is located in the Ṣirāt Mountains inland from the Red Sea coast of Saudi Arabia. The city has an area of about 10 square miles (26 square kilometres).

A historic city closely tied to tradition, Mecca has benefitted from the Saudi Arabian oil economy of the 20th century, although income from the pilgrimage is still important. The city has undergone vast improvements. The area around the religious shrines has been cleared, the mosque has been enlarged, housing and sanitation have been improved, and transportation facilities have been enhanced. As a result, Mecca can accommodate the continually increasing number of pilgrims.

Physical and human geography

THE LANDSCAPE

The city site. Mecca is situated at an elevation of 909 feet (277 metres) above sea level in the dry beds of the Wādī Ibrāhīm and several of its short tributaries. It is surrounded by the Ṣirāt Mountains, the peaks of which include Jabal Ajyad, which rises to 1,332 feet, and Jabal Abū Qubays, which attains 1,220 feet, to the east and Jabal Qu'ayq'ān, which reaches 1,401 feet, to the west. Jabal Hirā' rises to 2,080 feet on the northeast and contains a cave in which Muḥammad sought isolation and visions before he became a prophet. It was also in this cave that he received the first verse (*ayāt*) of the holy Qur'ān. South of the city, Jabal Thawr (2,490 feet) contains the cave in which the prophet hid from the people of Mecca before he moved to Medina.

Entrance to the city is gained through four gaps in the surrounding mountains. The passes lead from the northeast to Minā, 'Arafāt, and aṭ-Ṭā'if; from the northwest to Medina; from the west to Jidda; and from the south to Yemen (Ṣan'ā'). The gaps have also defined the direction of the contemporary expansion of the city.

Climate. Because of its relatively low-lying location, Mecca is threatened by seasonal flash floods despite the low amount of annual precipitation. There are less than five inches (130 millimetres) of rainfall during the year, mainly in the winter months. Temperatures are high throughout the year and in summer may reach 113° F (45° C).

Plant and animal life. Plant and animal life are scarce and consist of species that can withstand the high degree of aridity and heat. Natural vegetation is sparse, and includes tamarisks and various types of acacia. Wild animals include wild cats, wolves, hyenas, foxes, mongooses, and kangaroo rats (jerboas).

Mecca's religious shrines

The city layout. The city centres upon the al-Ḥaram mosque and the sacred well of Zamzam, located inside the mosque. The compact built-up area around the mosque comprises the old city, which stretches to the north and southwest but is limited on the east and west by the nearby mountains. The main avenues are al-Mudda'ah and Sūq al-Layl to the north of the mosque and as-Sūq as-Saghīr to the south.

Since World War II, Mecca has expanded along the roads through the mountain gaps to the north, northwest, and west. Among the modern residential areas are al-'Azīzīyah and al-Faysalīyah along the road to Minā and aẓ-Ẓāhir, az-Zahra'ā, and Shāri' al-Manṣūr along the roads to Jidda and Medina. Expansion has been accompanied by the construction of new streets in the old city and by the transformation of Mecca into a modern city, with fountains, built since the 1950s, in its four main squares. The

square mosque is magnificent in its size and architecture. It was enlarged from 313,520 to 1,724,032 square feet (29,127 to 160,168 square metres) and can accommodate more than 300,000 worshippers at one time. Houses near the mosque have been razed, and it is now surrounded by open spaces and wide streets, which can be crossed through underground walkways, built to ease traffic.

Housing. Mecca's houses are more compacted in the old city than in the modern residential areas. Traditional buildings of two or three stories are built of local rock. The villas in the modern areas are constructed of concrete. Slum conditions can still be found in various parts of the city; the slum inhabitants are mainly poor pilgrims who remained in Mecca after arriving on pilgrimage.

THE PEOPLE

The population density in Mecca is high. Most of the people are concentrated in the old city, while densities in the modern residential areas are the lowest in the city.

During the month of pilgrimage, the city is swollen with more than 2,000,000 worshippers from other parts of the country and from other Muslim nations.

Residence in Mecca is permitted only to followers of Islām. It is, however, one of the most cosmopolitan cities in the world, containing people from the various Islāmic countries in Asia and Africa. People of the same national origin tend to live together in certain parts of the city.

THE ECONOMY

Arable land and water are scarce, and food must be imported. Vegetables and fruits are brought in daily from the surrounding *wadi*s (watercourses), such as Wādī Fāṭimah, from the aṭ-Ṭāʾif area to the east-southeast, and from the southern agricultural areas, such as Bilād Ghāmid and Bilād Zahrān. Foodstuffs are imported from abroad mainly through the port of Jidda, 45 miles (72 kilometres) to the west on the Red Sea.

Industry. Industry is limited and includes the manufacture of textiles, furniture, and utensils. The overall urban economy is commercial.

Transportation. Transportation and facilities related to the pilgrimage are the main service industries. Mecca has no airport nor water or rail services. It is well served, however, by the Jidda seaport and airport and by intercity truck, bus, and taxi services. A local bus system was established in 1979. Asphalt roads link Mecca with the main cities of Saudi Arabia and neighbouring countries.

The problems of the pilgrimage

Because of the improvement of services, the number of pilgrims has increased. This annual influx brings a good income to the city, but it also results in a population of more than 2,000,000, all of whom need accommodations, food, water, electricity, transportation, and medical services.

In accordance with the prescribed route, all pilgrims have to be transported from Mecca to ʿArafāt, a distance of about 12 miles, during the early morning of the ninth day of the month of Dhū al-Ḥijjah. During the night of the same day, they must travel to Minā, which is almost two miles from Mecca; after three days, all are returned to Mecca. This problem has been met by the construction of a good road network, an adequate supply of vehicles, and traffic control.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The governor of the city is the emir of Makkah *manāṭiq idārīyah* (province), who is responsible for the maintenance of law and order in both the city and the *manāṭiq;* he is appointed by the king and is immediately responsible to the Minister of the Interior. The municipal council is responsible for the functioning of the municipality; the council was formed after World War II, and has 14 members, who are locally elected and are then approved by the Minister of the Interior. Mecca is also the capital of Makkah *manāṭiq idārīyah* which includes the cities of Jidda and aṭ-Ṭāʾif.

Public utilities. Mecca depends upon the surrounding *wadi*s for drinking water. The waters of ʿAyn Zubaydah (Zubaydah Spring), built in the 8th century, flow through tunnels from Wādī Nuʿmān, about 20 miles to the southwest. ʿAyn al-ʿAzīzīyah sends its waters through pipelines from Wādī ash-Shāmīyah, about 60 miles to the northeast. Water and electricity have reached almost all houses. Electricity is generated at an oil-fuelled power station located on the road to Medina.

Education. Free education is provided for both girls and boys from primary to university level. The Umm al-Qura University (founded 1979) is located in Mecca, as well as two university colleges—the Madrasat Ahl al-Hadith and the Saudi Arabian Institute for Higher Education.

Education services

Health. Health services and medical care are free and adequate. Despite careful checking, pilgrims may bring various illnesses, particularly cholera and cerebro-spinal infections, into the city; the health services, however, have been able to keep such problems under control.

History

Ancient Mecca was an oasis on the old caravan trade route that linked the Mediterranean world with South Arabia, East Africa, and South Asia. The town was located about midway between Maʾrib in the south and Petra in the north, and it gradually developed by Roman and Byzantine times into an important trade and religious centre. It was known to Ptolemy as Macoraba.

According to Islāmic tradition, Abraham and Ishmael, his son by Hagar, built the Kaʿbah as the house of God. The central point of pilgrimage in Mecca before the advent of Islām in the 7th century, the cube-shaped stone building has been destroyed and rebuilt several times. During biblical times, the city was ruled by a series of Yemeni tribes. Under the Quraysh, it became a type of city-state, with strong commercial links to the rest of Arabia, Ethiopia, and Europe. Mecca became a place for trade, for pilgrimage, and for poetry festivals.

The city gained its religious importance with the birth of Muḥammad about 570. The prophet was forced to flee from Mecca in 622, but he returned eight years later and took control of the city. He purged Mecca of idols, declared it a centre of Muslim pilgrimage, and dedicated it to Allāh. Since then, the city has remained the major religious centre of Islām. As the ancient caravan route fell into decline, Mecca lost its commercial significance and has since lived mainly on the annual pilgrimages and the gifts of Muslim rulers.

The period of foreign control

Mecca remained virtually independent, although it acknowledged the power of Damascus and later of the ʿAbbāsid caliphate of Baghdad, Iraq. In 1269 it came under the control of the Egyptian Mamlūk sultans. In 1517 dominion over the holy city passed to the Ottoman Turks in Constantinople (now Istanbul). The local city rulers were chosen from the *sharīf*s, or descendants of Muḥammad, who retained a strong hold on the surrounding area. With the fall of Turkish power after World War I, control of Mecca was contested between the *sharīf*s and the Āl Saʿūd (the Wahhābīs) of central Arabia. King Ibn Saʿūd entered the city in 1925, and it became part of the Kingdom of Saudi Arabia, and the capital of Makkah *mānaṭiq idārīyah.*

Under the Āl Saʿūd, a puritanical moral code and law and order were enforced, and the facilities for pilgrims were improved. With the exploitation of Saudi Arabia's oil resources since World War II, Mecca has had a high level of economic development. (A.S.A.)

MEDINA

Medina (Arabic al-Madīnah) was known as Yathrib before Muḥammad's residence there. The name was thereafter changed to Madīnat Rasūl Allāh (City of the Messenger of God) or al-Madīnah al-Muhawwarah (The Most Glorious City). The city is situated in the Hejaz region of western Saudi Arabia, about 250 miles (401 kilometres)

north of Mecca and 85 miles inland from the Red Sea. Medina is second only to Mecca as the holiest place of Muslim pilgrimage.

Physical and human geography

THE LANDSCAPE

The city site. Medina lies 2,050 feet (625 metres) above sea level on a fertile oasis. It is bounded on the east by an extensive lava field, part of which dates from a volcanic eruption in AD 1207. On the other three sides, the city is enclosed by arid hills belonging to the Hejaz mountain range. The highest of these hills is Mt. Uḥud, which rises to more than 2,000 feet above the oasis.

The city layout. Medina is a sacred area, and only Muslims are permitted to enter the city. The airport, however, lies just outside the sacred limits, and a good view of the city can be obtained by foreigners from aircraft landing at the airport.

In Turkish times, there was a small military landing ground at Sultanah, to the south near the garrison's barricks, but the area is now occupied by the king's palace and its extensive satellites. There too are the ruins of the tomb of ʿAmr ibn al-ʿĀṣ, the celebrated conquerer for early Islām of Palestine and Egypt. The tomb of Aaron is located on the highest point of Mt. Uḥud.

Other religious features of the oasis include the mosque of Qubāʾ, the first in Islāmic history, from which the Prophet was vouchsafed a view of Mecca; the Mosque of the Two Qiblahs, commemorating the change of the prayer direction from Jerusalem to Mecca, at ar-Rimāḥ; the tomb of Ḥamza, uncle of the Prophet and of his companions who fell in the Battle of Uḥud (625), in which the Prophet was wounded; and the cave in the flank of Uḥud in which the Prophet took refuge on that occasion. Other mosques commemorate where he donned his armour for that battle; where he rested on the way thither, and where he unfurled his standard for the battle of the ditch (*khandaq*); and the ditch itself, dug around Medina by Muḥammad, in which the rubble of the great fire during the reign (1839–61) of Sultan Abdülmecid I was dumped. All these spots are the object of pious visitation by all Muslims visiting Medina; they are forbidden to non-Muslims. In addition the city is also the site of the Islāmic University, established in 1961.

But the cynosure of all pilgrims is the Prophet's Mosque, which Muḥammad himself helped to build. Additions and improvements were undertaken by a succession of caliphs, and the chamber of the Prophet's wives was merged in the extension during the time of the Umayyad caliph al-Walīd ibn ʿAbd al-Malik. Fire twice damaged the mosque, first in 1256 and again in 1481, and its rebuilding was variously undertaken by devout rulers of several Islāmic countries. Sultan Selim II (1566–74) decorated the interior of the mosque with mosaics overlaid with gold. Sultan Muḥammad built the dome in 1817 and in 1839 painted it green, this being the accepted colour of Islām. Sultan Abdülmecid I initiated a project for the virtual reconstruction of the mosque in 1848 and completed it in 1860. This was the last renovation before the modern expansion planned by King ʿAbdul-ʿAziz in 1948 and executed by King Saud in 1953–55. The mosque now includes a new northern court with its surrounding colonnades, all in the same style as the 19th-century building but of concrete instead of stone from the neighbouring hills. The *Qafas* (cage), to which female worshippers were formerly restricted, has been dismantled, while, apart from minor repairs, the southern (main) part of the mosque has remained intact. It comprises the three ornamental iron structures representing the houses of the Prophet and containing respectively (according to general consensus) the tomb of the Prophet himself under the great green dome, those of the first two caliphs, Abū Bakr and ʿUmar, and that of the Prophet's daughter Fāṭimah, and also a section of the pillared southern colonnade, specially adorned to represent the palm grove (Al Rawdha) in which the first simple mosque was constructed.

The modernization of Medina has not been so rapid as that of Jidda, Riyadh, and other Saudi towns. Building development has involved the complete dismantlement of the old city wall and the merging of that historic area with the now built-up pilgrim camping ground (al-Manakh) and the Anbariya quarter, beyond the Abu-Jidaʿ torrent bed, which was formerly the commercial quarter and in which the Turks established the railway station and terminal yards. The foundations of the old city wall were found to be lower than the surface of accumulated silt and rubble; but unfortunately no attempt has been made to examine the excavations from the archaeological point of view. Nor has any archaeological work been done on the ruined sites of the old Jewish settlements, the largest of which was Yathrib (the Lathrippa or Iathrippa of Ptolemy and Stephanus Byzantius), which gave its name to the whole oasis until Islāmic times. There are also several interesting mounds (*ʿitm*), besides the village of al-Quraidha, which would certainly produce historical data of interest. The Islāmic cemetery of al-Baqiya (Baki al-Gharkad) was shorn of all the domes and ornamentation of the tombs of the saints at the time of the Wahhābī conquest of 1925; but the ruins so created have been tidied up, with simple concrete graves in place of the old monuments, and with a circuit wall.

THE PEOPLE

The residents of Medina are Arabic-speaking Muslims, most of whom belong to the Sunnah branch of Islām. The city is one of the most populous in Saudi Arabia, and it is common for Muslims who make the pilgrimage to settle in the city. Farming and pottery making are important occupations.

THE ECONOMY

Agriculture. To supplement the income derived from accommodating pilgrims, Medina has an economy based on the cultivation of fruits, vegetables, and cereals. The city is especially well known for its date palms, the fruits of which are processed and packaged for export at a plant built in 1953.

Mechanical pumps for irrigation, in use since Turkish times early in the 20th century, have virtually replaced the old draw wells. Drinking water is supplied by an aqueduct from a spring at the southern end of the oasis. In addition to the plentiful supply of subsoil water at no great depth, a number of important *wadis* meet in the vicinity of Medina and bring down torrents of water during the winter rains. Of these the most notable are the Wadi al-ʿAqiq from the western mountains and a wadi coming down from the aṭ-Ṭāʾif area to the south.

Industry. Although Medina was known in early Islāmic times for metal working, jewelry, and armory, these industries were never large-scale and most activity was connected with agricultural technology until the mid-20th century. Principal activities include automobile repair, brick and tile making, carpentry, and metal working.

Transportation. From 1908 to 1916 Medina was connected with Damascus by the Hejaz railway, destroyed during World War I. Reconstruction of this railroad is studied periodically, but has never taken place. Asphalt roads link the city with Jidda, Mecca, and Yanbuʿ (Medina's port on the Red Sea); another road extends north through the Hejaz and connects the city to Jordan. Al-Jiladain airport nearby provides transportation to Saudi Arabian centres and has links to Jordan, Egypt, and Syria.

History

The earliest history of Medina is obscure, though it is known that there were Jewish settlers there in pre-Christian times. But the main influx of Jews would seem to have taken place as the result of their expulsion from Palestine by the Roman emperor Hadrian in about AD 135. It is probable that the Arab tribes of Aws and Khazraj were then in occupation of the oasis; but the Jews were the dominant factor in the population and development of the area by AD 400. In that year Abi-kariba Asʿad, the Sabaean king of Yemen, visited the colony and imbibed the lore and teaching of the Jewish rabbis with the result that he adopted the religion of the Jews and made it the state religion of Yemen on his return, in supersession

The flight from Mecca

of the local paganism. The Yemeni Jews, who in 1949 emigrated to Israel, are mostly descendants of the Arabs then converted. On September 20, 622, the arrival of the Prophet Muḥammad at Medina, in flight from Mecca, introduced a new chapter into the history of the oasis. This flight, known as the *hijrah* (commonly transliterated higira), marks the beginning of the Muslim calendar. Soon thereafter, the Jews, at first treated with indulgence, were driven out of all their settlements in Hejaz. Medina became the administrative capital of the steadily expanding Islāmic state, a position it maintained until 661, when it was superseded in that role by Damascus, the capital of the Umayyad caliphs.

After the caliph's sack of the city in 683 for its factiousness, the native emirs enjoyed a fluctuating measure of independence, interrupted by the aggressions of the *sharīfs* of Mecca or controlled by the intermittent Egyptian protectorate.

The Turks, following their conquest of Egypt, held Medina after 1517 with a firmer hand, but their rule weakened and was almost nominal long before the Wahhābīs, an Islāmic puritanical group, first took the city in 1804. A Turko-Egyptian force retook it in 1812, and the Turks remained in effective control until the revival of the Wahhābī movement under Ibn Saʿūd after 1912. Between 1904 and 1908 the Turks built the Hejaz railroad to Medina from Damascus in an attempt at strengthening the empire and ensuring Ottoman control over the *ḥajj*, the obligatory Muslim pilgrimage to the nearby holy city of Mecca. Turkish rule ceased during World War I, when Ḥusayn ibn ʿAlī, the *sharīf* of Mecca, revolted and put the railroad out of commission, with the assistance of the British officer T.E. Lawrence ("Lawrence of Arabia"). Ḥusayn later came into conflict with Ibn Saʿūd, and in 1925 the city fell to the Saʿūd dynasty. (J.B.Gl./Ed.)

BIBLIOGRAPHY. Literature about Mecca is available mainly in Arabic. For pre-Islāmic and early Islāmic times, see M.A.A. AL-AZRAQI, *Akhbār Makkah,* written in the 9th century (1875, reprinted 1969); and A.I. AL-SHARIF, *Makkah wa-al- Madīnah* (1965). For the Middle Ages, see IBN JUBAYR, *Travels . . . (1183–1185* A.D.), written in the 12th century, trans. by R.J.C. BROADHURST (1952); and IBN BATUTA, *Travels, A.D. 1325–1354,* written in the 14th century, trans. by H.A.R. GIBB (1958). Other English-language accounts include JOHN L. BURCKHARDT, *Travels in Arabia* (1829); JOHN F. KEANE, *Six Months in Meccah* (1881); C. SNOUCK HURGRONJE, *Mekka,* 2 vol. (1888–89); Eng. trans. of vol. 2, *Mekka in the Latter Part of the 19th Century: Daily Life, Customs and Learning* (1931); SIR RICHARD BURTON, *Personal Narrative of a Pilgrimage to El-Medinah and Meccah,* 5th ed., 3 vol. (1906); ARTHUR WAVELL, *A Modern Pilgrim in Mecca and a Siege in Sanaa* (1912); ELDON RUTTER, *The Holy Cities of Arabia* (1928); J.B. PHILBY, *A Pilgrim in Arabia* (1946); DESMOND STEWART, *Mecca* (1980); JOHN SABINI, *Armies in the Sand: The Struggle for Mecca and Medina* (1981); and M.S. MAKKI, *Medina, Saudi Arabia: A Geographic Analysis of the City and Region* (1982). For statistics, see the *Statistical Yearbook* (annual), published by the Central Department of Statistics, Riyadh, Saudi Arabia.

Mechanics: Energy, Forces, and Their Effects

Mechanics is the science concerned with the motion of bodies under the action of forces, including the special case in which a body remains at rest. Of first concern in the problem of motion are the forces that bodies exert on each other. This leads to the study of such topics as gravitation, electricity, and magnetism, according to the nature of the forces involved. Given the forces, one can seek the manner in which bodies move under the action of forces; this is the subject matter of mechanics proper.

Historically, mechanics was the first of the exact sciences to be developed. Its internal beauty as a mathematical discipline and its early remarkable success in accounting in quantitative detail for the motions of the Moon, the Earth, and other planets had enormous influence on philosophical thought and provided impetus for systematic development of science into the 20th century.

The principles of mechanics have been applied to three general realms of phenomena. The motions of such celestial bodies as stars, planets, and satellites can be predicted with great accuracy thousands of years before they occur. (The theory of relativity predicts some deviations from the motion according to classical, or Newtonian, mechanics; however, these are so small as to be observable only with very accurate techniques, except in problems involving all or a large portion of the detectable universe.) As the second realm, ordinary objects on Earth down to microscopic size (moving at speeds much lower than that of light) are properly described by classical mechanics without significant corrections. The engineer who designs bridges or aircraft may use the Newtonian laws of classical mechanics with confidence, even though the forces may be very complicated, and the beautifully precise calculations of celestial mechanics usually cannot be duplicated. The third realm of phenomena comprises the behaviour of matter and electromagnetic radiation on the atomic and subatomic scale. Although there were some limited early successes in describing the behaviour of atoms in terms of classical mechanics, these phenomena are properly treated in quantum mechanics.

This article treats the fundamental parameters and concepts of classical mechanics and of certain key allied fields. It also considers those of quantum mechanics in some detail. For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 126–131.

(Ed.)

The article is divided into the following sections:

Classical mechanics 803
- The origins and foundations of mechanics 803
  - History
  - The fundamental concepts and quantities of mechanics
- Statics 804
  - Force
  - Friction
- Dynamics 807
  - Kinematics
  - Kinetics
  - Principles of momentum and energy
  - Celestial mechanics
    - Historical review
    - Modern method
    - Problem of two bodies and perturbations
    - The stability of the solar system
    - The classical problem of three bodies
    - The problem of *n* bodies
- Other concerns of mechanics 823
  - Inertia forces and the Coriolis force
  - Mechanics of complex systems

Mechanics of nonrigid bodies 827
- Impact, or collision 827
- Mechanics of elastic bodies 827
  - The phenomenon of elasticity
  - Mathematical theory of elasticity
- Deformation and flow 831
  - History
  - Continuum mechanics
  - Constitutive equations
  - Fracture
  - Molecular theories
- Tides and related phenomena 835
  - The causes of tides
  - Tidal analysis
  - Tidal dynamics and prediction
  - Tides in the atmosphere and solid Earth
  - Astronomical tides

Wave motion 843
- Development of wave concepts
- Types of waves
- The theory of waves
- Superposition of waves
- The behaviour of waves at boundaries
- Interference and diffraction
- Water waves

Ballistics 854
- History of ballistics theory
- Principles of ballistics
- Interior ballistics
- Exterior ballistics
- Terminal ballistics

Fluid mechanics 857
- Basic properties of fluids 858
- Hydrostatics 859
- Hydrodynamics 860
  - Bernoulli's law
  - Waves on shallow water
  - Compressible flow in gases
  - Viscosity
  - Navier–Stokes equation
  - Potential flow
  - Potential flow with circulation: vortex lines
  - Waves on deep water
  - Boundary layers and separation
  - Drag
  - Lift
  - Turbulence
  - Convection

Quantum mechanics 872
- Historical basis of quantum theory 872
  - Basic considerations
  - Early developments
    - Planck's radiation law
    - Einstein and the photoelectric effect
    - Bohr's theory of the atom
    - Scattering of X rays
    - De Broglie's wave hypothesis
- Basic concepts and methods 874
  - Schrödinger's wave mechanics
  - Electron spin and antiparticles
  - Identical particles and multielectron atoms
  - Time-dependent Schrödinger equation
  - Tunneling
  - Axiomatic approach
  - Incompatible observables
  - Heisenberg uncertainty principle
  - Quantum electrodynamics
- The interpretation of quantum mechanics 878
  - The electron: wave or particle?
  - Hidden variables
  - Paradox of Einstein, Podolsky, and Rosen
  - Measurement in quantum mechanics
- Applications of quantum mechanics 881
  - Decay of the K^0 meson
  - Cesium clock
  - A quantum voltage standard

Bibliography 883

CLASSICAL MECHANICS

As was noted above, classical mechanics deals with the motions of material bodies of ordinary size moving at speeds that are small compared with the speed of light. It differs not only from quantum mechanics, which describes the behaviour of particulate matter and electromagnetic "rays" of atomic and subatomic size, but also from relativistic mechanics, which deals with exceedingly high-speed phenomena. Relativistic mechanics constitutes an extension of classical mechanics, conforming as it does to the principles of special relativity formulated by the German-American physicist Albert Einstein (1905).

Classical mechanics is a basic part of physics and astronomy and the foundation of most branches of engineering; all rational procedures for the design of machines and structures are based on the laws of classical mechanics. The subject is also of interest to mathematicians because some basic mathematical procedures, such as the differential calculus, were invented for solving mechanical problems.

The bodies dealt with in classical mechanics are assumed to be so stiff that their deformations have no effect on their motions. The effects of deformations are treated in the mechanics of deformable bodies or elasticity; fluid mechanics deals with liquids and gases.

Subdivisions of classical mechanics

Classical mechanics is usually subdivided into statics, which deals with bodies at rest, and dynamics, which deals with bodies in motion. Dynamics is further subdivided into kinematics, which treats of motions without considering the forces involved, and kinetics, which seeks the relation between the forces and the motions.

Classical mechanics is the oldest of the physical sciences. It grew out of the search for an explanation of natural physical phenomena and curiosity regarding the conditions necessary for a state of rest or of motion. Another motivating factor was the urge to create more effective tools and mechanical devices.

The origins and foundations of mechanics

HISTORY

The foundations of statics were laid in the 3rd century BC by the Greek mathematician Archimedes, who derived formulas for the equilibrium of simple levers and centres of gravity. These derivations were restricted to parallel forces, which were only slightly more difficult to deal with than collinear forces. The treatment of nonparallel forces presented difficulties that were not overcome until force was conceived as a vector quantity, having both magnitude and direction, and describable by a directed line segment (arrow). About 2,000 years after the death of Archimedes, the 16th-century Dutch mathematician and inventor Simon Stevin solved the lever problem with nonparallel forces and by observation and intuition alone demonstrated the equilibrium of bodies on a double inclined plane; he showed how to add force vectors by constructing a parallelogram using the force vectors as the sides.

Forces as vectors

In his writings, Stevin makes the statement that what a simple machine gains in force it loses in distance, a clear indication that he understood the principle of virtual work, which was formally enunciated by the Swiss mathematician Johann Bernoulli in the 18th century. The work of Stevin gave a new impetus to the study of statics, which previously had only the theory of the lever as a working principle.

Although the pre-Christian Greeks, particularly Aristotle, tried to explain the behaviour of moving bodies, they were unsuccessful. There were two reasons for their failures: in the first place, they had no satisfactory means for measuring distance or time and, consequently, were unable to check their formulas experimentally; in the second place, they laboured under the false assumption that force was necessary to maintain motion rather than to change its direction or magnitude. It was left for Galileo early in the 17th century to lay the foundation of the science of dynamics by a combination of analytical and experimental procedures that is characteristic of the best scientific work.

One of Galileo's contributions deals with the motion of falling bodies; he disproved the Aristotelian theory that heavy bodies fall faster than light bodies. To verify his theory that the height fallen would be proportional to the square of the time and independent of the weight, he experimented with a smooth ball rolling down an inclined plane. He measured the time of the motion by weighing the water that flowed through a small hole at the bottom of a large tank filled with water.

Galileo developed the equation of motion of a projectile and showed that, neglecting the air resistance, the curved path was the result of two independent motions—namely, a horizontal motion at a constant speed and a vertical motion at a varying speed under the action of the force of gravity. In this case, he utilized the principle of compounding motions in accordance with the parallelogram law for force vectors introduced by Stevin.

Galileo anticipated the law of inertia, formulated by the English physicist Isaac Newton at a later date, when he stated that a body in motion and free from external influences would keep on moving at a constant speed in a straight line. Galileo was the first to recognize that it was acceleration (the rate of change of velocity), not velocity or position, that was determined by the external forces.

The effects of forces

Following Galileo in the 17th century, the next important contributor to the science of dynamics was the Dutch physicist Christiaan Huygens, who developed the equations of motion of the pendulum and invented the pendulum clock. He was the first to obtain the acceleration of gravity by pendulum observations and the first to encounter the term now known as the moment of inertia—*i.e.*, the measure of the inertia of a body in rotary motion.

Newton was born the year Galileo died (1642), and, apart from his important discoveries of the law of universal gravitation, the mathematics of the calculus, and his work in optics, he introduced the concepts of force and mass and formulated the three laws of motion that summarize, clarify, and extend the principles introduced by Galileo. The magnitude of his contributions is reflected in the fact that Newtonian and classical are synonymous terms in the field of mechanics.

Newton's second law equates the force acting on a body to the product of the mass of the body and the acceleration; because the latter is the time rate of change of the velocity, the resulting equations of motion are differential equations and require mathematical techniques for their solution. When possible, Newton used geometric methods, believing that they would be more easily understood by readers. During the 18th century the Swiss mathematician Leonhard Euler wrote a series of textbooks on mechanics in which he used the analytic methods of the calculus to solve the differential equations of motion. Euler also showed that any motion of a rigid body can be separated into a translation followed by a rotation about a chosen point, and he extended Newton's second law, which applies to bodies with mass but negligible dimensions (particles), to rigid bodies.

Also in the 18th century, the French mathematician Jean Le Rond d'Alembert wrote a famous textbook on mechanics and discovered a principle that bears his name. This was an entirely new and original method of solving problems in dynamics by applying the principles of statics. When a body is being accelerated, the forces acting on it are unbalanced, but by Newton's second law they are equal to the product of the mass and the acceleration of the body. By the simple expedient of reversing the mass-acceleration terms, the body is placed in equilibrium, and unknown forces can be determined by the equations that apply to bodies at rest.

Dynamics reduced to statics

One hundred years after the publication of Newton's laws, the French mathematician Joseph-Louis Lagrange developed a formula for deriving the equations of motion that is more universal in its application and for certain types of problems easier to use than Newton's equations. Lagrange's equation is a differential equation that deals

with energy rather than force, and with velocity rather than acceleration; the resulting equations of motion are the same as those derived by Newton's second law. Equations of a similar type were developed by the Irish mathematician William Rowan Hamilton in the 19th century. Both Lagrange's and Hamilton's equations have proved invaluable in the more advanced and theoretical treatments of dynamics.

THE FUNDAMENTAL CONCEPTS AND QUANTITIES OF MECHANICS

The descriptions of the states of rest or of motion of bodies that Newton's laws provide are inherently mathematical, and, consequently, the quantities appearing in the equations require explanation and, in some cases, precise definition.

A description of motion requires a frame of reference and means for locating bodies in the frame. The reference frames used in mechanics are known as coordinate systems, with mutually perpendicular axes (lines) emanating from a point known as the origin; for motion parallel to a fixed plane two axes are required, while for motion in space three are required. Strictly speaking, Newton's laws of motion are valid only in a coordinate system with origin at the centre of the solar system; this is known as a Newtonian, or inertial reference, frame. Although it rotates, the Earth can be considered an inertial reference frame in most cases.

Coordinate systems

The principal units of length used to describe the position of bodies are the metre in the Système Internationale d'Unités (SI) and the foot in the engineering (English) system. With time measured in seconds, the derived units of velocity (directed speed) are metres per second and feet per second. Acceleration, which is the time rate of change of velocity, is expressed as metres per second per second or feet per second per second.

The mass of a body is a measure of its inertia, or the resistance that it offers to having its speed changed; as an illustration of the difference between mass and weight, consider the following example. If a smooth, heavy body were resting on a flat sheet of ice, the mass of the body would influence the time taken to bring it up to a given speed by pushing it and also the time taken to stop it; the greater is the mass, the longer is the time. On the other hand, the weight of the body would influence the effort required to lift it.

Newton's second law states that the acceleration $\boldsymbol{a}$ imparted to a body by a given force is equal to the force $\boldsymbol{F}$ divided by the mass m; *i.e.*, $\boldsymbol{a} = \boldsymbol{F}/m$. This means that, because acceleration has the dimensions distance per second per second, the quotient $\boldsymbol{F}/m$ must have the same dimensions; either $\boldsymbol{F}$ or m can be chosen as the fundamental unit, and the other will be the derived unit.

The SI system is known as an absolute system, and mass in kilograms is the fundamental unit; the derived unit of force is known as a newton, and from the expression for $\boldsymbol{a}$ it is the force that acting on a one-kilogram mass will produce an acceleration of one metre per second per second.

The engineering system is known as a gravitational system, and force in pounds is the fundamental unit; the derived unit of mass is known as a slug, and it is that mass that, when acted on by a one-pound force, will have an acceleration of one foot per second per second.

The relation between mass and weight

The relation between mass and weight can be obtained by applying Newton's equation to a falling body. The force acting on the body is $\boldsymbol{W}$, the weight, and the average value of the acceleration of gravity g on the surface of the Earth is 9.81 metres per second per second in the SI system and 32.2 feet per second per second in the engineering system. Then, in the SI system $\boldsymbol{W}/m = 9.81$, which means that a kilogram weighs 9.81 newtons; in the engineering system $\boldsymbol{W}/m = 32.2$, which means that a slug weighs 32.2 pounds.

The SI and the engineering systems are related in the following way; one pound force equals 4.4 newtons, and one slug mass equals 14.59 kilograms.

The acceleration of gravity varies slightly from place to place on the surface of the Earth, and in interstellar space it may approach zero. Thus, whereas the mass of a body at the speeds considered in classical mechanics is a constant, the weight of a body will vary with its position relative to the centre of the Earth. At the same place, the masses of bodies are proportional to their weights, so that weighing is a convenient means of measuring masses.

Newton's equations of motion deal with the gross motion of a body (*i.e.*, the motion of the body as a whole). There is one point in a body the motion of which is characteristic of the gross motion; it is called the centre of mass. If the body rotates, the rotation can always be referred to the centre of mass, and additional equations must be deduced to deal with the rotation. When the gross motion of a body is more significant than its rotary motion, the body may be treated as a particle. The Earth, for example, can be treated as a particle when describing its motion around the Sun; it cannot be treated as a particle when dealing with terrestrial problems.

Centre of mass

Statics

The forces in the members of a machine or structure must be known before the dimensions required to carry them can be calculated; thus, methods for determining such forces are of extreme importance to all designers of mechanical devices. Assuming that the main loads that a machine or structure is expected to carry are specified, the forces acting in the connected elements of the machine and the reactions from the supports can be determined by various analytic and graphic procedures. When a machine or structure is in a state of equilibrium (at rest), the science of statics provides the means for calculating a limited number of the unknown forces.

The science of statics assumes that the bodies with which it deals are perfectly rigid and asserts that, when at rest, the resultant (sum) of all the forces acting on a body must be zero and that there must be no tendency for the forces to turn the body about any axis. These conditions are independent of one another, and their expression in mathematical form constitutes the equations of equilibrium. When all the forces lie in the same plane there are only three equations so that only three unknown forces can be calculated. If there are more than three unknown forces, it signifies either that there are more restraints than are necessary to keep the body from moving or that there are more members (as in a structure) than are necessary to support the applied loads; the unnecessary restraints and members are said to be redundant. In such cases, the rigid-body assumption must be discarded, and the elastic properties of the members must be considered. Although the procedures for finding the forces in structures with redundant members are strictly static rather than dynamic in nature, it is customary to designate such structures as statically indeterminate. Unless otherwise noted, the following material deals with rigid-body statics only.

FORCE

In statics a force may be defined as any action that tends to maintain or alter the position of a body or to distort it. There are many kinds of forces, such as a push or a pull, gravity (which is manifested as weight), steam or gas pressure in an engine cylinder, atmospheric pressure, wind pressure on a building, resistance to motion in air or water, magnetic and electrostatic attraction, nuclear attraction, and pressure and friction between contacting bodies.

The effect of a force on a body is determined by its point of application, its direction, and its magnitude. Because it has direction as well as magnitude, force is a vector quantity and can be shown graphically as a directed line segment; *i.e.*, a line with a length equal to the magnitude of the force, to some scale (such as one inch equals 100 pounds), inclined at the proper angle, and with an arrowhead at one end to indicate sense (one of two opposite directions). When it is necessary to emphasize the vectorial nature of a quantity it will be indicated in boldface type.

The representation of forces by vectors implies that they are concentrated at a single point or along a single line. Physically this is impossible; even when two metallic spheres are pressed together, the contacting surfaces flatten, and the force is distributed over a small area. On a loaded member of a machine or structure, the force pro-

Internal forces

duces an internal force or stress that is distributed, in some manner, over the cross section of the member. The force of gravity is distributed throughout the volume of a body. Nevertheless, when the equilibrium of a body is the main consideration, it is usually valid as well as convenient to assume that the forces are concentrated at a point. In the case of the gravity force, the total weight of a body is assumed to be concentrated at its centre of gravity.

Basic principles. There are three main principles, or axioms, that facilitate the solution of problems in statics—namely, the parallelogram of forces, action and reaction, and the transmissibility of a force.

The parallelogram of forces, formulated by Stevin in 1586, states that, if two forces whose lines of action intersect are acting on a body, their action is equivalent to the action of a single force whose vector representation is the diagonal of the parallelogram of which the two original force vectors are sides.

In Figure 1A the flexible cords *AB* and *BC* are attached to a ring at point *B* from which a weight is suspended by the cord *BD*. The forces acting on the ring are the tension W in the cord *BD* (acting down and equal to the weight) and the tensions T_1 and T_2 in the cords *AB* and *BC*, respectively. Because the system is in equilibrium, the resultant force acting on the ring—*i.e.*, the sum of W, T_1, and T_2—must be zero. This requirement enables T_1 and T_2 to be determined as functions (*i.e.*, in terms) of W. In Figure 1B the vector W is drawn to any convenient length to represent the tension in the cord *BD;* to complete the triangle of forces, T_1 and T_2 are drawn parallel to *AB* and *BC*, respectively. When vectors are added, they are placed tail to tip, so that as long as the sense of W is known, the arrowheads can be placed on T_1 and T_2 to satisfy this requirement.

In Figure 1A the vector BE, acting upward, is the diagonal of the parallelogram formed by T_1 and the parallel side *GE* and T_2 and the parallel side *FE*. Because triangle *BEF* is similar to the triangle of forces in Figure 1B, the vector BE must be equal in magnitude to W. It is shown acting upward to indicate that it is the resultant of T_1 and T_2 and could serve as their replacement as far as their effects on the ring at *B* are concerned. The ring would then be in equilibrium under the action of two forces—namely, W acting upward and W acting downward.

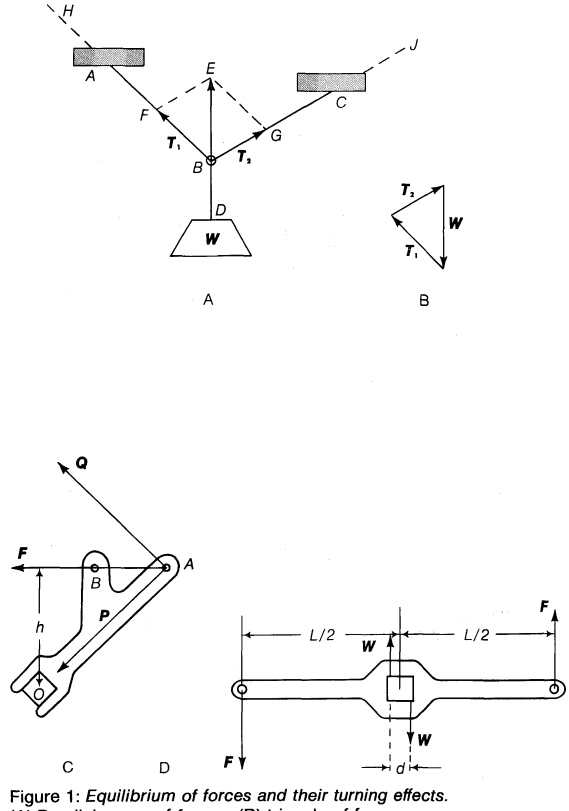


Figure 1: *Equilibrium of forces and their turning effects.* (A) Parallelogram of forces, (B) triangle of forces, (C) moments, (D) couples (see text).

The principle of action and reaction, which is also known as Newton's third law of motion, states that, when one body exerts a force on another body, the second body exerts a force on the first body; the two forces are collinear (act along the same line), equal in magnitude but opposite in sense.

The weight in Figure 1A pulls down on the ring *B* while the ring pulls up on the weight with an equal and opposite force. Contact forces between bodies always occur in pairs; a book resting on a table pushes down on the table with a force that is equal and opposite to the upward force of the table on the book. The principle is also true for some of the forces that act at a distance, such as gravitational attraction; a flying airplane pulls up on the Earth with a force equal and opposite to its weight.

Principle of transmissibility

The principle of transmissibility states that the equilibrium of a body is not altered when the point of application of a force is moved to some other point on its line of action—*i.e.*, the line along which the force acts. In Figure 1A the position of the weight would be unaltered if the cord *BD* were lengthened and attached to the weight at its centre or if *A* and *C* were moved to points such as *H* and *J* or *F* and *G*.

Moment of a force and a couple. The concept of moment, or torque, as it is used in mechanics is the tendency of a force to rotate the body to which it is applied about a point or axis. Of the three forces of equal magnitude applied to the wrench in Figure 1C, the force Q, which acts at right angles to *OA*, is clearly the most effective in tending to turn the square bolt head to which the wrench is applied about the point *O;* force P is completely ineffective. The moment of a force about a point *O* is equal to the product of the magnitude of the force and the length of a line from *O* at right angles to the line of action of the force. Thus the moment of the force F in Figure 1C is equal to Fh, and from the principle of transmissibility F would have the same effect if applied at either point *A* or point *B*.

A more effective wrench than the one in Figure 1C is shown in Figure 1D. If the forces F are equal, they constitute a "couple," and the turning effect, or moment, of the couple is FL. To keep this wrench in equilibrium, another couple, Wd, the reaction from the bolt head, must be acting and must be equal to FL. On the other hand, because the torque is applied to the wrench in Figure 1C by a force, there must be a force equal and opposite to F (or Q) acting on the wrench in addition to a balancing couple.

Application of equilibrium equations. Figure 2A shows one of the two davits used to suspend a lifeboat. The load applied to the davit at *A* by the boat is W, the reaction R_B at *B* is assumed to be horizontal because the support is frictionless, and the support at *C* is a ball-and-socket joint that can carry both horizontal and vertical reactions, such as H_C and V_C. It is required to find the reactions as functions of W, h, and L.

When applying the condition that when a body is in equilibrium the resultant force acting on the body must be zero, it is usually convenient to resolve (separate) the forces into sets of orthogonal (mutually perpendicular) components and assert that the summation of the forces in each orthogonal direction must be zero. Because the direction of a vector involves both inclination and sense, a vector or the component of a vector along or parallel to any fixed line, such as the axis of a coordinate system, has a fixed inclination and can be described by a number, either positive or negative, to indicate its magnitude and whether it points in one or the other of two opposite directions (sense). Consequently, such vectors, although they are shown as directed lines, can be considered as scalars (quantities that require only a number for their specification). In the following examples, force components along chosen directions will be taken as scalars and be added, subtracted, multiplied, and divided like ordinary numbers.

Scalars

For the davit, using the x and y coordinate axes as convenient directions, Σ for summation, F as a general term for forces, and F_x and F_y for the forces acting along x and y, the two force equations of equilibrium would

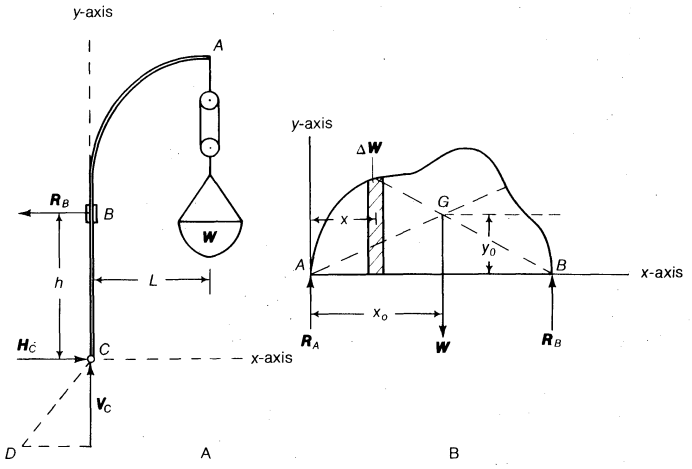


Figure 2: *Supporting reactions on stationary bodies.*
(A) Lifeboat davit, (B) flat plate (see text).

be $\Sigma F_x = 0$ and $\Sigma F_y = 0$. An additional moment equation, expressing the requirement that the davit must not rotate about any chosen point in the plane of the forces, must be written; choosing point C and using M to designate moment, the moment equation would be $\Sigma M_C = 0$. In Figure 2A the reaction R_B and the two components of the reaction at C, namely, H_C and V_C, are shown acting in assumed directions; the correctness of these assumptions will be verified later. Because positive directions for x and y are to the right and up, respectively, R_B is negative, while H_C and V_C are positive. Clockwise movements will be considered positive.

From $\Sigma F_y = 0 = V_C - W$ it follows that $V_C = W$; from $\Sigma M_C = 0 = WL - R_B h = 0$ it follows that $R_B = WL/h$; and from $\Sigma F_x = 0 = H_C - R_B$ it follows that $H_C = R_B = WL/h$. Because all three unknown forces came out positive, it proves that the assumed directions were correct; if H_C, for example, had been assumed in the negative x direction, it would have come out negative in $\Sigma F_x = 0$. The total force at C, which is the resultant of V_C and H_C, is a vector DC, pointing up and to the right.

Centre of gravity. Figure 2B shows a plate of weight W resting on horizontal supports at A and B. If the reactions R_A and R_B were required, the summation of vertical forces F_V would yield only $\Sigma F_V = R_A + R_B - W = 0$; to find R_A and R_B it would be necessary to write a moment equation about either A or B that would include the moment of the weight of the plate. The moment of the plate about A could be obtained by adding up the moments of a large number of small slices of the plate, symbolized by ΔW, whose moment arms x are taken at the centre lines of the slices. It would be more convenient if a point such as G, the centre of gravity, could be found at which the entire weight of the plate could be concentrated. The distance x_0 from A to the line of action of W can be found by equating the moment Wx_0 to the sum of the moments obtained by the slicing method; this gives $Wx_0 = \Sigma(\Delta W)x$, in which the summation is taken over the whole body. Solving this equation for x_0 yields $x_0 = \Sigma(\Delta W)x/W$. The vertical coordinate y_0 of G can be obtained by slicing the plate parallel to AB. The coordinate of G perpendicular to the plane of the paper would be half the thickness of the plate from either surface of the plate.

If the upper surface of the plate were a curve that could be described by a mathematical equation, the coordinates of G could be obtained by the integral calculus. The summation sign Σ becomes an integral sign $\int$, the slices become infinitesimally thin, and the calculated value of x_0 is exact. For a plate of any shape the centre of gravity can be obtained by suspension. If the plate in Figure 2B were suspended from A, the line AG would be vertical, while if suspended from B the line BG would be vertical; point G is at the intersection of these two lines.

Centre of gravity by suspension

For bodies with three planes of symmetry, such as a sphere or a circular cylinder, the centre of gravity is at the geometric centre of the body. For bodies with less symmetrical shapes, such as cones, segments of a circle, and triangles, formulas for the centre of gravity are to be found in handbooks.

FRICTION

Friction is of considerable importance in everyday life as well as in mechanics. Without friction, walking would be impossible, no ladder would stand against a wall, an automobile could not be driven or stopped.

Between two bodies in direct contact there is always some friction, which tends to hinder or even prevent the sliding of one body on the other. Consider the steel block resting on the steel table in Figure 3A. The weight W, acting down, and the reaction N from the table, acting up, must be equal and opposite. If a small horizontal force P is applied to the block and the block does not move, there must be a friction force F, equal and opposite to P at the surface of contact between the block and the table. If the force P is increased, it will eventually cause the block to slide on the table.

Experiments show that, when slipping impends, the value of F is always a fixed ratio of the normal force N and that this ratio depends on the materials and the roughness of the contacting surfaces. Thus, when slipping impends, $F = fN$, in which f is the coefficient of static friction. For

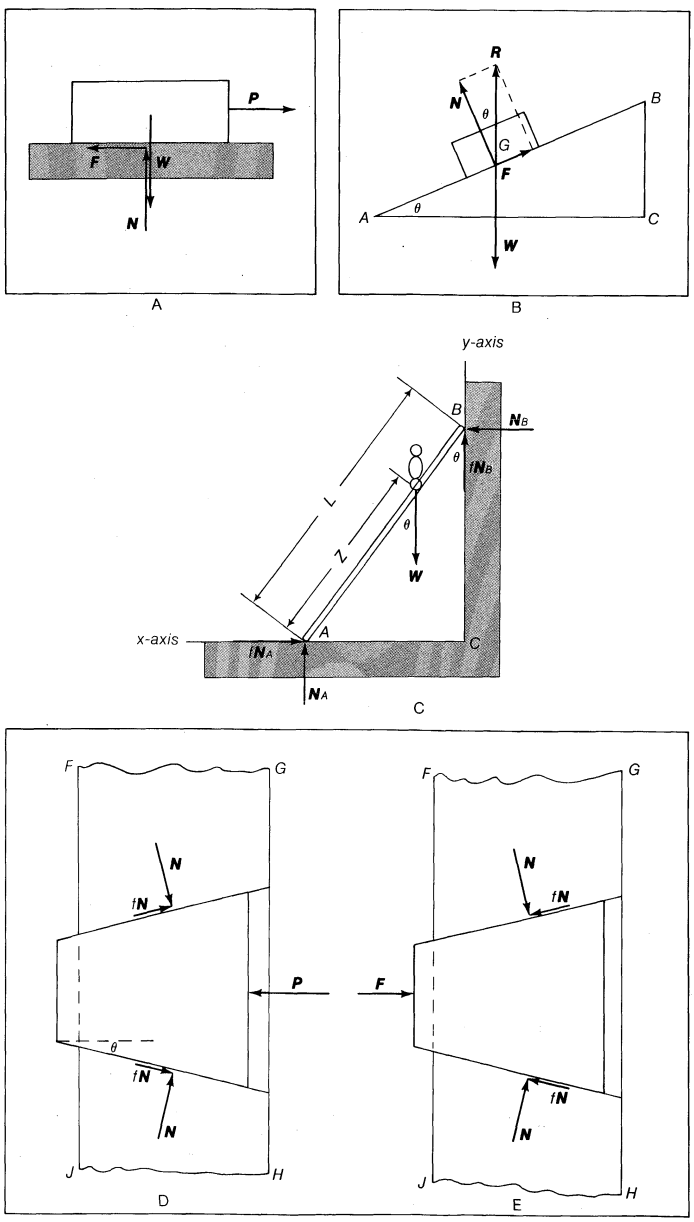


Figure 3: *Friction forces and their effects.*
(A) Sliding block, (B) the angle of repose, (C) a sliding ladder, (D,E) a wedge (see text).

dry steel on steel the value of f is given as 0.15 in some handbooks, but it varies greatly with surface roughness and whether or not surface films such as oxides are present.

The value of f can be determined experimentally by means of an adjustable inclined plane, as in Figure 3B. The angle θ is slowly increased until the block starts to slide down the plane. There are two forces acting on the block: the weight $\boldsymbol{W}$ and the reaction $\boldsymbol{R}$ from the plane, both of which must pass through the centre of gravity G. Resolving $\boldsymbol{R}$ into the components $\boldsymbol{N}$ and $\boldsymbol{F}$, and noting that triangle ABC and the triangle formed by $\boldsymbol{R}$, $\boldsymbol{N}$, and $\boldsymbol{F}$ are similar, $BC/AC = \boldsymbol{F}/\boldsymbol{N} = \tan\theta$, which by definition is equal to f; the angle θ is known as the angle of repose.

Between a rubber tire and a dry concrete road f could reach a value of 1.00 so that $\boldsymbol{F} = \boldsymbol{N}$, and, in Figure 3B, BC would equal AC and θ would equal 45°. This means that an automobile with locked brakes could stand on an incline of 45°, but for greater angles it would slide downhill.

Dry, or Coulomb, friction

The first complete investigation of dry friction was made by the French physicist Charles-Augustin Coulomb in 1781. He discovered that the total amount of friction that can be developed between two contacting bodies is proportional to the normal pressure and independent of the area of contact. If a brick is pulled along a table, the frictional force is the same whether the brick is lying flat, on its side, or on its end. Coulomb also found that after slipping has started, the frictional force is less than the friction when slipping impends. Later experiments indicate that at extremely high pressures abrasion may alter the surface conditions and cause f to reach unpredictable values. Dry friction is usually called Coulomb friction, to distinguish it from viscous friction, which occurs in oil films.

In the operation of machines, friction may be either a useful agent or the cause of lost power. The operation of brakes and clutches, the grip of belts on pulleys, and the tractive effort of wheels on rails are all applications in which the maximum amount of friction is desired. In bearings and sliding connectors the friction, which causes heat and loss of power, is reduced to a minimum by the addition of lubricants whose main function is to minimize metal-to-metal contact.

The ladder. A ladder AB of length L in Figure 3C rests on a horizontal floor at A and a vertical wall at B. It is required to find the maximum distance Z that a man of weight $\boldsymbol{W}$ can climb up the ladder before slipping occurs at A; the weight of the ladder, being small compared with the man's weight, is to be neglected.

In addition to the man's weight, there is a force $\boldsymbol{N}_A$ normal (at right angles) to the floor at A and a force $\boldsymbol{N}_B$ normal to the wall at B. There must also be a friction force $f\boldsymbol{N}_A$ at A to balance $\boldsymbol{N}_B$; the friction force $f\boldsymbol{N}_B$ at B is not necessary to maintain equilibrium, but it appears because the wall is assumed to be rough. To simplify calculation, it is assumed that the coefficient of friction f is the same between the floor and the ladder as between the wall and the ladder.

To determine $\boldsymbol{N}_A$ and $\boldsymbol{N}_B$ in terms of $\boldsymbol{W}$, it is necessary only to equate the sum of the forces in the vertical (y) direction and in the horizontal (x) direction to zero. To obtain a relation between Z, L, θ, f, and the forces, however, it will be necessary to write a moment equation about a convenient point such as A. Then:

$$\Sigma\boldsymbol{F}_y = \boldsymbol{N}_A + f\boldsymbol{N}_B - \boldsymbol{W} = 0,$$
$$\Sigma\boldsymbol{F}_x = \boldsymbol{N}_B - f\boldsymbol{N}_A = 0.$$

The summation of moments is given by:

$$\Sigma\boldsymbol{M}_A = \boldsymbol{W}Z\sin\theta - \boldsymbol{N}_B L\cos\theta - f\boldsymbol{N}_B L\sin\theta = 0$$

(sin, cos, tan, and cot are ratios of the various sides of a right-angled triangle). Solving these equations yields the following:

$$\frac{Z}{L} = \frac{f\cot\theta + f^2}{1 + f^2}.$$

As long as $f\cot\theta$ is equal to or greater than one, Z/L will be equal to or greater than one, and the man can climb to the top of the ladder without fear of the ladder slipping. When slipping impends, $f\cot\theta = 1$ or $f = 1/\cot\theta = \tan\theta$. From Figure 3C, $\tan\theta = AC/BC$ so that as long as AC/BC is less than f the ladder will not slip. For timber on stone f is approximately 0.40, which corresponds to an angle θ of 21°48′; if θ is greater than this, the ladder will slip. On rough ground f could be 1.0, which corresponds to a safe angle θ of 45°. It is worth noting that θ is the same as the angle of repose, in Figure 3B, at which friction just holds the block from sliding down.

The wedge. In Figure 3D a tapered plug is being driven into a matching hole in the member $FGHJ$ by a force $\boldsymbol{P}$; the progress of the plug is opposed by the two normal (at right angles) reactions $\boldsymbol{N}$ and the two friction forces $f\boldsymbol{N}$. The summation of forces in the horizontal direction yields the equation $\boldsymbol{P} - 2\boldsymbol{N}\sin\theta - 2f\boldsymbol{N}\cos\theta = 0$, which, solved for $\boldsymbol{N}/\boldsymbol{P}$, gives

$$\frac{\boldsymbol{N}}{\boldsymbol{P}} = \frac{1}{2(\sin\theta + f\cos\theta)}.$$

The ratio $\boldsymbol{N}/\boldsymbol{P}$ is the mechanical advantage, a measure of the force-amplifying property of a wedge. The magnitude of the ratio increases as the denominator of the previous equation decreases. The magnitude of the denominator depends on both the angle θ and the coefficient of friction f; the smaller is the angle and the smaller is the friction coefficient, the smaller is the denominator. As θ approaches zero, $\sin\theta$ approaches zero and $\cos\theta$ approaches unity. This means that no matter how small θ becomes, the term $f\cos\theta$ can only be reduced by lowering f.

If the plug in Figure 3D is to be pulled out, the force $\boldsymbol{P}$ (now labeled $\boldsymbol{F}$) and the frictional forces will be reversed, as shown in Figure 3E. For this case it can be shown that $\boldsymbol{F} = 2\boldsymbol{N}(f\cos\theta - \sin\theta)$.

When $f\cos\theta = \sin\theta$, or $f = \tan\theta$, the magnitude of $\boldsymbol{F}$ is zero. This means that the plug will fall out of its own accord when $\boldsymbol{P}$ is removed. For all angles for which $\tan\theta$ is less than f, the taper is a "sticking" one, and $\boldsymbol{F}$ in the previous equation will be greater than zero. Sticking tapers are used for holding such cutting tools as drills and reamers in machine-tool spindles.

Dynamics

KINEMATICS

The position vector and motion. Motion is change of position, and to describe the motion of a body A, for example, relative to another body B, it is necessary to establish a reference frame on body B and means for describing the position of any point in body A in this frame. A convenient frame is a system of rectangular coordinate axes—*i.e.*, graduated lines emanating from a point called the origin and at right angles to one another. For three-dimensional (space) motion three axes are necessary, for two-dimensional (plane) motion two are required, while for one-dimensional (line) motion only one axis is necessary.

In Figure 4A the position of point P on its path in the OXY coordinate system can be described either by the coordinates x and y or the length of the vector $\boldsymbol{S}$ and the angle θ. Like all vectors, $\boldsymbol{S}$ has both magnitude, as given by its length, and direction, which is a combination of inclination and sense (one of two opposite directions). When a vector, such as $\boldsymbol{S}$, is used to describe the position of a point, it is known as a position vector.

Magnitude and direction

As the point moves on its path, the head of vector $\boldsymbol{S}$ will move with it, and the magnitude and direction of $\boldsymbol{S}$ will change. These changes are the only two that a vector can experience. The coordinates x and y will change in magnitude as the point moves and serve as orthogonal components of the vector $\boldsymbol{S}$.

Velocity. The speed of the point P is the time rate at which the point moves along the path; the velocity of the point is a vector quantity and as such has both magnitude, which is equal to the speed, and direction, which, as shown in Figure 4A by the vector $\boldsymbol{v}$, is tangent to the path and pointing in the direction in which the point is moving.

By definition, the velocity of point P can be expressed either as the time rate of change of the vector $\boldsymbol{S}$ or the vector sum of the changes in x and y. Because $\boldsymbol{S}$ changes in both magnitude and direction, the velocity $\boldsymbol{v}$ must be the vector sum of these changes. As will be shown below, changes in the magnitude of a vector are always collinear

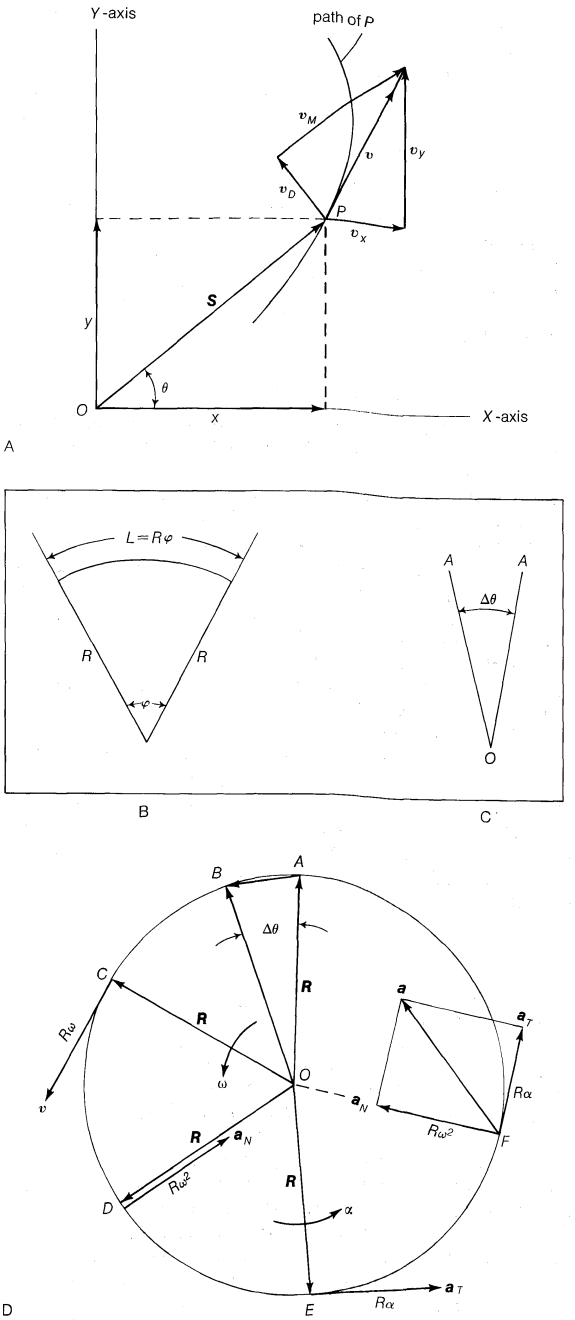


Figure 4: *Motion of a point on a plane (flat) curve.*
(A) Position and velocity vectors, (B) radian measure, (C) rotary motion, (D) motion on a circle.

with the vector, whereas changes in direction are at right angles to the vector.

The vector $\boldsymbol{v}_M$ in Figure 4A represents the rate at which the magnitude of $\boldsymbol{S}$ is changing and $\boldsymbol{v}_D$ the rate at which its direction is changing. Both of these changes are functions of $\boldsymbol{S}$; *i.e.*, they are expressed mathematically in terms of $\boldsymbol{S}$.

The coordinates x and y are quantities that have magnitude and sense only; they are positive when they lie above and to the right of the origin O and negative when they lie below or to the left of O. Because they do not rotate and change in magnitude only, x and y can be considered as scalars or directed lines or one-dimensional position vectors. The velocities resulting from their rates of change are usually called the scalar components of $\boldsymbol{v}$. Velocities obtained from scalar quantities, or the magnitude of a velocity vector, are usually indicated by $\boldsymbol{v}$. Then $\boldsymbol{v}_x$ and $\boldsymbol{v}_y$ are the velocities resulting from the changes in x and y and their resultant is the vector $\boldsymbol{v}$.

Acceleration. Being a vector, the velocity $\boldsymbol{v}$ may change in both magnitude and direction as the point moves, and the vector sum of these changes is defined as the acceleration $\boldsymbol{a}$ of P. The acceleration can also be defined in terms of the sum of the changes in $\boldsymbol{v}_D$ and $\boldsymbol{v}_M$, or the changes in $\boldsymbol{v}_x$ and $\boldsymbol{v}_y$. Because $\boldsymbol{v}_D$ and $\boldsymbol{v}_M$ move with $\boldsymbol{S}$, they may change in both magnitude and direction, and the acceleration vector will have four components if they are used, and they will all be functions of $\boldsymbol{S}$; if $\boldsymbol{v}_x$ and $\boldsymbol{v}_y$ are used, there will be only two components, parallel to the X- and Y-axes and functions of x and y. These components of $\boldsymbol{a}$ are written as $\boldsymbol{a}_x$ and $\boldsymbol{a}_y$. If a point moves on a curved path, the direction of its velocity must change, and its acceleration can never be zero, even if the speed is constant.

Components of acceleration

Because the coordinate system in Figure 4A can be located in any position relative to the path, there are an infinite number of coordinates and position vectors that can be used to describe the position of the point; the choice must be made by the analyst. For a straight path, a directed line lying in the path and changing only in magnitude is obviously the best choice. For a point moving on a circle, a position vector with its tail at the centre of the circle and changing only in direction is most suitable. The use of position vectors that change in both magnitude and direction is to be avoided, if possible.

Rotation about a fixed axis. When a body rotates about a fixed axis, the angular position of any line on the body passing through and perpendicular to the axis changes with respect to time, and the time rate at which the line sweeps out angles is known as the angular velocity of the body.

Because it permits a simple mathematical relationship between the length L of the arc of a circle (*i.e.*, segment of the circumference) and the angle subtended (enclosed) at the axis by the arc, measurement of angles in classical mechanics is expressed in radians, as explained below. In Figure 4B if φ is measured in radians, the arc length L is given by $L = R\varphi$. This means that, when the length of the arc equals the radius, or $L = R$, φ is one radian, or the angle at the centre of a circle subtended by an arc equal to the radius is one radian. In one revolution the arc length is equal to the circumference so that $L = 2\pi R$, and thus $2\pi R = R\varphi$ or $\varphi = 2\pi$ radians; this means that in one revolution there are 2π (*i.e.*, approximately 6.28) radians.

Measurement of angles

If the line OA in Figure 4C is on a body rotating about a fixed axis O perpendicular to the page and sweeps out an angle $\Delta\theta$ in a time interval Δt, the average angular velocity $\boldsymbol{\omega}$ during the interval is $\boldsymbol{\omega} = \Delta\theta/\Delta t$ radians per second. As the time interval Δt is reduced, the average velocity approaches the instantaneous velocity. The limiting value of $\Delta\theta/\Delta t$ as Δt approaches zero is defined in the differential calculus as the instantaneous angular velocity, and the equation is written $\boldsymbol{\omega} = d\theta/dt$. If θ is expressed as a function of t, exact values of $\boldsymbol{\omega}$ can be obtained by procedures developed in the calculus. If the angular velocity changes by an amount $\Delta\boldsymbol{\omega}$ in time Δt, the average angular acceleration $\boldsymbol{\alpha}$ is given by $\boldsymbol{\alpha} = \Delta\boldsymbol{\omega}/\Delta t$, and the instantaneous acceleration, in calculus notation, is $\boldsymbol{\alpha} = d\boldsymbol{\omega}/dt$. Because $\boldsymbol{\alpha}$ is the rate of change of $\boldsymbol{\omega}$, which is itself the rate of change of θ, $\boldsymbol{\alpha}$ can also be described as the second derivative of θ and written as $\boldsymbol{\alpha} = d^2\theta/dt^2$.

Motion on a circular path. To describe the motion of a point on a circular path, a position vector coinciding with the radius and changing only in direction is the best choice.

In Figure 4D, when the point moves from position A to position B on the circle of radius $OA = R$, the vector $\boldsymbol{R}$ of magnitude R changes by the amount $\boldsymbol{AB}$; *i.e.*, $\boldsymbol{AB}$ is the vector that must be added vectorially to $\boldsymbol{R}$ to obtain the vector $\boldsymbol{OB}$. If $\Delta\theta$ is measured in radians, the length of the arc AB equals $R\Delta\theta$, and it is approximately equal to the length of the vector $\boldsymbol{AB}$, which is the chord of the arc, and also the magnitude of the change in the vector $\boldsymbol{R}$ because of its rotation through the angle $\Delta\theta$. If $\Delta\theta$ takes place in time Δt, then R $(\Delta\theta/\Delta t)$ is approximately the magnitude of the average time rate at which the vector $\boldsymbol{R}$ is changing. As $\Delta\theta$ decreases, the lengths of the chord and the arc AB approach one another and the angle between $\boldsymbol{R}$ and $\boldsymbol{AB}$ approaches 90°. In the limit, as Δt approaches zero, $\Delta\theta/\Delta t$ (using the notation of the calculus) becomes $d\theta/dt$, which is recognizable as $\boldsymbol{\omega}$, the angular velocity of $\boldsymbol{R}$ in radians per second. Then $R\omega$ is the magnitude of the vector that

represents the instantaneous time rate of change of $\boldsymbol{R}$; this is the magnitude of the velocity, or the speed, of the moving point and is written as $v = R\omega$. The velocity vector $\boldsymbol{v}$ is inclined at 90° to $\boldsymbol{R}$, with a sense compatible with the direction of $\boldsymbol{\omega}$. At point C in Figure 4D, a position vector $\boldsymbol{R}$ rotating in a counterclockwise direction and the corresponding velocity vector $\boldsymbol{v}$ of magnitude $R\omega$, are shown in the proper relationship. As shown below, this relationship can be described in a more convenient way.

Vectors that rotate

When a vector rotates, its time rate of change resulting from the rotation is a vector whose magnitude is equal to the product of the magnitude of the rotating vector and its angular velocity; the direction of the vector is found by turning the rotating vector through 90° in the direction of the angular velocity. This procedure applies to all rotating vectors regardless of the kinematic property that they represent. As shown below, it can be used to find the acceleration resulting from the rotation of $\boldsymbol{v}$.

The vector $\boldsymbol{v}$ in Figure 4D must always change in direction, and because it is always at right angles to $\boldsymbol{R}$ its angular velocity is $\boldsymbol{\omega}$. The vector that represents the rate at which its direction changes will have a magnitude equal to $R\omega \times \omega$ or $R\omega^2$ and its direction will be obtained by turning $\boldsymbol{v}$ through 90° in the direction of $\boldsymbol{\omega}$. As shown at D in Figure 4D, this vector, labeled $\boldsymbol{a}_N$ and with magnitude $a_N = R\omega^2$, is collinear with $\boldsymbol{R}$ but with the opposite (inward) sense. If $\boldsymbol{R}$ rotated clockwise, $\boldsymbol{v}$ would point up at point C and when rotated clockwise it would point toward the centre of the circle, just like $\boldsymbol{a}_N$ at point D.

The vector $\boldsymbol{a}_N$ always points toward the centre when a point moves on a circular path. The subscript N signifies that the vector lies along the normal (at right angles) to the tangent to the path; $\boldsymbol{a}_N$ is also known as the centripetal acceleration.

The relation $v = R\omega$ when solved for ω yields $\omega = v/R$, and, if this is substituted for ω in $R\omega^2$, the magnitude of the normal acceleration can be written as v^2/R. When a point moves on a curved but noncircular path, the normal component of acceleration has a magnitude v^2/ρ, in which ρ is the radius of curvature of the path—*i.e.*, the radius of the circle that matches the curve at the point.

Any changes in the speed of the point in its orbit about O must come from changes in $\boldsymbol{\omega}$, which are measured by the angular acceleration $\boldsymbol{\alpha}$ of $\boldsymbol{R}$. Consequently, the magnitude of the other component of the acceleration of the point is $a_T = R\alpha$, because it represents a change in the magnitude of $\boldsymbol{v}$, it must be collinear with $\boldsymbol{v}$; *i.e.*, tangent to the path and with a sense compatible with the direction of $\boldsymbol{\alpha}$, as shown by $\boldsymbol{a}_T$ at point E in Figure 4D. The resultant (total) acceleration $\boldsymbol{a}$ of the point, as shown at F, is the vector sum of $\boldsymbol{a}_N$ and $\boldsymbol{a}_T$.

Simple harmonic motion. The device shown in Figure 5A is a Scotch yoke; it consists of a T-shaped yoke with a slot in which the block can slide and an integral piston P that slides in a cylinder in the frame, which is shown crosshatched. The crank of radius R is pivoted to the frame at O and to the block at B. Rotation of the crank about O with an angular velocity $\boldsymbol{\omega}$ causes the block to slide up and down in the slot and the yoke to move back and forth (left and right in Figure 5) with a translatory motion while guided by the piston.

The directed line $\boldsymbol{S}$, with its tail at O and head at the centre of the slot, is convenient for describing the position of the yoke on its straight path. From Figure 5A it is clear that the length of $\boldsymbol{S}$ measured from O depends on the length of $\boldsymbol{R}$ and the angle θ and that $\boldsymbol{S}$ will point to the left when θ exceeds 180°; to distinguish between the directions of $\boldsymbol{S}$, it will be assumed that $\boldsymbol{S}$ is positive (+) when it points to the right and negative (−) when it points to the left.

From trigonometry the relation between $\boldsymbol{S}$, R, and θ is given by $\boldsymbol{S} = R \sin \theta$. When θ is zero and 180°, $\sin \theta$ and $\boldsymbol{S}$ are zero; when $\theta = 90°$, $\sin \theta = 1$ and $\boldsymbol{S} = R$; when θ is between 180° and 360°, $\sin \theta$ is negative. Consequently, the magnitude and direction of $\boldsymbol{S}$ are taken care of automatically by $\sin \theta$.

Because the vector $\boldsymbol{S}$ does not rotate, its time rate of change, which is the velocity of the yoke, can be obtained from the calculus by taking the derivative of $\boldsymbol{S} = R \sin \theta$ with respect to time. The velocity can also be obtained by utilizing the relationships shown in Figure 4D for a point moving on a circular path. Point B in Figure 5A moves on a circular path of radius R, and its velocity, of magnitude $R\omega$, is perpendicular to R and pointing down. The velocity of the yoke is the horizontal component of $R\omega$, namely, $\boldsymbol{v} = R\omega \cos \theta$, pointing to the right in the position shown and collinear with $\boldsymbol{S}$.

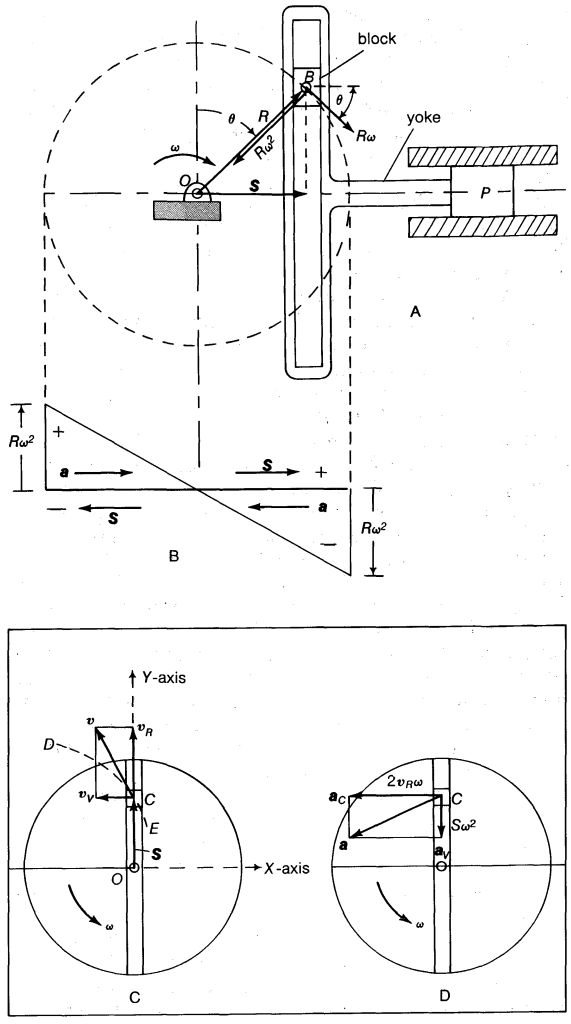


Figure 5: *Harmonic and relative motion.*
(A) Scotch yoke, (B) harmonic motion, (C) relative velocity, (D) relative acceleration.

If $\boldsymbol{\omega}$ is constant, the acceleration of B has a magnitude $R\omega^2$ and is collinear with R and pointing toward O. The acceleration of the yoke is the horizontal component of $R\omega^2$—namely, $\boldsymbol{a} = -R\omega^2 \sin \theta$, pointing to the left in the position shown and collinear with $\boldsymbol{S}$.

Because $\boldsymbol{S} = R \sin \theta$, the horizontal component of the acceleration can be expressed in terms of $\boldsymbol{S}$, as $\boldsymbol{a} = -\boldsymbol{S}\omega^2$. This important relationship is illustrated in Figure 5B. The motion of the yoke, which is known as a simple harmonic motion, can be described in words as follows: if the displacement ($\boldsymbol{S}$) of a reciprocating point is measured from the centre of its path and the acceleration is proportional to the displacement and in the opposite sense, then the motion is called simple harmonic. In Figure 5B, when the displacement $\boldsymbol{S}$ is to the right (+), the acceleration is to the left (−), and, when $\boldsymbol{S}$ is to the left (−), the acceleration is to the right (+).

Simple harmonic motion

Simple harmonic motion is an important form of periodic motion; it is the motion of a point on the prong of a tuning fork, the bob of a simple pendulum swinging on a small arc, and a weight hung from a spring and vibrating vertically.

The period, or time required for one complete to-and-fro motion of a body executing simple harmonic motion, is

equal to the time required for one revolution of the crank in Figure 5A. Because there are 2π radians in one revolution, the speed of the crank in revolutions per second is $\omega/2\pi$ and the time for one revolution $T = 2\pi/\omega$ seconds. Then, from $a = -S\omega^2$, $\omega = \sqrt{-(a/S)}$ and $T = 2\pi/\sqrt{-(a/S)}$ or $T = 2\pi\sqrt{-(a/S)}$. Because S and a are always of opposite sign, the quantity under the radical is always numerically positive. (A.Co./Ed.)

Vibrations. The word vibration comes from the Latin *vibrare,* meaning "to shake or brandish." This involves the idea of a motion that is not always in the same direction but in which the moving object goes back and forth. A vibration is a motion that can, and usually does, result when an elastic material medium is displaced from its equilibrium or rest state. When the displacing force is removed, the elastic medium tends to snap back, to regain its rest state, and then to overshoot it. This tendency leads to the production of vibrations.

A simple example is provided by a vertical spring, one end of which is attached to a ceiling, the other end supporting an object. When undisturbed, the object hangs at rest. If the object is pulled down to stretch the spring and it is then released, the elasticity of the spring tries to restore it to its original length and hence pulls up the object with an accelerating force. When the spring reaches its original length, the object has some momentum that it cannot lose instantaneously. It therefore continues to move upward, compressing the spring until it comes to rest with the spring now shorter than it was originally. The spring's elasticity then reasserts itself and, in striving again to regain its original length, it expands and pushes the weight down, this time accelerating it downward. When the original length is once more reached, the object possesses momentum in the downward direction. In coming to rest it stretches the spring. The process in which the elastic spring tends to regain its original state and continually overshoots the mark will go on repeatedly (at any rate for some time) with the result that the object vibrates up and down.

Examples of vibrating bodies

Examples of such vibrations are legion. Some, like the vibrations of violin and piano strings, give pleasure to the ear by the sounds they produce. Others, like those of the pneumatic drill, are useful technologically, though painful to human ears in their vicinity. Still others, like the torsional or twisting vibrations of the drive shaft of an engine, can lead to dangerous fatigue and if intense enough may even produce fracture of the metal. Similar dangers may accompany the vibrations of the hull of an aircraft.

Other examples of the vibrations of solids commonly experienced are those of the loudspeaker cone in the radio or record player, the vibrations of the crystal in a microphone, and the vibrations of drumheads and bells. Nature provides the vibrations of insect wings and bodies, producing the interesting, if at times monotonous, sounds of summer.

Because air is an elastic fluid, it is able to vibrate and transmit harmonic sounds. Humans have been able to develop a host of musical instruments that owe their properties to the vibration of air. These include pipe organs, flutes, woodwinds, and brass instruments like horns and trumpets. Finally, there is the vibration of air in the ear, responding to the vibration of the vocal cords leading to speech, without which human society could scarcely have developed.

Technologically, it is practically impossible to build a machine that transfers energy from one place to another without having its operation accompanied by vibrations of some kind. Such vibrations involve the wastage of precious energy, metal fatigue, and, usually, noise. Vibration control is, therefore, an important element in modern technology.

The general nature of vibrations. Disturbances leading to vibration can be of various kinds. The simplest is a change in the physical dimensions of an object, such as an increase in the length of a string or rod or the displacement of a taut string or rod at right angles to its length. It may be a twist in a cylindrical rod. In the case of a liquid or a gas, the disturbance is commonly a change in density (*i.e.,* a squeeze or a rarefaction).

Vibration as periodic motion

Considered quantitatively, a vibration is a periodic motion—one that repeats itself over and over again in more or less equal time intervals. The simplest and most regular type of periodic motion is that called simple harmonic. This is exhibited graphically in Figure 6, in which the vibrational disturbance, represented by ξ, is plotted as a function of the time t. If it is assumed that the vibratory motion begins at $t = 0$ when $\xi = A$, in the passage of time ξ takes on a succession of values, including zero at O_1, negative values from O_1 through O_2 to O_3 (corresponding to a decrease in dimensions if the initial ξ is an increase in dimensions), zero again at O_3, and back to its original value of A at O_4. In the plot shown here, the curve is called a cosine curve, discussed in texts on trigonometry and analytic geometry. As t continues to increase, ξ continues to vary periodically in the fashion shown in Figure 6.

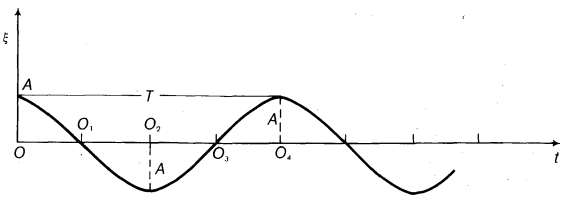


Figure 6: Simple harmonic vibration with displacement ξ plotted against time, t; the period is T (see text).

This simple harmonic vibration is characterized by certain important properties. The maximum displacement from equilibrium (that is, here the state of zero displacement) is called the amplitude of the vibration. In the drawing it is A. The time taken for the disturbance to come back to its original value is called one period of the motion, denoted by T in Figure 6. Although ξ has the same absolute magnitude A at O_2 when $t = T/2$, it has not there attained its original value, since at O_2, $\xi = -A$; *i.e.,* the displacement is in the opposite direction from that at $t = 0$. In the period T the motion is said to have undergone one complete oscillation or cycle.

Frequency

The number of cycles the motion goes through per unit of time (usually taken as one second) is called the frequency, a quantity of greater practical importance than the period. Frequency is commonly denoted by f and is the reciprocal of the period T, that is, $f = 1/T$.

The unit of frequency is the cycle per second, now usually called the hertz (Hz), named for the German physicist Heinrich Hertz. In the vibration of the object on the end of a spring, as described earlier, the frequency may be low enough (*e.g.,* from 0.5 to 2 Hz) so that the eye can readily follow the vibratory motion. With a much stiffer spring (*i.e.,* one harder to stretch) or a smaller mass, or both, the frequency may become so high, however, that all the eye can detect is a blur. If one strikes an A tuning fork (440 Hz), the eye is unable to detect the vibration at all. In this case, one can feel the vibration by touch, but more significantly one can hear the sound produced when the vibrations of the fork are communicated to the air. Solid vibrators in air at frequencies from about 20 Hz to about 20,000 Hz are audible to human ears; in this range the actual motion of such vibrations can be made visible only by means of a device called a stroboscope, which produces short flashes of light in synchronization with the vibration frequency. A quartz crystal may be made to vibrate at such high frequencies (*e.g.,* 10^5 Hz to 10^6 Hz or even higher) that one can neither hear nor see it vibrate. Such vibrations have many practical applications, such as the production of ultrasonic radiation and, by coupling to appropriate mechanical devices, the production of high-speed dental drills.

The frequency of vibrations depends essentially on two factors: the mass of the vibrator and its elasticity. In general, for a given mass, the greater the elasticity, the higher the frequency will be; whereas for a given elasticity, the larger the mass, the lower the frequency. A simple illustration is the object vibrating at the end of the stretched spring as discussed earlier. If the mass of the spring is negligible compared with the mass of the vibrating object, the mass of the system is effectively the mass of the object.

The elasticity of the vibrator is provided by the stiffness of the spring. The stiffness constant is defined as the ratio of the static force required to stretch the spring from its equilibrium length to the increase in length. The greater the stiffness constant, the more elastic the spring is said to be. The vibration frequency can be calculated from the equation in which stiffness is denoted by k, the effective mass of the vibrator by m, and frequency is denoted by f:

$$f=\frac{1}{2\pi}\sqrt{\frac{k}{m}}. \qquad (1)$$

It is apparent that this agrees with the verbal statement made above. For example, if $m = 1{,}000$ grams and $k = 10^5$ dynes/cm, the resulting frequency is about 1.5 Hz.

Most vibrating systems encountered in nature or fabricated by humans are more complicated than the mass–spring combination, but the general statement about frequency still holds, though the assignment of mass and elasticity factor may involve considerable mathematical analysis. Thus, in the case of a simple stretched membrane like a circular drumhead of area S, mass per unit area ρ, and surface tension F (the force per unit length with which the membrane on one side of any straight line in the surface pulls on the part of the membrane on the other side), one can still express the lowest frequency of vibration of the membrane when the centre is depressed and then released by a formula of the form of (1) if the effective stiffness of the membrane (its elasticity factor) is $2\pi F$, and the effective mass of the membrane is one-third the actual mass, namely $\rho S/3$. The lowest frequency then becomes

$$f=\frac{1}{2\pi}\sqrt{\frac{2\pi F}{\rho S/3}}=\frac{1}{2\pi}\sqrt{\frac{6\pi F}{\rho S}}=\frac{1}{2\pi}\sqrt{\frac{6F}{\rho a^2}}, \qquad (2)$$

in which a is the radius of the membrane. Again, the equation shows that the more massive the membrane for given F, the lower the frequency will be, and vice versa. This discussion, of course, does not tell the whole story, for when such a membrane is excited by mechanical or other means, it will in general vibrate simultaneously with a number of frequencies in addition to the lowest one (the fundamental) given by (2). These additional vibrations occur at the so-called natural frequencies, or normal modes, of vibration. Similar considerations apply to metal rods, plates, and shells. The harder it is to deform the vibrators—*i.e.*, the more elastic they are for the same mass—the higher is the vibration frequency. For any given elasticity factor, the greater the effective mass, the lower the frequency.

Damped and forced vibrations. The vibrations described in the previous section are known as natural or free vibrations. A body set to vibrating by a single act of excitation and then left entirely to itself never continues to vibrate indefinitely. The vibration gradually dies and is said to be damped. Even though the frequency does not change appreciably with time, the amplitude decreases more or less as shown in Figure 7, which is a more realistic representation of the behaviour of most actual vibrating systems than that in Figure 6. A standard measure of the damping, or decay, of free vibrations is called the logarithmic decrement, the natural logarithm of the ratio of successive amplitudes (*i.e.*, $\log_e A_1/A_2$ in Figure 7). Damping is usually the result of friction between the vibrating body and its supports and the surrounding medium, as well as of the internal friction associated with the relative motion of the different parts of the body itself. If the resultant frictional force is proportional to the velocity of the vibration (a common and reasonably well substantiated assumption)

Damping

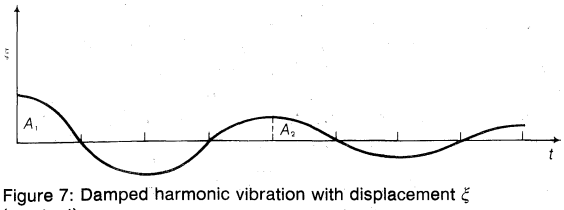


Figure 7: Damped harmonic vibration with displacement ξ (see text).

and the coefficient of this proportionality is R, then the logarithmic decrement (the measure of the decay) is directly proportional to R and inversely proportional to the effective mass of the vibrator. This means that the larger the friction for a given velocity of the vibrator, the faster the decay. The decay will be slow if the mass of the vibrator is large, and slow if the frequency is high. It is also inversely proportional to the natural frequency.

To keep a vibration going at constant amplitude—*i.e.*, to counteract the damping due to friction—energy must be supplied from some external source. Such an energy supply need not be vibratory in character. It can feed energy to the vibrator at a rate constant in time, and the vibration will be sustained at its natural frequency. Such a sustained vibration is called a self-excited vibration because its frequency depends on its own constitution and not on the external energy supply. An example is the vibration of a violin string produced and maintained by a steady bowing.

Self-excited and forced vibration

On the other hand, a vibrator can be forced to vibrate at a frequency other than its natural frequency by the application of a periodically varying force of deliberately chosen frequency. The resulting vibration is then called a forced vibration. An important practical example is the forced vibration of a piezoelectric quartz crystal placed in an alternating electric field.

The amplitude of such a forced vibration depends on the amplitude F of the force, the natural frequency of the vibrator, and the damping coefficient R as above defined. For convenience, it is customary to deal with the amplitude of the velocity of the vibrator rather than with that of the displacement. The velocity amplitude A is greater when the amplitude F is large, when the damping coefficient R and the effective mass of the vibrator are small. If f_0 is the natural frequency and f is the frequency of the oscillatory impressed force (and the damping is not too large), the amplitude of the velocity becomes approximately

$$A=\frac{F}{\sqrt{R^2+4\pi^2m^2(f-f_0)^2}}, \qquad (3)$$

in which m is the effective mass of the vibrator. If the velocity amplitude A is plotted as a function of f, the resulting curve is a bell-shaped curve symmetrical about the value $f=f_0$ and reaches its maximum value $A=A_0$ for $f=f_0$. When $A=A_0$, the vibrator is said to resonate with respect to the force, and f_0 is then called the resonance frequency of the vibrator. In terms of the equivalent stiffness k and the effective mass m, the resonance frequency is that given by equation (1). The resonance frequency and the natural frequency are practically identical as long as the damping is small—*i.e.*, not sufficiently great to reduce the amplitude of a free vibration to zero in a few periods of the motion.

Resonance

The curve for equation (3) is called the resonance curve of the vibrator. Examination of equation (3) discloses that, at resonance when $f=f_0$, $A_0=F/R$ and, if the damping coefficient R (or alternatively the logarithmic decrement to which it is proportional) is small, the maximum amplitude (*i.e.*, the maximum response of the vibrator, or resonance peak, as it is usually called) is very large. At the same time, under these conditions the value of A falls off from A_0 very quickly as f departs on either side from f_0. This is said to correspond to sharp tuning of the vibrator. On the other hand, for relatively large damping the resonance peak is low and the curve itself is flattened out; *i.e.*, the tuning is broad or coarse.

Another way of describing the resonance peak, which is much used by engineers in dealing with vibrations, is the so-called quality factor, or Q, which is defined as π divided by the logarithmic decrement. Small damping with a sharp resonance peak corresponds to a high Q, and conversely. A quartz crystal has very small damping and a very high Q (on the order of 10,000), whereas a vibrator made of lead, which manifests much internal damping, has a very low Q.

Energy and power in vibrations. The behaviour of vibrators can be effectively studied in terms of the concept of energy and the related concept of power (the rate of energy

expenditure). At any instant any vibrator will in general possess both kinetic energy of motion and potential energy of displacement or deformation, the total mechanical energy being the sum of the two. A vibrator cannot create energy. Its energy must be supplied by a source. Examples of such sources are the relative motion or slippage of rocks in unstable equilibrium in the Earth's mantle resulting in seismic vibrations and earthquake waves, the bodily motions of a human individual playing a musical instrument, electrical oscillation exciting a piezoelectric crystal like quartz, a flame producing vibrations in the air in a tube through the emission of heat, an explosion releasing chemical energy, expanding air as in a pneumatic tool, and an oscillating magnetic field producing vibrations in a magnetostrictive rod (whose dimensions change according to its state of magnetization).

Sources of excitation

In order to keep a vibrator going despite the decay associated with damping, energy must be supplied continually from the source. If the vibrator is to remain in a steady state with constant vibration amplitude, the average power delivered by the source must be equal to the average power dissipated by the vibrator. The latter is proportional to the square of the amplitude of the force produced by the source; it also depends in a more complicated way on the damping coefficient and on the ratio of the frequency at which the source is forcing the vibrator to vibrate to the resonance frequency. As might be expected, the average power input and its equal, the average power output, take on their maximum values when the forcing frequency is equal to the resonance frequency. To put the matter in another way, a vibrator absorbs energy from a source at maximum average rate at the resonance frequency. Thus, the vibrator prefers to vibrate at resonance, and at this frequency it loads the energy source to maximum extent.

As has already been suggested, if the damping is very small, a vibrator will draw correspondingly large average power from the source, especially at resonance. If the damping becomes effectively zero momentarily, or even negative, as can happen under certain peculiar circumstances, the power withdrawal may become so great as to lead to a runaway vibration that may destroy the vibrator. An example of this was the destruction of the Tacoma Narrows Bridge in the U.S. state of Washington in 1940, when wind supplied the power causing the bridge to vibrate at one of its torsional resonance frequencies without sufficient damping.

Energy transfer

An example of a useful vibrator, considered from the standpoint of energy transfer, is the so-called Helmholtz resonator. This is usually in the form of a hollow sphere, commonly, but not necessarily, of metal, with a circular opening whose diameter is $1/5$ to $1/10$ of the diameter of the sphere. It can be thought of as analogous to the object at the end of a spring. The mass of the air in the opening of the resonator is the analogue of the mass of the object, while the elasticity of the air in the volume of the resonator (acting as a kind of cushion) provides the stiffness of the system. If somehow the air in the mouth of the resonator can be excited to move, the resonator will act as an aerial or sound vibrator. This will happen if a source of sound, like a tuning fork vibrating at the resonance frequency of the resonator, is brought close to its opening. As a result of the excited vibration, the resonator will emit sound at its natural frequency, and the sound will be more intense than that produced by the tuning fork alone. The reason is that the presence of the resonator forces the fork to greater power output than would otherwise be normal; the resonator "loads" the fork as a source of power. Since the fork's supply of energy is limited (unless it is electrically driven at constant amplitude), its vibrations die down more quickly than they would in the absence of the resonator. Such resonators are useful devices in acoustics.

Vibrations as wave sources. One of the most important properties of vibrations in material media, both in nature and in technological applications, is their role in initiating and maintaining wave propagation. The part played by the vibrations of strings, rods, membranes, plates, and air columns in the production of sound, which is a form of compressional wave propagation in air, has already been noted.

The general connection between the behaviour of vibrations and waves is readily grasped, using the example of a solid vibrator surrounded by a fluid medium. The motion of the vibrator is communicated to the portion of the medium immediately adjacent to it, but if the medium is elastic, it strives to return to its original equilibrium state and has to continue its motion to do this. This motion is communicated to the next adjoining portion of the ambient medium, and so the disturbance is transmitted as a wave through the medium (see below *Wave motion*). Since the vibrations of the source are periodic, the wave will be periodic; at any place in the medium surrounding the source, the disturbance will vary periodically in time, with frequency equal to that of the vibrating source (as long as the source as a whole is at rest in the medium). From the standpoint of energy, the wave may be thought of as a means for the transfer of the energy output of the vibrating source to distant points in the medium. The source is said to radiate energy. Indeed, it is quite proper to think of the propagation of a periodic wave as the production of a vibration at every point of the medium.

Types of vibrational waves

There are two main types of vibrational waves: transverse and longitudinal. If one end of a finite, horizontally stretched, flexible string, with the other end fastened to a rigid support, is vibrated with an up-and-down motion at right angles to the length of the string, periodic waves will travel up and down the string. These waves are called transverse, since the displacement constituting the disturbance is at right angles to the equilibrium position of the string. Another important example of transverse vibrational waves is provided by the torsional, or twist, waves in a solid cylinder or cylindrical pipe. Light and radio waves are also transverse, though here the physical significance of the disturbance is quite different from the displacement of the material elastic medium.

If one vibrates one end of a solid rod in such a way that the displacement is always directed along the rod, a longitudinal vibrational wave results in which the wave displacement at every point in the rod is parallel to the axis of the rod. This may also be referred to as a compressional wave. Sound waves in air and in other fluids like water are longitudinal waves.

If the surface of water is vibrated at one point, vibrational surface waves (ripples) are propagated in all directions from the point of disturbance, and these surface waves have both transverse and longitudinal components. Vibrational surface waves can also be produced in solids.

Vibrational waves possess all the properties of waves in general. Thus, they can be reflected from barriers; they can be refracted (bent in the direction of propagation) when traveling across the boundary separating two media in which the wave velocities differ; they can be diffracted or bent in passing around the edges of obstacles; they can interfere with each other, as in the phenomenon of stationary waves in a finite stretched string; and they can be absorbed by the media in which they travel.

Behaviour of materials in vibration. It has been suggested that the vibration behaviour of an extended, though finite, solid can be complicated, since in general different parts of the material will not vibrate together, or in phase. The simplest case of vibration is that in which the solid vibrates in one of its so-called normal or harmonic modes. A flexible string, fastened at its two ends, is a good illustration. When vibrating in its fundamental or first harmonic mode, the string vibrates as a whole in a single segment, passing through all bow-shaped curves from the extreme arc on one side of the equilibrium position to the extreme arc on the other side and back again in one complete cycle, or period, of the motion. In this case, the length of the string is one-half the wavelength of the transverse wave running back and forth along the string.

A somewhat more interesting example is provided by Figure 8, which indicates schematically the string vibrating in its third harmonic mode. In this case, the vibration takes place in three segments. At its maximum displacement on one side, the string takes the form AGCDEFB, with the first and third segment up and the middle segment down. After a quarter of a period, the displacement in each segment has become zero and then proceeds to

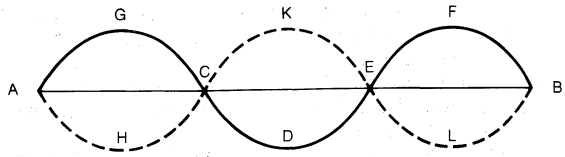


Figure 8: Vibration of a finite stretched string in its third harmonic normal mode (see text).

reverse itself, so that at the end of half a period the string assumes the pattern AHCKELB, with the first and third segments down and the middle one up. After another half period the original shape is reassumed, and so on. This mode of vibration is called the third harmonic because the frequency is three times that for the fundamental, or first harmonic, mode. Vibrations of this type are also referred to as stationary waves. In such modal vibrations there are certain points in the string that do not move at all—*e.g.*, in Figure 8, the points A, C, E, and B. (Of course A and B do not move, since the string is assumed to be fastened there.) Such points are called nodal points, or nodes. Points at which the displacement can reach its maximum value—*e.g.*, G, K, F—are called loops, or antinodes.

Stationary waves

As might be expected, the normal mode patterns in an extended elastic solid, like a flat or curved plate, are more complicated than those just described for a string. In place of nodal points there appear nodal lines or curves along which no motion takes place. These can readily be exhibited in the case of a square metal plate (excited to vibrate in one of its normal modes) by sprinkling fine sand on the plate. The sand tends to gather along the nodal lines. Figure 9 indicates the kind of pattern that may be expected in a square plate of uniform thickness.

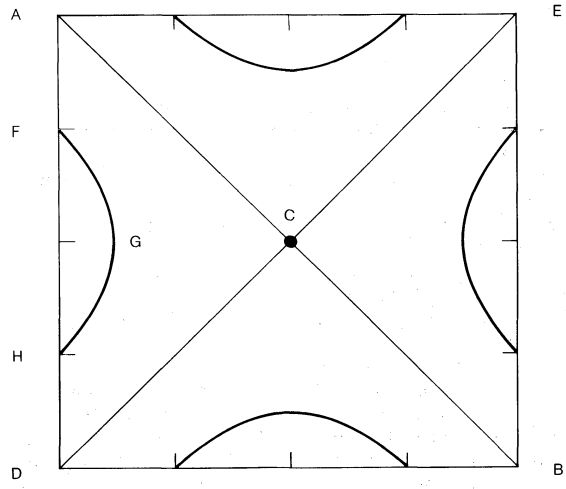


Figure 9: Typical normal mode pattern in the vibration of a square solid plate of uniform thickness rigidly fastened at the centre (C). The lines ACB and DCE are straight nodal lines, whereas the curves, such as FGH, are curved nodal lines.

The study of such normal mode patterns is of some technological importance because it gives an indication of where in the plate, for a given frequency of vibration, the magnitude of the displacement may be expected to be greatest and where least.

Whenever a solid is deformed over and over again as by a vibration, fatigue, or weakness, ensues, and this can become serious enough to lead to ultimate fracture if the intensity of the vibration is sufficiently great. Vibration damping, consequently, is an important technological problem. A vibration damper is a device that reduces the amplitude of the vibration disturbance, as explained earlier. It may also serve to reduce the communication of vibration energy from the vibrator to its surroundings and thus in a sense isolate it. A good example of a vibration isolator is provided by the stiff springs with which a piece of vibrating machinery is mounted to the floor of a building. The springs, due to their stiffness and frictional resistance, reduce the transmission of vibrating energy to the floor.

Friction can be used directly as a mechanism for damping undesirable vibrations. Thus the flat springs used in the suspension of vehicle bodies are usually laminated. Separation of the metal layers by a viscous material also assists the damping process. Such sandwich-type plates have found wide use in vibration prevention.

Hydraulic dampers are also in common use. These operate through the resistance that the viscosity of a liquid offers to a vibrator. Still another type of damper is the dynamic variety: If an appropriate mass is attached by a spring to an object whose vibrations are to be damped, and if the frequency of the mass and spring combination as a vibrating system is tuned to resonance with the frequency of the vibration whose damping is desired, the auxiliary mass will transfer the vibration from the object. This scheme can also be employed to damp torsional vibrations in the drive shafts of turbines.

The detection and utilization of vibrations. It is often desirable to detect vibrations that are not obvious to the ear or the eye, such as those produced at relatively great distances from an earthquake centre or those generated in various parts of a high-speed machine. In the former, a seismograph is employed. This instrument is basically a physical pendulum with a very massive bob, or hanging weight, so suspended as to respond to any vibration in the earth in its vicinity. The response is amplified by electromagnetic induction in a circuit surrounding the vibrating bob, with an optical lever attached to the associated galvanometer. Such instruments are used to detect vibrations resulting from seismic waves and those due to explosions set off deliberately, as in oil prospecting.

Vibration detectors

Vibration detectors for machinery vibration are usually referred to as pickups. These are, in general, mechanical–electrical conversion devices, to which the term transducer has been attached. One type employs the change in resistance of an electrical conductor when its size and shape are altered; another is the electromagnetic variety already referred to in connection with its use in the seismograph. A very important type of transducer used in the detection and utilization of sound vibrations at all frequencies is the piezoelectric crystal, mentioned earlier. All these devices may also be regarded as vibration amplifiers, useful particularly in the detection of weak vibrations. (R.B.L.)

Relative motion. Although all motion is relative, the term is usually reserved for motion on a moving path or, in other words, relative to a moving coordinate system. In Figure 5C, a large horizontal table rotating at a constant (nonvarying) speed ω about a vertical axis O carries a radial track on which a car C is moving away from the axis at a constant speed. The track is a moving path, and the OXY axes attached to the table form a moving coordinate system. Because the car is being carried around counterclockwise by the table as it moves outward along the radial track, the path that it traces in space will be a spiral curve, like ED. The motion of the car on its space path—*i.e.*, its absolute motion—can be described as the sum of its motion relative to the table and the motion of the table, and for this purpose the position vector $\boldsymbol{S}$ is convenient. Relative to the table, $\boldsymbol{S}$ changes only in magnitude at a time rate appropriately designated as the relative velocity $\boldsymbol{v}_R$. Because it rotates with the table, $\boldsymbol{S}$ also changes in direction at a rate that, as shown previously, is given by a vector, of magnitude $S\omega$, perpendicular to $\boldsymbol{S}$ and pointing to the left because of the direction of $\boldsymbol{\omega}$. All points on the table have circular space paths, and their speeds are equal to the products of their distances from the centre of the table and the angular velocity $\boldsymbol{\omega}$. Consequently, $S\omega$ can be interpreted as the speed of the point on the table that coincides with C at the instant considered; the corresponding vector is called the vehicle velocity $\boldsymbol{v}_V$, the table being the vehicle. The absolute velocity $\boldsymbol{v}$ of the car is equal to the vector sum of $\boldsymbol{v}_R$ and $\boldsymbol{v}_V$, and, as shown in Figure 5C, it is tangent to the space path.

Relative and absolute velocity

Because the car is moving at a constant speed there is no relative acceleration, and, consequently, $\boldsymbol{v}_R$ remains constant in magnitude; it does change in direction, however, because it rotates with the table at an angular velocity $\boldsymbol{\omega}$. The vector $\boldsymbol{v}_V$ also changes in direction but, in addition, it changes in magnitude because $\boldsymbol{S}$ is increasing in magnitude. Thus the acceleration of the car is the sum of the

three changes in the components of the velocity—namely, the changes in the direction and magnitude of $\boldsymbol{v}_V$ and the change in the direction of $\boldsymbol{v}_R$.

The directional change in $\boldsymbol{v}_V$ is obtained by turning $\boldsymbol{v}_V$ through 90° in the direction of $\boldsymbol{\omega}$; because $\boldsymbol{v}_V = \boldsymbol{S\omega}$ its magnitude is $S\omega \times \omega = S\omega^2$ and it is collinear with $\boldsymbol{S}$ and in the opposite sense; because of its similarity to $\boldsymbol{a}_N$ in Figure 4D, it can be interpreted as the acceleration of the point in the table coinciding with C at the instant considered; this is the vehicle acceleration $\boldsymbol{a}_N$.

Because $\boldsymbol{v}_V = \boldsymbol{S\omega}$ and $\boldsymbol{\omega}$ is constant, the change in the magnitude of $\boldsymbol{v}_V$ is the result of the change in the magnitude of $\boldsymbol{S}$, which has already been shown to be v_R. Consequently, the magnitude of this change in $\boldsymbol{v}_V$ is $v_R\omega$ and it is collinear with and has the same sense as $\boldsymbol{v}_V$ because $\boldsymbol{S}$ is increasing.

The directional change in $\boldsymbol{v}_R$ is obtained by turning $\boldsymbol{v}_R$ through 90° in the direction of $\boldsymbol{\omega}$; its magnitude is $v_R\omega$ and it will coincide with $\boldsymbol{v}_V$ in Figure 5C. It can be shown that the two vectors of magnitude $v_R\omega$, one resulting from the change in magnitude of $\boldsymbol{v}_V$ and the other from the change in direction of $\boldsymbol{v}_R$, will always coincide. The sum of these two components is the Coriolis component of acceleration $\boldsymbol{a}_C$, and its magnitude is $2v_R\omega$. In Figure 5D the absolute acceleration of the car $\boldsymbol{a}$ is shown as the vector sum of $\boldsymbol{a}_V$ and $\boldsymbol{a}_C$. If the car were moving at a variable speed, there would be a relative acceleration vector, collinear with $\boldsymbol{S}$ and pointing in the same direction as $\boldsymbol{S}$ if the relative velocity were increasing. If the table were being accelerated, there would be another component of vehicle acceleration, perpendicular to $\boldsymbol{S}$ and of magnitude $S\alpha$, in which $\boldsymbol{\alpha}$ is the angular acceleration of the table.

Coriolis component of acceleration

Summarizing, the absolute velocity is always equal to the sum of the relative and vehicle velocities. The absolute acceleration is the sum of the relative and vehicle accelerations, and, if the vehicle rotates, there is an added term, the Coriolis component (discovered in 1835 by the French mathematician Gustave-Gaspard Coriolis).

In the case considered, it is clearly convenient to describe the absolute motion in terms of the relative motion and the vehicle motion. To find $\boldsymbol{v}$ and $\boldsymbol{a}$ using absolute quantities, it would have been necessary to know the precise nature of the absolute (space) path. There are many applications, particularly in the analysis of machinery, for which either absolute or relative quantities can be used. In some cases absolute quantities provide the simpler solution.

KINETICS

Newton's laws of motion. Newton's three laws are the basic postulates governing the relations between the forces acting on a body and the motion of the body. Although they were formulated for the first time in usable form by Newton, they had been discovered experimentally by Galileo years earlier. The laws cover only the overall motion of a body—*i.e.*, the motion of its centre of mass—and not any rigid-body motion such as a rotation. As explained above, this is equivalent to assuming that the body is a particle.

The law of inertia. The first law states that, if a particle is at rest or moving at a constant speed in a straight line, it will remain at rest or keep moving at a constant speed in a straight line unless it is acted upon by a force. This postulate is known as the law of inertia, and it is basically a description of one of the properties of a force—namely, that it can change rest into motion or motion into rest or one kind of motion into another kind.

Before Galileo's time it was thought that bodies could move only as long as a force acted on them and that a body free from external influences would come to rest. Those who sought to find the forces that kept the planets moving did not realize that no force was necessary to keep them moving at a practically uniform rate in their orbits; gravitational force was necessary only to change the direction of motion.

The most common impediment to steady motion is friction. A power-driven vessel is quickly brought to rest by fluid friction when the driving engine is stopped; after being struck, a hockey puck will keep on moving for a longer time on smooth ice than on a rough road. No moving body is entirely free from the effects of friction, but it seems reasonable to assume that, if a moving body were free from all external influences, it would move at a constant speed in a straight line. When a body is at rest, the forces acting on it form a balanced system—*i.e.*, the net or resultant force is zero.

Bodies at rest

The law of force. Newton's second law is a quantitative statement of the changes that a force can produce in the motion of a particle. It states that the time rate of change of the velocity—*i.e.*, the acceleration—is directly proportional to the force and inversely proportional to the mass of the particle. The larger the force ($\boldsymbol{F}$), the larger the acceleration ($\boldsymbol{a}$); and the larger the mass (m), the smaller the acceleration; thus $\boldsymbol{a} = \boldsymbol{F}/m$.

Both force and acceleration are vector quantities and as such have both magnitude and direction. The acceleration produced by a force is in the same direction as the force; if several forces act on a body, it is their resultant (sum), obtained by adding the vectors tail to tip, that produces the acceleration.

The mathematical expression of the second law can be established by considering the results of Galileo's experiments with falling bodies; he found that the force of gravity $\boldsymbol{W}$, acting alone on a body, produces a constant acceleration equal to $\boldsymbol{g}$. If a force $\boldsymbol{F}$ acts on the same body instead of $\boldsymbol{W}$ and produces an acceleration equal to $\boldsymbol{a}$, it follows from the second law that the ratio of $\boldsymbol{a}$ to $\boldsymbol{g}$ is the same as the ratio of $\boldsymbol{F}$ to $\boldsymbol{W}$—*i.e.*, $\boldsymbol{a}/\boldsymbol{g} = \boldsymbol{F}/\boldsymbol{W}$, from which $\boldsymbol{a} = \boldsymbol{F}/(\boldsymbol{W}/\boldsymbol{g})$. The quantity $\boldsymbol{W}/\boldsymbol{g}$ is equal to m, the mass of the particle, and a measure of its inertia or reluctance to having its speed changed. As stated in the law, the larger the mass is, the smaller is the acceleration produced by a given force. The second law is usually expressed in the form

$$\boldsymbol{F} = m\boldsymbol{a}. \tag{4}$$

If the force acting on a body is resolved into components along coordinate axes, each component force equals the mass multiplied by the corresponding component of the acceleration.

The law of action and reaction. Newton's third law, frequently referred to as the law of action and reaction, asserts that forces, being the result of interactions of two bodies, always appear as equal and opposite pairs. For bodies in direct contact with one another the proposition seems obvious; it is also true for some forces that act at a distance, such as gravitational attraction. As noted previously, the property of forces expressed by the third law is particularly useful in statics.

Forces in pairs

Motion under a constant force. According to Newton's second law—namely, $\boldsymbol{F} = m\boldsymbol{a}$—if the force acting on a body passes through its centre of mass and remains constant in magnitude and direction, the body will move in a straight line with a constant acceleration. This means that the body will acquire equal increments of velocity in equal time intervals. Figure 10A shows how the velocity of the body increases from zero at the beginning of a time interval t to $\boldsymbol{v}$ at the end of the interval.

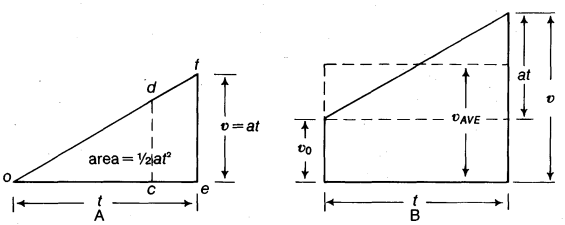


Figure 10: *Uniformly accelerated motion.*
(A) Initial velocity zero, (B) initial velocity v_0 (see text).

Because the line odf is straight, the ratios cd/oc and ef/oe, which represent the slope of the line, are equal, and either may be used to find the acceleration, which is the rate of increase of velocity per unit of time, or v/t. Using ef/oe, $\boldsymbol{a} = \boldsymbol{v}/t$; solving this equation for $\boldsymbol{v}$ yields $\boldsymbol{v} = \boldsymbol{at}$, which states that, if the velocity is zero at the beginning—*i.e.*, when $t = 0$—the velocity at the end of any time interval t is equal to the acceleration multiplied by the time.

The distance traveled in a given time by a body moving

at a constant velocity is equal to the velocity multiplied by the time. When the speed varies, it is equal to the product of the average velocity and the time. In Figure 10A the average velocity is $v/2 = at/2$, so that the distance S traveled during the interval is the average velocity times the time: $S = at/2 \times t = \frac{1}{2}at^2$, which is half the product of the acceleration and the square of the time; it is also equal to the area of the triangle in Figure 10A.

If the body were moving with an initial velocity v_0 before the force was applied to it, the relation between velocity and time—*i.e.*, the velocity–time (v, t) curve—would look like Figure 10B. In this case v is equal to the initial velocity v_0 plus the added part at—*i.e.*, $v = v_0 + at$. The distance traveled, S, is the area under the curve, which is the sum of the rectangle $v_0 t$ and the triangular piece, which is the same as the one in Figure 10A, namely, $\frac{1}{2}at^2$. The distance traveled can also be expressed as the average velocity $v_{ave} = (v_0 + v)/2$ multiplied by t. Summarizing:

$$v = v_0 + at \tag{5}$$

$$S = v_0 t + \tfrac{1}{2}at^2 \tag{6}$$

$$S = \left(\frac{v_0 + v}{2}\right)t. \tag{7}$$

Falling or rising bodies

If air resistance is neglected, these equations are valid for a falling or a rising body that is sufficiently close to the Earth that the acceleration of gravity g may be taken as a constant. For a body falling from rest, $v_0 = 0$ and $a = g$; for a body projected vertically upward with an initial velocity v_0,

$$v = v_0 - gt \tag{8}$$

$$S = v_0 t - \tfrac{1}{2}gt^2. \tag{9}$$

At the top of the rise $v = 0$, and setting $v = 0$ in equation (8) and solving for t yields the corresponding time

$$t = \frac{v_0}{g}. \tag{10}$$

Substituting this value of t in equation (9) gives $S = v_0^2/2g$, which is the height to which the body will rise.

Motion of a projectile. Consider a projectile of mass m fired from a gun with an initial velocity v_1 at an angle θ to the horizontal (Figure 11A). The projectile's motion may be considered as consisting of a horizontal part and a vertical part, the initial horizontal component of velocity being $v_H = v_1 \cos\theta$ and the initial vertical component $v_V = v_1 \sin\theta$. If air resistance is neglected, the only force on the projectile after firing is its weight, W, acting downward; there is no force in the horizontal direction. Consequently, the horizontal component of velocity will remain unaltered throughout the motion, while the vertical component will change because of the force of gravity—*i.e.*, the weight.

It is clear that the projectile will rise and then fall back to Earth after describing a curved path. During its travel the velocity v will remain tangent to the path, while v_H will remain constant, and v_V will vary as shown in Figure 11A.

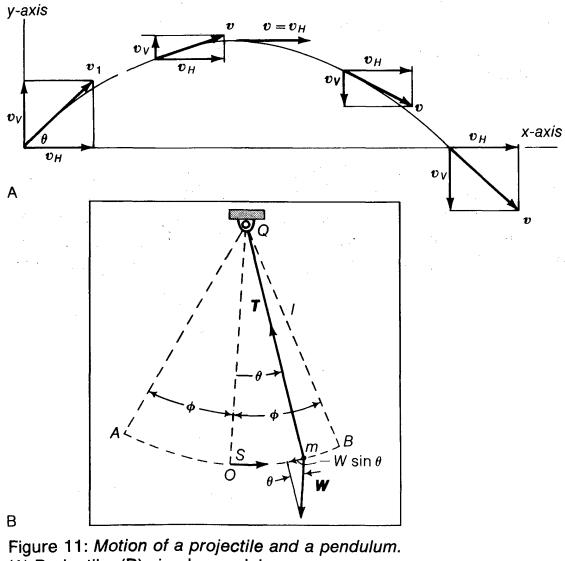


Figure 11: *Motion of a projectile and a pendulum.* (A) Projectile, (B) simple pendulum.

The distance covered in a given time by a body moving with a constant velocity is equal to the velocity multiplied by the elapsed time. Therefore x, the horizontal distance traversed in time t by the projectile, is given by $x = (v_1 \cos\theta)t$.

The vertical motion of the projectile is governed by the action of gravity and is entirely independent of the horizontal motion. It is an example of a body projected vertically upward with a velocity $v_1 \sin\theta$, and, consequently, equations (8) and (9) will apply. Letting $S = y$, $v = v_V$, and $v_0 = v_1 \sin\theta$,

$$v_V = v_1 \sin\theta - gt \tag{11}$$

$$y = (v_1 \sin\theta)\,t - \tfrac{1}{2}gt^2. \tag{12}$$

From equation (11) it may be seen that the projectile will have reached its greatest height when $v_V = 0$ and $t = (v_1 \sin\theta)/g$ and returned to Earth when $t = (2v_1 \sin\theta)/g$. Substituting this value of t in the equation for x gives $x = (2v_1^2 \sin\theta\cos\theta)/g$, which can be written $x = v_1^2 \sin 2\theta/g$.

Because $\sin 2\theta$ has its maximum value of unity when $2\theta = 90°$—that is, when $\theta = 45°$—it follows that the greatest horizontal range for a given initial velocity is obtained when $\theta = 45°$.

Ballistic calculations such as the foregoing are of little practical value when the initial velocities and the ranges are large. A precise calculation of the trajectory of a long-range projectile or guided missile must take into account air resistance, the variation in the value of g, the rotation of the Earth, and any spin of the projectile.

Long-range projectile motion

Simple pendulum. A simple pendulum (Figure 11B) consists of a small, heavy body of mass m called the bob, suspended from a fixed point Q by a cord of length l and of mass so small compared with m that it can be neglected.

If the bob is raised to position A in Figure 11B and let go, it will oscillate (swing back and forth) in a vertical plane through the angle φ on each side of the equilibrium position OQ. The forces acting on the bob are the tension T in the cord, acting upward, and the weight of the bob W, acting downward. The cord tension keeps the bob on its circular-arc path AB, and the resulting centripetal acceleration a_N, which is a consequence of the changing direction of the velocity of the bob, is directed toward Q. The tangential component of the acceleration a_T, which results from changes in the magnitude of the velocity, is produced by the component of the weight of the bob acting tangent to the path.

In Figure 11B the bob is shown in an intermediate position defined by the angle θ in radians, measured, for convenience, from the midposition OQ; the corresponding displacement S along the arc is given by $S = l\theta$, and it is arbitrarily considered positive when pointing to the right.

From Newton's second law, the summation of forces along the cord yields $T - W\cos\theta = ma_N$; from this equation the cord tension T can be calculated, but only after a_N has been determined by other means. Along the tangent the only force acting is the component $-W \sin\theta$ of W, which is negative because it points to the left; then, from Newton's second law, $-W \sin\theta = ma_T$. Now a_T can be expressed as a function of θ, namely, the second derivative of θ with respect to time, as described above under *Rotation about a fixed axis*. If this replacement for a_T is made, the previous equation becomes a differential equation; it is difficult to solve, but, if the oscillation of the pendulum is restricted to small angles, which it usually is, a simple solution can be obtained by replacing $\sin\theta$ by the angle θ in radians. On pendulum clocks the half oscillation φ is usually less than 10°, and for that angle the difference between $\sin\theta$ and θ in radians is never more than one-half of 1 percent.

Because $S = l\theta$ in Figure 11B, then $\theta = S/l$, and, if this is substituted for $\sin\theta$ in the previous equation and m is expressed as W/g, the weight W, being on both sides of the equation, will cancel out, and the equation will reduce to $-S/l = a_T/g$ or $a_T = -(g/l)S$.

The term g/l is a constant; *i.e.*, it does not vary during

the motion. Thus the tangential acceleration of the bob is proportional to its displacement (S) from the midposition and in the opposite sense. By definition, this is a simple harmonic motion, and the formula for the period (time for a complete cycle of two swings) of such a motion, previously derived for a Scotch yoke, is applicable. For the Scotch yoke $T = 2\pi - \sqrt{-S/a}$; for the pendulum $S/a_T = -l/g$ so that $T = 2\pi\sqrt{l/g}$.

A seconds pendulum is one that takes one second to make one swing (one-half cycle); then $T = 2$ seconds or $l = g/\pi^2$. Because g varies slightly on the surface of the Earth, the length of a seconds pendulum will be different at the North Pole and at the equator. An average value for l for a seconds pendulum is about one metre (slightly more than 39 inches). By measuring the length and period of a pendulum, the value of g can be calculated; this is one of the most accurate methods of determining g at any place on the surface of the Earth.

Determining the value of g

Force required for motion on a circle. When a particle moves on a circular path with a velocity that is changing in both magnitude and direction, it is convenient to treat these effects separately when seeking the forces that are their causes.

In Figure 12A, a particle of mass m in the form of a circular plate is attached at Q to a rigid bar of length R that rotates about a fixed axis at O with an angular velocity ω and an angular acceleration α. The mass of the bar is assumed to be so small compared with the mass of the particle that it can be neglected, while the total mass of the particle is assumed to be concentrated at Q. From Figure 4D the two components of the acceleration of the particle are $R\omega^2$, the centripetal or radial component directed inward along the bar, and $R\alpha$, the tangential component at right angles to the bar and with a sense compatible with a counterclockwise α (Figure 12A).

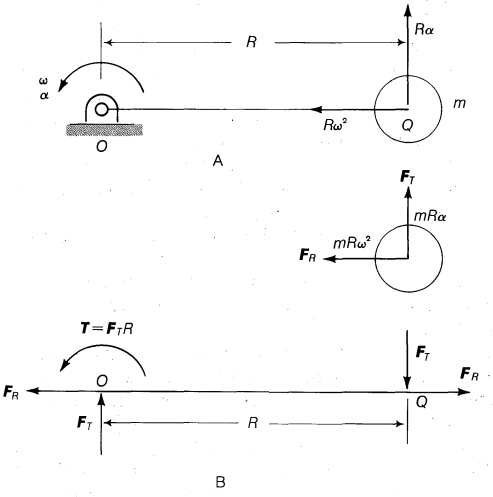


Figure 12: *Forces required for motion in a circle.*
(A) The acceleration, (B) the forces.

To determine the forces that are acting on the particle and on the bar, it will be convenient to draw free-body diagrams. A free-body diagram is a sketch showing a body isolated from all adjacent bodies and with all the forces acting on it that are significant to the problem under consideration.

Free-body diagrams

From Newton's second law the components of the acceleration of the particle are caused by the forces $F_R = mR\omega^2$ and $F_T = mR\alpha$; in Figure 12B these forces are shown acting on a free-body diagram of the particle.

From Newton's third law, action and reaction are equal and opposite. The forces F_R and F_T are exerted by the bar on the particle, and, consequently, the particle must exert equal and opposite forces on the bar; they are shown at point Q on a free-body diagram of the bar in Figure 12B. Because the mass of the bar is to be neglected, there are no inertial effects, and the forces on the bar must balance. The force F_R at Q is easily balanced by an equal and opposite force at O. To balance F_T there must be an equal and opposite force at O, but, because these forces are equal, opposite, and separated by a distance R, they constitute a clockwise couple that, for balance, requires that a counterclockwise couple or torque, $T = F_T R$, be applied (by some means) to the bar at O. Because $F_T = mR\alpha$, the torque can be expressed as $T = mR^2\alpha$.

If another particle of mass m were located diametrically opposite to the first particle on a bar at the same distance R from O, the forces created at O would be equal and opposite to those shown in Figure 12B, and they would cancel each other. The torque T, however, would be twice as much—namely, $2mR^2\alpha$.

The quantity mR^2 is a measure of the rotational inertia of a particle in the same way that m is a measure of its translatory inertia. Each particle in a rigid body has rotational inertia about any chosen axis, and the sum of all these inertias is called the moment of inertia, I, of the body.

Rotational inertia

For a rigid body $I = \Sigma mR^2$, or, using the calculus notation, $I = \int r^2 dm$, which means that every particle in the body is to be multiplied by the square of its distance from the chosen axis and the products added together. For a rigid body rotating about a fixed axis, for which its moment of inertia is I, the relation between the torque T and the resulting angular acceleration is given by

$$T = I\alpha. \tag{13}$$

This equation is analogous to $F = ma$ for translatory motion. For a circular cylinder of radius A and mass m, the moment of inertia, obtained most easily by means of the integral calculus, is equal to $mA^2/2$. It is a common and convenient practice to express the moment of inertia of a body as $I = mK^2$, in which K is the radius of gyration of the body with respect to the chosen axis; for the circular cylinder, $K^2 = A^2/2$.

Kinetics of general plane motion. As originally formulated, Newton's second law, $F = ma$, deals with the motion of particles. The law is also valid for the motion of the centre of mass of all bodies—*i.e.*, for a point in a body that is equivalent to the centre of gravity for the bodies and the motions dealt with in classical mechanics. When a body rotates, the causes and effects of the rotations must be treated separately from the motion of the centre of mass; Newton's law can be extended to take care of rotations.

If the resultant of all the forces acting on a body passes through the centre of mass of the body, it will not cause the body to rotate. In the most general case, the resultant force does not pass through the centre of mass, and, consequently, it will cause the body to rotate. By applying Newton's second law to each particle of the body, it can be shown that in all cases the centre of mass behaves as if the resultant force passed through it (whether it does or does not), while the moment, or torque, of the resultant force about the centre of mass is equal to the moment of inertia of the body about an axis through the centre of mass multiplied by the angular acceleration of the body.

If G is the centre of mass and a_G its acceleration, I_G the moment of inertia of the body about G, T_G the torque of the resultant force about G, and α the angular acceleration of the body, then

$$F = ma_G \tag{14}$$

$$T_G = I_G\alpha. \tag{15}$$

Forces and torques are the basic causes in mechanics, while accelerations are their effects. In Figure 13A the resultant force F acting along the line PQ on the body of mass m creates a torque, $T_G = Fh$, about the centre of mass G. The equivalent force and torque system are

Forces and torques

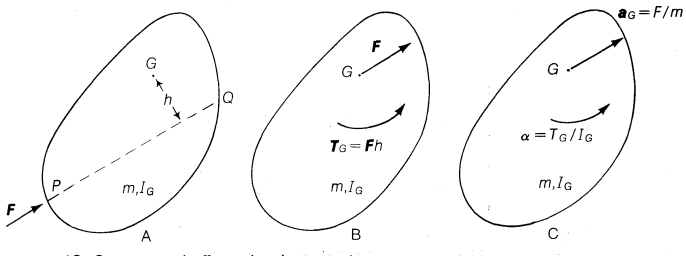


Figure 13: *Causes and effects in plane motion.*
(A) Prime cause, (B) resulting cause, (C) effects.

shown in Figure 13B, whereas the effects are shown in Figure 13C. It should be noted that, while equations (14) and (15) express relationships that exist at any instant during the plane motion of a rigid body, the magnitudes and directions of $\boldsymbol{a}_G$ and $\boldsymbol{\alpha}$ may be changing continuously on account of changes in $\boldsymbol{F}$.

Motion of centre of mass

Since the position of point P relative to G in Figure 13A affects the magnitude and direction of $\boldsymbol{T}_G$ and $\boldsymbol{\alpha}$ but not the magnitude and direction of $\boldsymbol{a}_G$, it follows that the motion of the centre of mass of a body is entirely independent of any rotation of the body. This means that the centre of mass moves as if it were a particle with the mass of the entire body and acted upon by the resultant force. Neglecting air resistance, any rotational motion of a projectile, for example, will not affect the motion of its centre of mass. Even after the projectile explodes, the centre of mass of the fragments will continue moving on the same path the shell was traveling before the explosion.

Equations (14) and (15) were derived by specifying the instantaneous acceleration of one point in a body and the angular velocity and acceleration of the body and then determining the resultant of the force system required to produce the accelerations. A similar procedure is necessary when making a dynamic analysis of the mechanism of a machine.

PRINCIPLES OF MOMENTUM AND ENERGY

Newton's second law, $\boldsymbol{F} = m\boldsymbol{a}$, describes the relation between the acceleration of a particle and the force acting on it. This equation works both ways; *i.e.*, if the acceleration is known, the force required to produce it can be calculated, while if the force is known, the acceleration that it produces can be calculated. The relationship defined by the equation is a causal one between the force, which is the cause, and the acceleration, which is the effect.

Although the acceleration $\boldsymbol{a}$ of a particle at any instant can be calculated directly from the formula if $\boldsymbol{F}$ and the mass of the particle are known, this instantaneous value of acceleration gives no information regarding the velocity of the particle. To obtain an adequate description of a particle's motion—*i.e.*, its velocity and displacement (position)—the manner in which $\boldsymbol{F}$ varies must be known. In general, $\boldsymbol{F}$ may vary with the elapsed time or with the distance that it moves or with its velocity.

If $\boldsymbol{F}$ is known as a mathematical function of t, S (distance), or $\boldsymbol{v}$, it may be possible to integrate the equation $\boldsymbol{F} = m\boldsymbol{a}$ (which is basically a differential equation) by the methods of the integral calculus and so obtain equations for the displacement and velocity of the particle. When the force is constant (nonvarying), the velocity and displacement can be found without using the calculus; when the force varies with time or distance, the results of the integration are embodied in two important and useful relations. One of the relations, known as the principle of impulse and momentum, describes what a force can do in a given time; the other, the principle of work and energy, describes what a force can do in a given distance. Both of these principles are extensions of Newton's second law, and they can be derived in a fairly satisfactory manner without using the calculus.

Impulse and momentum. When a constant force is applied for a time t to a particle moving with a velocity $\boldsymbol{v}_0$, the velocity of the particle increases to some value $\boldsymbol{v}$, as shown in Figure 10B. From this figure, if $\boldsymbol{a}$ is the acceleration of the particle, $\boldsymbol{v} - \boldsymbol{v}_0 = \boldsymbol{a}t$. From Newton's second law, $\boldsymbol{F} = m\boldsymbol{a}$ or $\boldsymbol{a} = \boldsymbol{F}/m$. Substituting $\boldsymbol{F}/m$ for $\boldsymbol{a}$ in the previous equation gives $\boldsymbol{v} - \boldsymbol{v}_0 = (\boldsymbol{F}/m)t$, which can be written $\boldsymbol{F}t = m(\boldsymbol{v} - \boldsymbol{v}_0)$. The term $\boldsymbol{F}t$, which is the product of the force and the time during which it acts, is called the impulse of the force. The product of the mass of a particle and its velocity is known as the momentum of the particle, so that $m(\boldsymbol{v} - \boldsymbol{v}_0)$ is the change in momentum. Hence, the impulse of a force equals the change in momentum produced by it; this is the principle of impulse and momentum, and it simplifies the solution of many problems that would be difficult to solve by $\boldsymbol{F} = m\boldsymbol{a}$.

The impulse of a force

The impulse $\boldsymbol{F}t$ can be visualized as the area of a rectangle t units of time long and F units of force high. When the force $\boldsymbol{F}$ is not constant, the impulse is the average force during the time interval multiplied by the time interval; this average is the imaginary (constant) force whose impulse equals the impulse of the real (varying) force. In Figure 14A the force varies from $\boldsymbol{F}_1$ at the beginning of the time interval dc to $\boldsymbol{F}_2$ at the end of the interval. The average force $\boldsymbol{F}$ is the force that when multiplied by dc gives the area enclosed by $abcd$. Consequently, the impulse of a force is the area under the force–time ($\boldsymbol{F}$, t) curve.

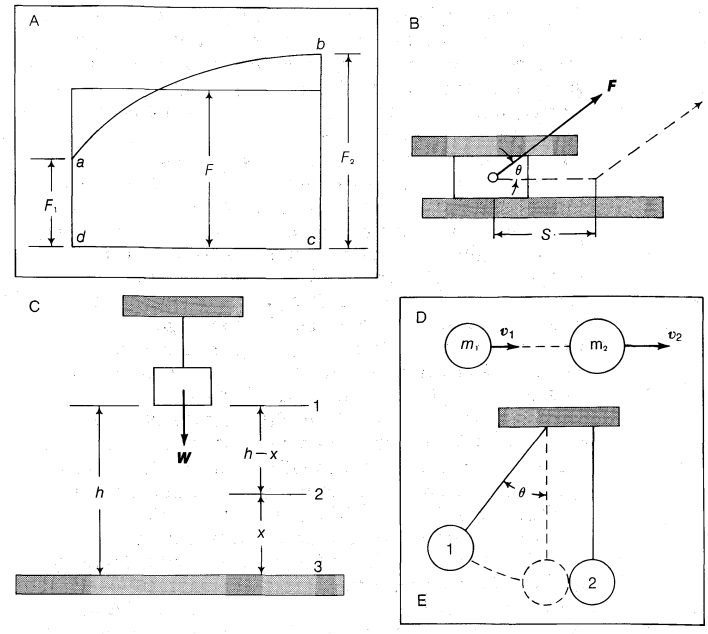


Figure 14: *Work, energy, and impact.*
(A) Average force F, (B) work done by a force, (C) potential energy, (D,E) impact, or collision.

The impulse–momentum equation is especially useful in dealing with the interactions between pairs of particles, since in such cases it may not be necessary to calculate the impulse. As an example, consider a projectile of weight $\boldsymbol{W}_1$ and a gun of weight $\boldsymbol{W}_2$, which is free to recoil. During the short interval of the explosion, the equal and opposite forces $\boldsymbol{F}$ acting on the gun and projectile, produced by the gas pressure, vary in an unknown manner, and it would be difficult to calculate the impulses of these forces. Because the forces and the times are equal, however, the impulses on gun and projectile must be equal and opposite. Thus, if the initial velocities are zero and the final velocities are v_1 and v_2 for the projectile and gun, respectively, then $\boldsymbol{F}t = (\boldsymbol{W}_1/\boldsymbol{g})\boldsymbol{v}_1 = (\boldsymbol{W}_2/\boldsymbol{g})\boldsymbol{v}_2$, from which $\boldsymbol{v}_1/\boldsymbol{v}_2 = \boldsymbol{W}_2/\boldsymbol{W}_1$. Thus the velocities of the projectile and gun after firing are in opposite directions and inversely proportional to their weights.

The projectile and the gun can be thought of as a system of two particles on which no external force acts; the forces $\boldsymbol{F}$ are internal forces. In general, when no external force is applied to a system of particles, the momentum of the system remains unchanged. This is known as the principle of conservation of momentum; the solar system furnishes an illustration of this principle. Because the nearest stars exert no appreciable attractions, the members of the system move under the action of their mutual attractions only, and the momentum of the system does not change. Consequently, the mass centre of the system moves in a straight line with a uniform velocity, just like a particle with a mass equal to the sum of the masses of the Sun and planets.

Conservation of momentum

Work and energy. When a force moves a body through a given distance in the direction in which it acts, the changes in the velocity of the body that result from this action are easily calculated if the force is constant. From equation (7), $\boldsymbol{S} = \frac{1}{2}(\boldsymbol{v} + \boldsymbol{v}_0)t$; solving this equation for t yields $t = 2\boldsymbol{S}/(\boldsymbol{v} + \boldsymbol{v}_0)$. If this value of t is substituted in the impulse–momentum equation, $\boldsymbol{F}t = m(\boldsymbol{v} - \boldsymbol{v}_0)$, the result is $\boldsymbol{F}\boldsymbol{S} = \frac{1}{2}m(\boldsymbol{v}^2 - \boldsymbol{v}_0^2)$.

The quantity $\boldsymbol{F}\boldsymbol{S}$ is the work done by the force. One-half the product of the mass of a body and the square of its velocity is called the kinetic energy of the body, or

The work done by a force

$E_k = \frac{1}{2}mv^2$. The quantity of $\frac{1}{2}m(v^2 - v_0^2)$ is, therefore, the gain in kinetic energy of the body, and, from the previous equation, the work done on a body is equal to the gain in kinetic energy; this is the principle of work and energy, and it simplifies the solution of many problems difficult to solve by $F = ma$.

The work FS can be visualized as the area of a rectangle S units of distance long and F units of force high. If the force F varies, the work is the average force for the distance multiplied by the distance moved; this average is the imaginary (constant) force that would do the same work as the real (varying) force. In Figure 14A the force varies from F_1 at the beginning of the displacement dc to F_2 at the end of the movement. The average force F is the force that when multiplied by dc gives the area enclosed by *abcd*. Consequently, the work done by a force is the area under the force–distance (F, S) curve.

When a force does not act in the same direction in which it moves, it is only the component of the force in the direction of the motion that performs work. In Figure 14B, work $= FS \cos \theta$. If a force moves in a direction opposite to that in which it acts, it does negative work.

In the treatment of impulse and momentum, it was shown that the momenta of a projectile and gun are equal and opposite because they are acted upon by equal and opposite forces for the same length of time. Because the mass of the gun greatly exceeds the mass of the projectile, however, the final velocity of the gun will be considerably less than that of the projectile. This means that the distance that the gun moves during recoil will be less than the distance that the projectile moves during the firing period. Consequently, the force, acting for a longer distance on the projectile than on the gun, will do more work on it, and the kinetic energy of the projectile will be greater than that of the gun.

A particle that is in such a state that it can do work is said to possess energy; this energy may be kinetic and the result of the particle's motion, or it may be the result of other circumstances, such as position above the Earth's surface or elastic deformation.

Potential energy

Consider, for example, the weight of mass m and weight W suspended by a cord at a height h above the ground (Figure 14C). If the cord were cut, the weight would fall to the ground, and the gravity force would do work. Thus, by virtue of its position above the ground, the weight has the capacity for doing work. This is known as its potential energy relative to the ground.

The work done by gravity on the weight must be equal to the kinetic energy of the weight just before it strikes the ground. If, as shown in Figure 14C, the initial position of the suspended weight is designated by 1, an intermediate position by 2, and the final position by 3, then $Wh = \frac{1}{2}mv_3^2$.

It is worth noting that the previous equation provides the most direct method of determining the final velocity of a freely falling body in terms of the initial height. Substituting W/g for m, $Wh = \frac{1}{2}(W/g)v_3^2$ and solving for v_3 yields

$$v_3 = \sqrt{2gh}. \tag{16}$$

In position 2 in Figure 14C the weight will have a potential energy of Wx relative to the ground and kinetic energy of $\frac{1}{2}mv_2^2$, which must be equal to the work done by gravity as the weight fell from position 1 to position 2, namely, $W(h - x)$. If E represents the total energy, P potential energy, and U kinetic energy, then, at position 2,

$$E_2 = P_2 + U_2 = Wx + W(h - x) = Wh. \tag{17}$$

Now $E_1 = Wh$ and $E_3 = \frac{1}{2}mv_3^2 = Wh$, so that in all three positions of the weight its total energy is the same. In position 1 the energy is all potential and in position 2 it is partly potential and partly kinetic, while in position 3 it is all kinetic. A system of particles in which the sum of the potential and kinetic energies remains constant is said to be a conservative system and to satisfy the law of conservation of mechanical energy; the forces involved are said to be conservative forces. When the motion of a body is opposed by friction, mechanical energy is lost; because the energy is not conserved, the force of friction is said to be a nonconservative force.

The work-energy equation can be used to approximate the force required to stop a moving particle, or the "force" of a blow. If a wooden pile is being driven into the ground by successive blows of a hammer of weight W falling from a height h on the head of the pile, the "force" of the blow will depend on the nature of the ground and the amount that the pile penetrates the ground on each blow. If the penetration d is small compared with the height h and the weight of the pile is small compared with the weight of the hammer, then the average "force" of the blow, or resistance to penetration R, will be obtained by equating the kinetic energy of the hammer to the work done by the force R during the penetration. The kinetic energy of the hammer when it strikes the pile is equal to its potential energy at the height h, namely, Wh, and the work done on the pile is Rd. Then $Wh = Rd$, from which $R = Wh/d$. This result shows that for given values of W and h the force R depends on d, which depends on the nature of the ground; the harder it is to penetrate the ground, the larger the value of R. If the ground acted like a compressed coil spring, the maximum force would be doubled.

Momentum and energy in rotary motion. Because it is applicable only to the motion of particles or the centres of mass of bodies, the equation $Ft = m(v - v_0)$ is known as the linear impulse–momentum equation. In much the same way that $F = ma$ for particles was shown to become $T = I\alpha$ when applied to the rotary motion of bodies, the linear impulse–momentum equation can be adapted to rotating bodies. In order to do this in a simple way, it will be expedient to utilize the perfect analogy between the mathematical description of motion on a straight path and rotation about a fixed axis.

Linear impulse–momentum equation

If the position of a particle on a straight path is described by a position vector that lies in the path, the vector will change only in magnitude as the particle moves, the first time rate of change being the velocity and the second the acceleration of the particle; all vectors will lie in the path.

The angular position, relative to a fixed radial line, of a line on a body rotating about a fixed axis can be described by a vector that lies along the axis and points in a direction that a right-hand screw would advance in a fixed nut if it were attached to and rotated with the body. As the body rotates the vector will change in length, the first rate of change being the angular velocity of the body and the second change the angular acceleration, just like the vectors for a point moving on a straight path.

As a consequence of the similarity of their vector representation, θ (in radians) in rotary motion is analogous to S in linear motion, ω is analogous to v, and α to a. The diagrams in Figure 10 can be adapted for rotation by a constant torque by changing v to ω and a to α. From equations (14) and (15), T is analogous to F and I to m.

Angular impulse–momentum equation

Using these relationships, the angular impulse–momentum equation can be written as $Tt = I(\omega - \omega_0)$. If the angular impulse Tt imparted to a rotating body for an interval of time is zero, then the angular momentum of the body must remain constant during the interval. This is known as the conservation of angular momentum. A figure skater may start to revolve slowly at a speed ω_1 and a moment of inertia I_1 with arms extended. By drawing the arms in close, the moment of inertia is reduced to I_2 and the speed increased to ω_2. From the previous equation, with zero impulse $I_1\omega_1 = I_2\omega_2$ or $\omega_2 = (I_1/I_2)\omega_1$.

The dimensions of linear impulse are force multiplied by time, and those of angular impulse, moment multiplied by time. Consequently, the linear momentum and the angular momentum of a body cannot be added.

The equation for work and kinetic energy of rotation can be written down at once by using the analogies; it is $T\theta = \frac{1}{2}I(\omega^2 - \omega_0^2)$. The work done by a torque is equal to the torque multiplied by the angle that it turns through in radians. Because θ is dimensionless, $T\theta$ has the same dimension as FS, and the kinetic energy of rotation can be added to the kinetic energy of translation. (A.Co.)

CELESTIAL MECHANICS

Celestial mechanics is concerned with the motion of natural and artificial bodies in space. The application of mathematics to the motion of artificial satellites, spacecraft,

rockets, and even missiles is often called astrodynamics and is considered a branch of astronautics. Dynamic astronomy deals with the forces and motions of celestial bodies and includes stellar and planetary dynamics. The terminology of these scientific areas is not well defined, and there is an undesirable and restrictive traditional association of celestial mechanics with the dynamic study of the solar system. In this section, a systematic treatment of the pertinent dynamic and mathematical principles is applied to the motion of bodies in space, regardless of their origin.

Close relation between dynamics and celestial mechanics

The laws of dynamics are described by differential equations, and an understanding of these laws is fundamental to the study of celestial mechanics. The close relation between dynamics and celestial mechanics suggests that Newton may be the founder of celestial mechanics because his laws of motion, which express the principles of mechanics, are applicable to any force field. His universal law of gravitation is one example of how his formulation and application of the calculus to dynamic problems permits an orderly and scientific treatment of the behaviour of celestial bodies. In fact, this is the only known language for treating problems in celestial mechanics, and without it the computation of orbits of space vehicles would not be possible (see above *Classical mechanics*).

A century before Newton, Johannes Kepler obtained his three laws of planetary motion empirically, but they may be derived from Newton's laws of motion and gravitation; the brilliant contributions to celestial mechanics of Nicolaus Copernicus, Tycho Brahe, and Galileo also include special cases of Newton's results, anticipating Newton, though it was he who comprehended and formulated the abstract laws.

Operational processes of celestial mechanics

The operational processes of celestial mechanics use as input either observations or, in the case of space vehicles, certain mission requirements. Next, the equations are formulated and solved. The output may be the time dependence of the position of a satellite, a set of initial conditions applicable to firing tables, the prediction of the orbit of a comet, or the location of a planet at any given time. In 3000 BC the Babylonian priest-astrologers made naked-eye observations of what were then considered to be the planetary motions (the five visible planets, the Moon, and the Sun were all considered to be planets of the Earth) in order to keep nations and kings out of trouble, to regulate the calendar, and to aid travelers. Modern preparation of lunar and planetary ephemerides (listings of locations of celestial bodies for regular time intervals), of nautical almanacs, of guidance and firing tables, and of refined relativity corrections is fundamentally the same kind of output that Mesopotamia requested from its high priests of celestial mechanics. The computations they performed might be considered clumsy today, but even the modern—but precomputer—use of Ernest W. Brown's lunar theory allowed hardly enough time for prediction of the Moon's motion. Two men were fully occupied in establishing one new position of the Moon every 12 hours. The balance between observations and computations has always been precarious, and today the enormous number of radar observations and their high precision may tax even advanced electronic computers.

Historical review. Early celestial mechanics was primarily concerned with the behaviour of the major components of the solar system. Observations were relatively crude, and sometimes completely erroneous dynamic ideas or models devoid of dynamics could offer fairly satisfactory predictions—at least for short times. This situation has a counterpart today: inadequate knowledge of the gravitational field of the Moon prevents long-term predictions of the orbits of lunar satellites.

The early Babylonian priest-astrologers were mainly observers. They invented the signs of the zodiac and recognized the precession of the vernal equinox, but there seems to be no evidence that either they or the early Egyptian astronomers established a consistent theory of planetary motions. The Greek period (from about 400 BC) of celestial mechanics is of more interest to philosophers than to astronomers. Plato's problem regarding the representation of the motion of celestial bodies by combining uniform circular motions had, at that time, two philosophical solutions: one culminating in Aristotle's 55 spheres and the other in the Ptolemaic system. Plato's problem is now seen as one of Fourier analysis, which, using trigonometric series as approximations, is today still the basic tool in treating planetary and lunar theories. The concept of the rotating celestial sphere with the Earth at its centre (even now, occasionally, a useful convention) and the concept of concentric shells used by Aristotle were attempts to represent planetary as well as stellar motion relative to the Earth. The model was intended to describe the geometry of the motion without reference to causes. Hipparchus, inventor of spherical trigonometry, emphasized the importance of observations and made Claudius Ptolemaeus' (*c.* AD 200) system feasible. The Ptolemaic system was still a purely kinematic approach to celestial mechanics, devoid of dynamic concepts. The Ptolemaic representation of the apparent planetary motions by epicycles centred at the Earth is approximate and complicated since the planets actually move around the Sun.

The Ptolemaic system

Arguments favouring a heliocentric system were mentioned by some early Greek astronomers, notably Aristarchus (*c.* 270 BC), but the first consistent theory with the Sun placed centrally in the solar system was accomplished in the 1500s by Nicolaus Copernicus. Tycho Brahe later proposed the somewhat irrelevant view that the Sun and Moon moved on circles centered at the Earth, all other planets revolving around the Sun; yet, Brahe's excellent observations were fundamental to Kepler's work.

Kepler's laws

Kepler's contributions may be considered the last of purely kinematic significance in celestial mechanics. His laws describe the motions of the planets without reference to the governing forces. His original three laws, formulated in 1609, were the ones numbered (1) and (2) below, together with a statement according to which the planes of motion of the planets pass through the Sun. His original fourth law, proposed in 1618, is shown as (3) below. The laws are as follows: (1) the orbit of each planet is an ellipse, with the Sun at one focus of the ellipse; (2) the line joining the Sun to the planet sweeps through equal areas in equal times; (3) the squares of the periods of revolution of any two planets are in the same proportion as the cubes of their mean distances from the Sun.

The first and second laws are true only if perturbations (*i.e.,* disturbances caused by additional forces) of the two-body problems are neglected. The third law is only approximately true even when perturbations are not present because it assumes that the masses of the planets may be neglected as compared to the mass of the Sun.

The ellipse of the first law (Figure 15) is a plane curve such that for any point G on the ellipse, the sum GS + GS′ is the same. In fact, S′G + SG = AP. The point P on the orbit closest to the Sun, S, is called the perihelion. The furthest, A, is the aphelion. The focuses are S and S′. The semimajor axis of the ellipse is half the distance between A and P; that is, AC or CP, usually denoted by *a;* C is the centre of the ellipse. The eccentricity of the ellipse, *e,* is the focal length CS divided by the semimajor axis CP. The line of apsides is defined by the major axis. The mean distance of a planet from the Sun is the length of its semimajor axis.

The second law reveals empirically that the shaded areas on Figure 15 are equal, provided the same time has elapsed between corresponding locations of the planet. In other words, consider a planet located at perihelion P and a month later at point F. Now, if the planet is (at a later

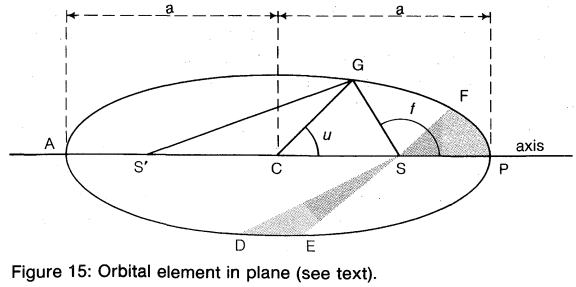


Figure 15: Orbital element in plane (see text).

time) at point D and one month later at E, then the areas of the sectors SPF and SDE are equal. Inasmuch as points F and P are closer to S than D and E, it can be concluded that the velocity of the planet is greater at perihelion than anywhere else. (Aphelion passage is the slowest and perihelion the fastest.) The angular velocity of the planet (rate of change of its angular position) varies on its orbit; the average angular velocity is called the mean motion of the planet.

The third law relates the semimajor axes of the planets to their mean motions or periods of revolution around the Sun. Once the period of revolution of a planet is known, its mean distance from the Sun can be computed relative to another planet's mean distance. The Earth's mean distance, the astronomical unit (a.u.), is used as the scale of the solar system. Consequently, any planet's mean distance in astronomical units may be obtained by computing the $^2/_3$ power of its period of revolution expressed in years. Thus, Jupiter's orbital period is approximately 11.86 years, the $^2/_3$ power of which is 5.2. The mean distance of Jupiter from the Sun (its semimajor axis) is 5.2 times that of Earth.

Kepler's contemporary Galileo, well-known for his observations, which included the discovery of four of Jupiter's satellites, made his major contribution by introducing concepts of modern dynamics. Newton's first two laws of motion were anticipated by Galileo, whose observations uncovered the fact that bodies maintain uniform motion (or are at rest) when no force is acting on them. To demolish the Aristotelian myth that rest is a "natural state" and that motion per se requires force were major steps made by Galileo before Newton's discoveries.

The Newtonian approach to dynamics

As noted earlier, Newton's three laws of motion are concerned with the effect of force on motion. To reiterate, they are (1) every body continues in its state of rest, or of uniform motion in a straight line, unless compelled by a force to change that state; (2) the rate of change of momentum (velocity × mass) is proportional to and along the line of the force; (3) to every action corresponds an equal and opposite reaction. The first law may be considered a corollary of the second. The force acting on a body is proportional to the rate of change of momentum; consequently, if there is no force, the body's momentum is constant.

Newton's law of universal gravitation states that two bodies attract each other with a force that is along the line connecting the bodies, is proportional to the product of the masses of the bodies, and is inversely proportional to the square of the distance between the bodies. The law is universal since it is applicable to the motion of a planet around the Sun as well as to the falling of a stone (or apple). The law of gravitation speaks about two bodies, the Sun and a planet, or the Earth and the stone, but the law is valid for computing the interactions of any number of bodies. The motion of two (isolated) bodies subject to mutual gravitational forces can be computed by solving the pertinent differential equations. The solution in the planetary case is an ellipse on which the motion takes place according to Kepler's laws. This result, however, is no longer empirical; it does not follow from observations as in Kepler's case, but it is obtained by analytic tools using the force as input. In fact, the motion of double stars, comets, or satellites might be computed by carefully evaluating the forces and solving the appropriate differential equations. If there are only two bodies participating (called the problem of two bodies), the solution is known to be either an ellipse, hyperbola, or parabola, but such general statements about the solution of problems in which more than two bodies participate cannot be made. The celebrated unsolved problem of celestial mechanics is the problem of three bodies that interact gravitationally with each other (see below).

Progress in celestial mechanics after Newton's work was rapid because the basic force law, as well as the equations to be solved, were now ready for application. Powerful analytic methods were developed over the next two centuries.

The Newtonian approach to dynamics is to solve the equations of motion by finding explicit expressions of the positions of the bodies as functions of the time, but such satisfying and luxuriously elegant results can be obtained only for the simplest problems of celestial mechanics. In practice, one must resort to successive approximations and series solutions, since the simple closed-form solutions are not available. Often these series represent solutions valid only for a short interval of time or, in the language of mathematics, the series are either not convergent or have a small radius of convergence. The other fundamental difficulty of series solutions is that they are unable to furnish the properties of the general solution. Nevertheless, the success of classical series solutions in obtaining approximate representations of the motion is unquestionable.

Qualitative methods of Poincaré

The next step in the development of dynamics was the application of qualitative methods. The unsurpassed master of such methods was the French mathematician and astronomer Henri Poincaré, who revolutionized the fundamental concepts of celestial mechanics. The qualitative approach does not furnish numerical descriptions of the motion; consequently, Poincaré's contributions—and those of George D. Birkhoff, A. Wintner, and A.N. Kolmogorov—were of more theoretical than immediate practical or observational interest.

Modern method. The qualitative, quantitative, and formalistic methods of celestial mechanics have clearly defined places in the aforementioned development of dynamics. The most powerful mathematical methods belong to the modern qualitative approach, and the use of electronic computers permits the full utilization of the quantitative and formalistic approaches. The formalistic approach is the classical series solution of problems of celestial mechanics used by the great mathematicians and scientists from Leonhard Euler to Dirk Brouwer. The capability of high-speed digital computers to perform algebraic manipulations, series expansions, and multiplications of series (in addition to numerical work) makes the classical formalistic approach once again fashionable, since high-order expansions may be obtained with high reliability and with great speed. The quantitative approach is popular with astronomers and space engineers because often a particular solution of a problem is sought rather than the general behaviour of the dynamic system. Examples are planetary ephemerides, representing particular solutions of the astronomers' *n*-body problems (involving the interaction of any number, *n*, of bodies) and lunar trajectories, representing particular solutions of the engineers' problem.

The recent popularity and importance of celestial mechanics are due not only to the great advances made in the development of electronic computers but also to successful launching of space probes and satellites. Computers perform computations that some years ago would have taken a lifetime of hard labour with desk calculators. The concept of experimentation in its original and numerical sense has entered celestial mechanics. Physical experimentation in classical astronomy was not possible; the universe could only be observed. However, it is now possible to launch satellites on orbits designed to obtain information regarding the Earth's gravitational field, and this is an experiment in astronomy. Modern computers also permit numerical experiments. From the point of view of dynamics, it is essential to survey the totality of possible motions and thereby to establish information about the future behaviour of a system of interest. For instance, a stone thrown with various initial velocities will have different orbits. Once a method is established whereby the orbit can be predicted from the direction and speed with which the stone was thrown, the possible motions that result can be surveyed. In space research this concept may assume the greatest interest, as the following example demonstrates.

Space missions as experiments

The orbit of an Apollo spacecraft was designed in such a way that astronauts could land on the Moon and return safely. There were many ways to do this, as many possible orbits for satisfying the requirements of such a mission exist (hitting a can with a stone and a target with a missile are similar types of "missions"). Knowing the totality of such orbits permits selection of the optimum path: the shortest flight time, minimum amount of fuel used, largest payload delivered, and so forth. Thus, from theoretical and practical viewpoints, comprehension of all

possible orbits is essential. The computer is able to furnish precisely such information when analytic methods fail.

The problem of three bodies (Earth, Moon, space vehicle) is an unsolved one in celestial mechanics. That is to say, there are no closed-form formulas describing the orbit of the space probe, but computers offer numerical solutions and help to find the totality of orbits. The dynamic system (in the case of an Apollo flight) consists of the three bodies, the motions of two of which are known since the orbits of the Earth and of the Moon are available from precomputed ephemerides. The motion of the space vehicle may be numerically integrated (*i.e.*, the differential equations may be solved) for any specific set of initial conditions including the firing date, the location, and the direction and magnitude of the initial velocity. Thousands of such trajectories can be calculated in minutes on a computer, and the "best" can be chosen.

In fact, it is possible to experiment numerically to see which planetary or satellite orbits are stable, what initial conditions result in orbits that repeat themselves (periodic orbits), or under what conditions a triple-star system will break up into a binary and an escaping star. The last example refers to the case when two of three stars form a two-body system, revolving around each other, while the third star departs from this binary.

New developments in computational techniques are necessary to satisfy the operational requirements imposed by attempts to solve new problems. An outstanding new problem resulting from space-probe activities is related to the accurate computation of orbits with close approaches.

Orbits with close approaches

Close approaches are common occurrences—in fact, often they are operational requirements—for space probes, while natural celestial bodies seldom have close approaches. The consequence is far-reaching, and it can be comprehended when the structure of the Newtonian gravitational force field is considered. The force acting between two bodies becomes large when their mutual distance decreases. The relative velocities and accelerations also become large, and computational inaccuracies might occur just at the time when increased accuracy is usually mandatory. When the space probe is in the immediate vicinity of a target planet, the accuracy of the orbit prediction must be great; while it is coasting between centres of gravitation, the preciseness of its orbit is not critical. Close approaches may occur sometimes with asteroids or with comets, and ingenious ad hoc methods have been devised for accurate orbit computations by modifying the classical methods. Hansen's method of partial anomalies, one such technique, splits the orbit into segments and describes the motion by different variables on each segment. Changes of relative distances between bodies in the solar system are not as great as in the case of space probes. The Newtonian force field shows a "singular" behaviour when the participating bodies are close: the force becomes large and the slightest computational errors are amplified. Such a singularity may be eliminated in most problems in celestial mechanics by proper selection of variables to describe the motion. This process of mathematical regularization was investigated by Tullio Levi-Civita, Poincaré, and K.F. Sundman.

These mathematical methods were not used much in classical celestial mechanics until the peculiar orbit requirements of space exploration made their applications essential. The theory of regularization, or smoothing of the singular behaviour of the force field, has now become a characteristic approach to celestial mechanics. It stabilizes computational processes, speeds up computer solutions, and simultaneously reduces computational errors. In a special form, the method was already known to Kepler, who made use of the eccentric anomaly (see the angle GCS = u in Figure 15), in addition to the true anomaly (angle GSP = f) and the mean anomaly. (This latter is the angular measurement of time expressed in radians along the orbit so that, after the planet completes one orbit around the Sun, its mean anomaly increases by 2π.) Introducing the eccentric anomaly (instead of the time) as a variable to describe the motion is the basic method of regularization. Sophisticated regularization techniques may be considered generalizations of the use of the eccentric anomaly and consist of the introduction of new dependent as well as independent variables. These methods have been applied with considerable success to calculations of lunar trajectories as well as to other problems in modern space dynamics.

Regularization techniques

Problem of two bodies and perturbations. The orbits of two gravitationally attracting bodies are ellipses, hyperbolas, or parabolas, depending on the initial conditions of the motion. The orbits of natural and artificial planets and satellites would be ellipses if perturbations did not influence these orbits; any effects that result in deviations from the elliptic orbits are called perturbations. The gravitational effect of Jupiter, for instance, on the motion of the Earth around the Sun is a perturbation, because, if the Earth's motion were only under the Sun's influence, its orbit would be an ellipse. One of the classic triumphs of celestial mechanics was the discovery of the planet Neptune in 1846 from perturbations of Uranus.

Deviations from elliptic orbits may be caused by effects other than that of a third body. The orbit of an Earth satellite is a perturbed ellipse because the Earth does not consist of homogeneous, concentric, spherical shells, and its mass distribution is distorted. Such oblateness perturbations on the elliptic orbits of natural and artificial satellites can be evaluated today without difficulty, and it is known that the orbits are not stationary ellipses. Another perturbation on the orbits of close Earth satellites is caused by the atmosphere. Atmospheric effects tend to circularize elliptic orbits and progressively reduce the satellite's altitude until it hits the surface of the Earth or is destroyed by heat upon reentering the lower atmosphere. Solar radiation pressure is another perturbation that may become important for large, light, balloon-type satellites. An originally circular orbit is distorted by the radiation effect into an ellipse with decreasing perigee distance until the satellite is destroyed.

Calculating perturbations

In order to describe the effects of perturbations, it is necessary to introduce the conventional parameters describing a satellite orbit in space (see Figures 15 and 16). In addition to the previously mentioned orbital elements, the semimajor axis (a) and the eccentricity (e), which together determine the shape and size of the elliptic orbit, four more quantities (also called parameters) are needed. The inclination (i) of the orbital plane to a fixed plane such as the equatorial or the ecliptic plane (in which the Earth moves) and the location of the ascending node Ω on the ecliptic define the plane of the orbit. The fifth orbital element, ω, determines the orientation of the ellipse in its plane by the location of the pericentre, the point of the orbit closest to the centre of attraction. The last parameter specifies the location of the body on its orbit by the time, T, of its perihelion (or perigee, etc.) passage or the time when any other position along the orbit is occupied. The ascending node Ω is the point where the body in orbit

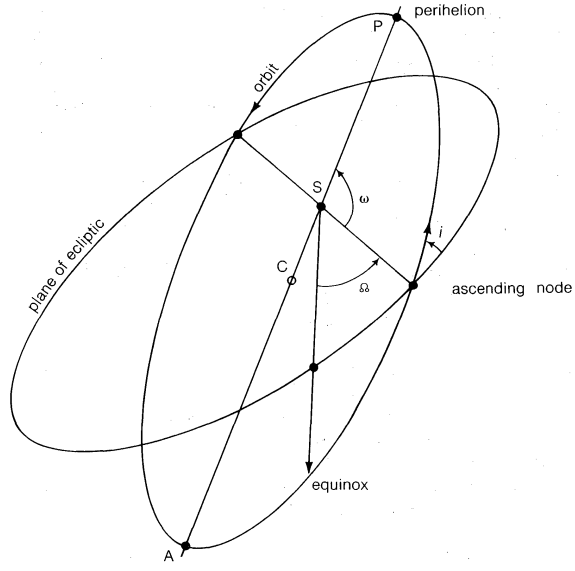


Figure 16: Orbital elements in space (see text).

passes from north to south of the ecliptic plane, and it is described by its angular position measured from a point on the ecliptic plane, such as the vernal equinox, at a given time. The six quantities a, e, i, Ω, ω, and T, or any equivalent set, define the unperturbed orbit of a planet around the Sun or of a satellite around a planet. Perturbations will result in changes in these orbital parameters. At any instant one may think of the body moving on a specially varying ellipse called the osculating ellipse that simulates the actual momentary path quite closely. If the perturbations suddenly stopped, the body would then continue traveling on the same ellipse.

Secular and periodic effects

Those changes in the shape and position of the two-body orbit that progress always in the same direction with time are called secular, while certain others, varying in intensity and direction, are called periodic. Secular effects are of great importance as they accumulate with time.

The principal effects of the oblateness of a planet on the motion of its satellites are a backward motion of the node Ω and a motion of the apsidal line ω. Hence, by carefully measuring satellite orbits, the gravitational field of planets can be determined. Specially designed satellites are observed to obtain information about the internal mass distribution of the Earth. This procedure is called satellite geodesy.

A few numerical results concerning an artificial Earth satellite may be of interest. Consider an orbit inclined at 30° to the Earth's equatorial plane, with perigee altitude of 160 kilometres and with apogee altitude of 840 kilometres. The eccentricity is 0.05 and the semimajor axis is 6,875 kilometres. The orbital period of this satellite is 92 minutes; therefore, it performs 15.7 revolutions per day. The angle of the ascending node changes 7° per day (*i.e.*, the nodal line regresses by 7° per day), or in 51.4 days the orbital plane precesses 360° (*i.e.*, it makes a complete revolution). The argument of the perigee changes 11° per day.

An interesting special case occurs when a satellite is placed in a synchronous orbit in the plane of the equator. The period is 24 hours, and consequently without perturbations the satellite is stationary relative to the rotating Earth. The slight ellipticity of the equator, however, causes a resonance effect between the Earth and the satellite that might result in instability of the satellite's orbit under certain conditions. If the satellite is placed on the line that is the extension of the minor axis of the equator, it is stable. At other locations the satellite will drift away from its stationary position. The ellipticity of the equator measured by the difference between the major and minor semiaxes is 64 metres. The major axis of the ellipse is oriented approximately at 36° west.

The stability of the solar system. Perturbations vary the orbital elements of the motion of celestial bodies. These variations may be computed using Lagrange's equations that express the time rate of change of the orbital elements. These differential equations are solved numerically on a computer or analytically by successive approximations. The equations are exact, but both the numerical and the analytic methods provide only approximate solutions and present fascinating and important unsolved dilemmas regarding the long-time future of the solar system. If, from such approximate calculations, for instance, the eccentricity of the Earth's orbit is shown to increase proportionally with time, the calculations will predict that the orbit eventually will change so much that the Earth may depart from the solar system or fall into the Sun. Yet, when higher-order approximations are computed, an opposite or compensating trend may be seen. It is essential to realize the limitations of the analytic methods, as well as the restricted applicability of computers when long-time effects are to be evaluated. At present, questions that are related to the long-time stability of the solar system cannot be answered either analytically or computationally.

Use of Lagrange's equations

These questions involve a fundamental unsolved problem of celestial mechanics and also of dynamics: whether or not some characteristics of unperturbed, highly idealized elliptic orbits will survive slight perturbations that last for a long time. Certain modern-day mathematical results allow an approach, in principle, to the stability problem for long time periods in an entirely different manner, according to which the quasi-periodic behaviour of the solar system will prevail. Very small perturbations will not disturb its stability provided that conditions excluding certain resonances in the system are satisfied. Unfortunately, the application of this principle to celestial mechanics encounters two serious difficulties. First, the theory contains certain assumptions whose validity for the solar system is questionable. Second, Newtonian gravitational forces are not the only effects influencing the orbits of the members of the solar system. Atmospheric drag, radiation pressure, relativity effects, and others must be considered as well. These forces are negligible at any instant, but their cumulative actions for several times 10^9 years may influence planetary motions significantly. In addition, there may be unknown effects and perturbations; these naturally cannot be expressed or represented in the equations, and it must be emphasized that such unknown factors can have only very small instantaneous effects; practically speaking, they cannot be detected.

Influences on orbits

The classical problem of three bodies. Because of its historical, dynamic, astronomical, and astronautic significance, the problem of three bodies, the "most celebrated of all dynamical problems" (E.T. Whittaker, 1904), is described separately. This is a strikingly simple problem to state, but its solution is still not available. Three bodies attract each other according to the Newtonian gravitational law. The bodies are thought of as point masses; *i.e.*, their gravitational fields correspond to those of the concentric homogeneous spheres mentioned previously. These bodies are free to move in space and are initially supposed to be moving in any given manner. The problem is to determine their subsequent motion. The solution cannot yet be expressed mathematically in finite terms. The number of articles and memoirs published on this problem is well over 1,000 and includes contributions of the greatest mathematicians from Newton to the present. The motion of the Moon in the field of the Sun and the Earth is a classic example. The motion of a planet around the Sun as perturbed by Jupiter, the motions of the satellites of the planets perturbed by the Sun, and, in astronautics, the orbits of space vehicles to the Moon as affected by the Moon and the Earth are others.

A version of the three-body problem of practical importance occurs when one of the three participating bodies is much less massive than the other two. The approximate ratios of the masses of the Sun, Earth, and Moon are 300,000:1:0.01. The mass ratios of the Earth, Moon, and a space vehicle are about $100{:}1{:}10^{-19}$; *i.e.*, the mass of a six-ton space vehicle is represented in this ratio by 18 zeroes and a one following a decimal point. When one of the participating bodies has such a small mass, its effect on the other two bodies may be neglected. In this way, the problem is split into a two-body problem (which is solvable) and the problem of determining the motion of the third (small) body. Once the first problem is solved, the third body is introduced into the known and moving (time-dependent) gravitational fields of the two large bodies, often called primaries, and an attempt is made to determine the motion of this single small body. The motions of the primaries will not be measurably influenced by the small body; the motion of the small body is governed entirely by the primaries. This problem, originated by Euler in the 18th century, is known as the restricted three-body problem, but even in this simplified form, the problem is not solved.

Restricted three-body problem

Unsolved problems in dynamics may have certain characteristics, which, even though the complete solutions are unknown, will promote the understanding of the behaviour of the system. Such characteristics are usually either qualitative or are given in closed-form mathematical expressions, representing some, but not all, of the relations between the variables involved. For instance, in the gravitational problem of any number of bodies, the total energy (the sum of the kinetic and potential energies) is constant, for there is no dissipation of energy. This is the principle of the conservation of energy, and its mathematical expression is known as the integral of energy. The kinetic energy is proportional to the masses and to the

Conservation of energy

squares of the velocities of the participating bodies; the potential energy is related to the relative positions of the bodies. Therefore, the conservation of energy principle can be expressed by an equation relating the velocities to the positions of the participating bodies at any instant. While it does not represent the solution of the problem, it does contribute to the understanding of the behaviour of the bodies. Using such an expression, the American mathematician George W. Hill arrived at the conclusion that the Moon will never depart from the Earth, provided the assumptions of the restricted problem of three bodies are valid. Regions limiting possible motions of the bodies may be established in this way without solving the problem.

An interesting application of the three-body problem is Lagrange's prediction (1772) of the existence and location of the Trojan planets, which, 134 years later, were actually observed. These minor planets are found around the triangular (equilibrium) points of a three-body system, Jupiter and the Sun occupying the other corners of an equilateral triangle. At the points, there are no resultant forces acting on the asteroids, so they are in equilibrium.

The configuration is shown schematically in Figure 17. The line of syzygies is L_1L_3, connecting the primaries (m_1 = Sun and m_2 = Jupiter). Point C is the fixed centre of mass around which the Sun and Jupiter revolve approximately in circles. As a consequence, the rigid triangles $m_1L_4m_2$ and $m_1L_5m_2$ also revolve, and L_4 and L_5 also describe circles. The minor planets "librate" (oscillate) around L_4 and L_5 in stable orbits. Lagrange also showed that, if the third particle is placed at any of the three collinear points (L_1, L_2, L_3), the ever present perturbations will eject it; in other words, these points represent unstable solutions. The stability of L_4 and L_5 when m_1 = Earth and m_2 = Moon is influenced by the Sun, among other perturbing effects. Mathematical and observational evidence for the possible existence of material particles around the Earth–Moon triangular libration points is still lacking.

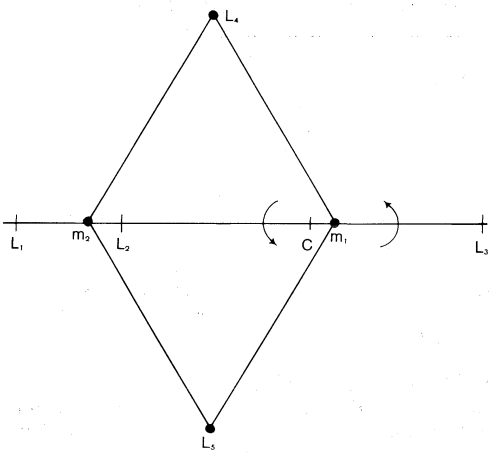


Figure 17: The Lagrangean points L_1–L_5. $L_1L_2L_3$ constitute a straight line; $L_4m_1m_2$ and $L_5m_1m_2$ are equilateral triangles (see text).

Periodic solutions

Other special solutions of the problem of three bodies are periodic solutions, in which the three bodies periodically return to their original position with their original velocities and repeat their motion. Poincaré's unproved conjecture in celestial mechanics states that all solutions may be approximated by periodic solutions. This lends considerable importance to the numerical and analytical investigation of periodic orbits. The planetary, lunar, and satellite theories are all based on the idea of using trigonometric functions and periodic solutions as reference orbits, which, when improved, may represent the actual solutions.

When such periodic terms are used in the solution, they introduce small divisors resulting in large amplitude motions. This phenomenon, described by Laplace in 1780, is related to resonance, and one of the most famous examples is "the great inequality" in the motion of Jupiter and Saturn. These planets are nearly in resonance since their mean motions are approximately in the ratio of 5:2.

Another important problem of celestial mechanics is connected with the possibility of capture. This formidable mathematical problem may be stated in a variety of ways depending on the orientation of interest. A practical application is the origin of the Earth–Moon system. From the point of view of celestial mechanics, it is essential to know if the Moon could have been captured by the Earth, being originally an independent planet with an orbit around the Sun close to the orbit of the Earth. If the assumptions of the restricted problem are valid, this could not have happened, but with tidal effects or dissipative mechanisms operating, the answer could be affirmative.

Capture problem

The systematic numerical survey of the totality of possible solutions in the restricted problem was begun in 1913 by Svante Elis Strömgren at the Royal Copenhagen Observatory. The corresponding survey for the much greater variety of possible motions in the general three-body problem was made possible only by the use of high-speed computers, and it is in progress at the present time.

The problem of *n* bodies. This is a generalization of the three-body problem for the case when the number of bodies (n) is larger than three. All participating bodies attract each other according to the Newtonian law of gravitation and are free to move in space. Initially all bodies have arbitrary positions and velocities. The problem is to find their subsequent motions.

Equations like the integral of energy are again available to establish certain properties of the motion. Nevertheless, the mathematical complications limit the possibility of finding explicit solutions analytically, and, once again, numerical solutions become of considerable interest. Elaborate computers have been used to attack the problem when $n = 10$, 25, 50, and up to several hundred. The complexity of these programs is revealed when it is realized that for 100 bodies the simultaneous solution of 600 first-order differential equations is required. At every step of the computation almost 5,000 mutual distances and the corresponding inverse squares must be computed. Any time two or more of the participating bodies perform close approaches, the accuracy of the computational process is jeopardized. Such computations could not have been performed before the use of electronic computers.

Depending on the number of bodies and on the initial conditions, the n-body problem corresponds to various physical systems. The solar system has been modeled in this way, and planetary ephemerides have been obtained for thousands of years. If n is sufficiently large (the order of a hundred), the problem corresponds to the motion of stars in a globular cluster. Computations have shown the dynamic formation of binaries and the phenomenon of escaping stars. Trends regarding the stability of the system also are indicated. It is known that, due to close approaches, a large inherent instability is present in the calculations that makes the results of the numerical integrations questionable. (V.G.S.)

Other concerns of mechanics

INERTIA FORCES AND THE CORIOLIS FORCE

It was pointed out by Jean Le Rond d'Alembert that Newton's second law could be interpreted in a different way by writing it in the form $\boldsymbol{F} + (-m\boldsymbol{a}) = 0$ and treating the term $(-m\boldsymbol{a})$ as if it were a force. In this form $(-m\boldsymbol{a})$ is called an inertia force to distinguish it from such real forces as the force of gravity or the force exerted by one body on another by direct contact. The inertia-force concept makes it possible to reduce a problem in dynamics to a problem in statics by writing the second law in the form $\Sigma\boldsymbol{F} = 0$ and including the inertia forces in the summation. The term $(-m\boldsymbol{a})$ is sometimes called a reversed effective force and also a fictitious force.

Fictitious forces

When a particle is moving on a curved path at a constant speed, the force and the centripetal acceleration are both directed toward the centre of curvature of the path. The corresponding outward inertia force, equal in magnitude to the centripetal acceleration, is called the centrifugal force. Neglecting gravity, a stone whirling in a circle in a horizontal plane at the end of a string has only one real force acting on it—namely, the tension in the string. The string tension keeps the stone on its circular path, and,

because it continually alters the direction of the velocity of the stone, it creates the centripetal acceleration. Using the inertia-force viewpoint, the stone is in equilibrium—*i.e.*, has zero acceleration—under the influence of the inward string tension and the outward centrifugal force. A real force, however, equal to the centrifugal force in both magnitude and direction, is acting on the member at the centre of the circle to which the string is attached.

When a body is at rest, the external forces (loads and reactions) acting on it are in equilibrium. If the body is assumed to be cut in a region that is stretched, bent, or twisted, the stresses acting on the cut cross section must balance the external forces on the portions of the body on either side of the cross section. When a machine or structural member is being accelerated, the external forces are not in equilibrium, and the stresses on any cross section do not balance the loads and reactions. By employing d'Alembert's principle this difficulty can be overcome, and a problem in dynamics becomes a problem in statics, for which the equations of equilibrium can be used.

Application of d'Alembert's principle

D'Alembert's principle is particularly useful in deriving the equations for the stresses in bodies rotating about fixed axes. On a disk rotating at a constant speed, for example, each cubic element of the disk is assumed to be in equilibrium under the action of the stresses acting on its sides and the outwardly directed centrifugal (inertia) force. To a person standing on it, the particles of the disk do appear to be in equilibrium because they do not move relative to the disk. In the theory of elasticity (see below *Mechanics of elastic bodies*), inertia forces and gravitational forces are classified as body forces; in the rotating disk, the centrifugal forces, because they vary with the distance of the element from the axis, have the same effect on the stresses as varying gravitational forces. From a similar point of view, a centrifuge or a cream separator is a system in equilibrium to an observer rotating with it and can be more easily explained in terms of centrifugal force than in terms of centripetal force.

In the treatment of relative motion in Figure 5D, the absolute acceleration $\boldsymbol{a}$ is shown as the resultant of the Coriolis component $2\boldsymbol{v}_R\boldsymbol{\omega}$ and the centripetal component $\boldsymbol{a}_V = S\boldsymbol{\omega}^2$, in which S is the magnitude of the vector $\boldsymbol{S}$. If m is the mass of the car, the track must be exerting a force on the car equal to $m2\boldsymbol{v}_R\boldsymbol{\omega}$, collinear with the acceleration vector and in the same direction; this is the Coriolis force. There must also be a centripetal force $m\boldsymbol{a}_V$, collinear with $\boldsymbol{a}_V$ and in the same direction, generated by the mechanism that drives the car on the track.

The Coriolis force is always present when a particle moves on a rotating body or when a body rotates about an axis that is fixed on another rotating body. The torque in a fluid coupling or converter is created by the Coriolis forces that result from the relative motion of the fluid; on an automobile making a turn to the right, elements in the upper half of the rotating flywheel experience a Coriolis acceleration forward and those in the lower half an acceleration backward. The Coriolis forces corresponding to these accelerations are produced in the crankshaft bearings; they are known as gyroscopic couples. The behaviour of gyroscopes can be explained in terms of Coriolis forces.

Because of its rotation about an axis joining the poles, the Earth is a rotating frame of reference, and the Coriolis and centripetal components will appear in the acceleration equations for particles moving in certain directions and in certain locations on and near the Earth's surface. Because the Earth rotates at only 0.000694 revolution per minute (0.0000729 radian per second), the forces associated with these accelerations are small and can usually be neglected; they do have some geophysical importance, however.

Like the expansion of a rotating disk caused by the centrifugal forces, the centrifugal forces on the Earth's crust are responsible for the equatorial bulge; the forces are greatest at the equator because the centrifugal force ($-mR\omega^2$) depends on the distance R from the polar axis.

The Coriolis drift

In Figure 5D the real Coriolis force from the track is pushing to the left on the car; the inertia force from the car is pushing to the right against the track, and, if the track were not there, the car would slide to the right. This deviation is known as the Coriolis drift, and it has a marked effect on the paths of long-range projectiles and the direction of trade winds and ocean currents.

To complete the treatment of relative motion and the corresponding forces, it will be helpful to reexamine the implications and validity of Newton's second law, $\boldsymbol{F} = m\boldsymbol{a}$. As stated, the law applies to the absolute motions of particles—*i.e.*, their motions in a reference frame or coordinate system attached to the fixed stars. This is called either a Newtonian frame or an inertial frame. If the motion of the particle is described in terms of a moving coordinate system—*i.e.*, one attached to a moving vehicle—it may be necessary to modify Newton's equation.

In general, when a particle moves on a nonrotating vehicle, its absolute acceleration can be expressed as the vector sum of the absolute acceleration of the point on the vehicle coinciding with the particle at the instant considered ($\boldsymbol{a}_V$) and the acceleration of the particle relative to the vehicle ($\boldsymbol{a}_R$). Thus $\boldsymbol{a} = \boldsymbol{a}_V + \boldsymbol{a}_R$, it being understood that the plus sign signifies a vector addition. Then Newton's second law takes the form $\boldsymbol{F} = m(\boldsymbol{a}_V + \boldsymbol{a}_R)$.

If the vehicle is moving with a uniform velocity—*i.e.*, on a straight line at a constant speed—then $\boldsymbol{a}_V = 0$ and $\boldsymbol{F} = m\boldsymbol{a}_R$. This means that Newton's law applies without correction to, for example, motions relative to a train moving on a straight track at a constant speed.

If the vehicle is accelerating, then $\boldsymbol{a}_V$ is the same for all points in the vehicle and can be added vectorially to $\boldsymbol{a}_R$ to obtain the absolute acceleration $\boldsymbol{a}$. Newton's equation, however, can be written as $\boldsymbol{F} - m\boldsymbol{a}_V = m\boldsymbol{a}_R$; in this form it signifies that the law applies to accelerations relative to an accelerating nonrotating vehicle, provided an inertia force of magnitude $-m\boldsymbol{a}_V$ is added to the left side of the equation.

If the vehicle rotates, the Coriolis component $\boldsymbol{a}_C$ appears, and the absolute acceleration $\boldsymbol{a} = \boldsymbol{a}_V + \boldsymbol{a}_R + \boldsymbol{a}_C$. When Newton's law is written in the form $\boldsymbol{F} - m\boldsymbol{a}_V - m\boldsymbol{a}_C = m\boldsymbol{a}_R$, it signifies that the law applies to acceleration relative to a rotating vehicle, provided the inertia forces $-m\boldsymbol{a}_V$ and $-m\boldsymbol{a}_C$ are added to the left side of the equation.

In Figure 5D there is no relative acceleration $\boldsymbol{a}_R$, and the vehicle acceleration $\boldsymbol{a}_V$ is simply the centripetal acceleration $S\boldsymbol{\omega}^2$; if the vehicle were being accelerated, there would be another component $S\boldsymbol{\alpha}$. Consequently, $\boldsymbol{F} - m\boldsymbol{a}_V - m\boldsymbol{a}_C = 0$, and to an observer on the table the inertia forces $-m\boldsymbol{a}_V$ and $-m\boldsymbol{a}_C$ balance the two components of $\boldsymbol{F}$, one from the bottom of the track and one from the side of the track; *i.e.*, the car is in relative equilibrium.

On the surface of the Earth $-m\boldsymbol{a}_V$, the centrifugal force, and $-m\boldsymbol{a}_C$, the reversed Coriolis force, are both small and can usually be neglected. Then $\boldsymbol{F} = m\boldsymbol{a}_R$, and Newton's second law can be applied, without correction, to motions relative to the Earth.

MECHANICS OF COMPLEX SYSTEMS

The principle of virtual work. The solution of problems in statics by the equations of equilibrium—*i.e.*, by force and moment summations—can become tedious for complex structures and mechanisms. To determine forces by the equilibrium method it is usually necessary to dismember the structure and treat each component separately; when the number of components is large, many equations must be solved. In addition, even if only one or two unknown forces are to be found, in moving from one component to the next, many unwanted forces may appear and require evaluation. The method of virtual work is an alternate approach to statics problems that is well suited for complex systems and for cases in which only a few unknown forces are to be found. It can also be used for the dynamic analysis of mechanisms, provided the inertia-force concept of d'Alembert is utilized to create a state of equilibrium in each component.

Applicability of the virtual work method

The principle of virtual work, formulated by Johann Bernoulli in 1717, states that, if a rigid body or a system of connected rigid bodies is in equilibrium, then the total virtual work of the external forces acting on the body or system of bodies is zero for any arbitrary virtual displacement of the body or the system consistent with the constraints.

When the force $\boldsymbol{F}$ moves a very short distance, $\Delta\boldsymbol{S}$, in

Virtual displacements

the direction in which it is acting, the mechanical work done is $F\Delta S$. The displacement ΔS is "virtual" if it is not actually experienced but merely imagined in the course of an analysis; it is usually assumed to be infinitesimal in order that the geometry of the system be maintained and that F and ΔS be always in the same direction. The product of F and a virtual displacement is virtual work. The principle can be illustrated by means of the simple lever in Figure 18A.

For the lever, a virtual displacement consistent with the constraints is a rotation through a small angle $\Delta\theta$, in radians, about the fulcrum f. The virtual displacements of P and W are $a\Delta\theta$ and $b\Delta\theta$, respectively. Because P goes down during the displacement, its virtual work is positive; the work of W is negative because W and $b\Delta\theta$ are in opposite directions. The sum of the virtual works is thus given by $P(a\Delta\theta) - W(b\Delta\theta) = 0$, from which $P/W = b/a$. In this case, the principle of virtual work offers no real advantage; it seldom does for a single rigid body.

The device in Figure 18B consists of eight weightless pin-connected bars that are held at A and B and subjected to a load P; the reactions at A and B are required. When point B is displaced a distance Δ to the right by a force Q, the displacement will be consistent with the restraint at A. Then, from the nature of the device, the corresponding displacement of P will be $\Delta/3$ to the right. The sum of virtual works is $Q\Delta - P\Delta/3 = 0$, from which $Q = P/3$, pointing to the right. By similar means the reaction at A can be shown to be $2P/3$, pointing to the right. This problem would be more difficult to solve by any other means.

The principle of virtual work, or, as it is sometimes called, the principle of virtual displacements, was originally known as the principle of virtual velocities. Because all of the displacements are simultaneous, their relative magnitudes are unaltered if divided by equal time intervals; the resulting quotients are average velocities that become instantaneous velocities when the time interval approaches zero. The use of instantaneous velocities instead of displacements has many advantages when the principle is applied to machine mechanisms and leads to a basic velocity–force–power relationship in machines.

If a virtual displacement ΔS takes place in a time interval Δt, then $F(\Delta S/\Delta t)$ is the average time rate at which the work is being done—or the average power—while $\Delta S/\Delta t$ is the average velocity of the force. In the limit, as Δt approaches zero, $\Delta S/\Delta t$ becomes v, the instantaneous velocity, and Fv the instantaneous power.

Mechanical advantage

On such simple machines as the lever, the wheel and axle, and the screw, the ratio of the load (output force) to the effort (input force) is called the mechanical advantage (MA). The ratio of the motion of the effort to the motion of the load is known as the velocity ratio (VR). If friction is neglected, the mechanical advantage is the reciprocal of the velocity ratio. Because of friction, the mechanical advantage is always less than the velocity ratio, and the ratio MA/VR is the mechanical efficiency of the machine. The efficiency of machines that rotate continuously is most easily obtained by measuring the instantaneous output and input powers and dividing the output by the input.

The principle of virtual work can also be applied to rotating bodies; it is only necessary to replace the force F by the torque T and ΔS by $\Delta\theta$ (in radians). Then the virtual work is $T(\Delta\theta)$, and instantaneous power becomes $T\omega$. In the case of the lever in Figure 18A, the torques are Pa and Wb, in which a and b are the lever arms, so that $(Pa)(\Delta\theta) - (Wb)(\Delta\theta) = 0$, which leads to the same result as before.

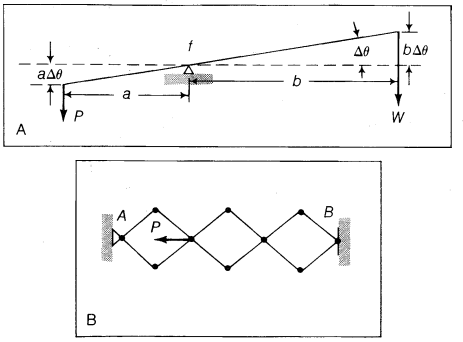


Figure 18: *Applications of the principle of virtual work.* (A) Simple lever, (B) linked structure.

Dealing with rotating bodies

Spinning tops and gyroscopes. In the previous sections the force systems and the motions of the particles and bodies were confined to parallel planes. There were two reasons for restricting the treatment to two dimensions; in the first place, most of the problems of interest to physicists and engineers can be solved by the equations of plane mechanics; in the second place, solutions to spatial (three-dimensional) problems in statics can be effected most easily by the formalized procedures of vector analysis, while some of the problems in spatial dynamics are so complicated that solutions have not yet been obtained.

In three-dimensional space, as in two-dimensional space, the centre of mass of a body moves as if it were a particle with the resultant of the external forces acting on it. In plane motion the sum of the torques about the centre of mass equals $I_G\alpha$, which is equivalent to assuming that the body rotates about an axis passing through the centre of mass and perpendicular to the plane of motion. In space motion there are three planes and three axes through the centre of mass about which the body may be rotating simultaneously; this is the reason for the complexity of space motion. In special cases, such as the motion of spinning tops and gyroscopes, satisfactory explanations of dynamic behaviour can be obtained by making simplifying assumptions.

It has been shown that when a body rotates about a fixed axis the torque impulse Tt is equal to the change in the angular momentum $I(\omega - \omega_0)$. Because t is the time interval during which the torque acts, it may be written as Δt, and because $I(\omega - \omega_0) = I\omega - I\omega_0$ is the corresponding change in the angular momentum, it may be written as $\Delta(I\omega)$. Then $T(\Delta t) = \Delta(I\omega)$, from which $T = \Delta(I\omega)/\Delta t$, or the torque is the time rate of change of the angular momentum. When a body rotates about a fixed point, a similar equation applies; in this case, T is replaced by M, the moment of the external forces about the point, and $I\omega$ is the resultant angular momentum of the body. Then, in the limit, as Δt approaches zero, $M = d(I\omega)/dt$.

In Figure 19A a top in the form of a disk of weight W, mounted on a shaft with a pointed end, is held in a frictionless support at O in the frame. The shaft rotates at a high speed ω in a clockwise direction viewed from O. If the top were stationary, the moment of the weight W, which is $M = WC \sin a$, would cause it to fall. Because the top rotates, however, the spin axis OS sweeps out the surface of a cone as the line QS (which is at right angles to OQ) rotates at an angular speed Ω in a counterclockwise direction, viewed from above, about point Q.

Precession of a top

The right-angled triangle OQS forms a plane that sweeps out the volume of the cone at the speed Ω, which is called the angular velocity of precession of the top.

The angular momentum $I\omega$ of the top is represented by the vector OS, coinciding with the axis of the top and pointing in the direction that a right-handed screw would advance if it were attached to the rotating axis. The vector OS, of magnitude $I\omega$, is changing in direction only, and, because its rate of change must equal the velocity v_S of point S on the circular path, it is also equal to the rate of change of QS, of magnitude $I\omega \sin a$. Then $v_S = QS\Omega = (I\omega \sin a)\Omega$ is the time rate of change of $I\omega$.

As the top rotates about its axis, the axis also rotates, and $I\omega$ is not the total angular momentum of the top; there are two other components of the momentum, and they would have to be considered for a complete description of the motion. If ω is very large compared with Ω, however, a satisfactory explanation of the motion can be obtained by assuming that $I\omega$ is the resultant momentum.

Using the right-hand screw rule, the moment M, of magnitude $WC \sin a$, is a vector acting through O, at right angles to the plane OQS and, therefore, parallel to v_S. Then, equating the moment to the rate of change of momentum, $WC \sin a = (I\omega \sin a)\Omega$, from which $WC = I\omega\Omega$. This result shows that the velocity of precession is the same for all values of a. Because of the attraction of the Sun

and the Moon on the Earth's equatorial bulge, the Earth's axis has a small velocity of precession. It requires 26,000 years for a complete revolution and was first explained by Newton.

Gyroscopes

A gyroscope is a small, accurately made flywheel that is usually supported in encircling rings known as gimbals in such a way that it can rotate freely about any axis. This is equivalent to mounting it at its centre of mass so that the distance C in Figure 19A is zero and the axis of spin remains immovable (theoretically) in space.

In Figure 19B a circular disk of moment of inertia I, rotating at a high angular velocity $\boldsymbol{\omega}$, is supported in bearings at A and B in the U-shaped frame. The frame is rotating about a vertical axis at C with a clockwise (viewed from above) angular velocity $\boldsymbol{\Omega}$, indicated, according to the right-hand rule, by a vector pointing down. This is a forced precession, and the change in the angular momentum vector $I\boldsymbol{\omega}$, like the change in any vector caused by rotation, will be perpendicular to $I\boldsymbol{\omega}$ and pointing into the paper for the given direction of $\boldsymbol{\omega}$. The magnitude of the change in $I\boldsymbol{\omega}$ is $I\omega\Omega$, and the moment $\boldsymbol{M}$ that produced the change must have the same magnitude and direction; it is shown by the vector $\boldsymbol{M}$ in Figure 19B. The additional loads on the bearings, the equal and opposite forces $\boldsymbol{P}$ that create the moment $\boldsymbol{M}$, are obtained from $PL = I\omega\Omega$. Note that the spin, precession, and moment axes are mutually perpendicular and that the spin axis, when precessing, always rotates toward the moment axis.

On some ships the controlled precession of a large gyroscope is used to produce a moment to counteract the rolling of the ship at sea. There are many other applications of the gyroscope.

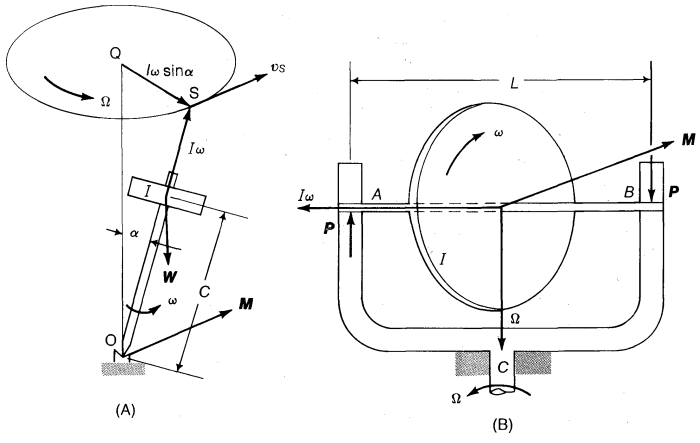


Figure 19: *Gyroscopic action.*
(A) Spinning top, (B) forced precession.

Lagrange's and Hamilton's equations. As the complexity of the problems in classical dynamics increases, it becomes more and more difficult either to formulate the equations of motion or to solve them. In many cases, the forces acting on the particles may be difficult to ascertain, and Newton's second law, which expresses a causal relationship between force and acceleration, may be difficult to formulate. Among the more powerful and more universal methods that have been developed for deriving the equations of motion for these more complicated situations, the equations of Lagrange and Hamilton are the most generally useful. These equations do not represent new theories, but only new ways of looking at dynamics, and the resulting equations of motion are the same as those derived by Newton's law.

Because it emphasizes the vectorial characteristic of the quantities involved, Newtonian dynamics is largely geometric; Lagrangian and Hamiltonian dynamics, on the other hand, require only analytic operations and deal with scalar (nonvectorial) quantities that require no geometric representation.

In Newtonian dynamics force and acceleration are the basic components, and in the coordinate systems only lengths and angles are used to describe the position of particles. In the Lagrange and Hamilton equations, the components are kinetic and potential energy, and the coordinates are generalized; *i.e.*, they are not restricted to lengths and angles but may be any quantities that describe the state of a mechanical system and can be varied independently of one another. In some cases, areas or combinations of lengths and angles may be used. It is common practice to represent the generalized coordinates by $q_1, q_2, q_3, \ldots, q_n$.

Generalized coordinates

Corresponding to each q there is a generalized velocity that is simply the derivative (rate of change) of q with respect to time, namely, dq/dt, and it is written as $\dot{q}$. If q represents a length, then $\dot{q}$ is a real linear velocity, while if q represents an angle, then $\dot{q}$ is an angular velocity. If q is neither a length nor an angle, $\dot{q}$ is not a real velocity but a defined one.

The derivation of Lagrange's equations is a lengthy, abstract mathematical exercise in differential calculus; the basic equations take several forms, depending on the nature of the loads acting on the system. For conservative systems (*i.e.*, those for which the law of conservation of energy is valid), the following equation holds for each generalized coordinate q:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = 0. \qquad (18)$$

The quantity $L = (T - V)$, in which T is the kinetic energy and V the potential energy of the system in generalized coordinates, is called the Lagrangian function. The term $\partial L/\partial \dot{q}$ is the derivative (rate of change) of L with respect to $\dot{q}$ when all other variables in the expression for L are held constant; it is called a partial derivative and is written with ∂ instead of d, to distinguish it from $dq/dt = \dot{q}$, which is called an ordinary derivative.

Partial and ordinary derivatives

When $\partial L/\partial \dot{q}$ has been evaluated, the first term in Lagrange's equation is obtained by taking its derivative with respect to time. The terms $\partial/\partial \dot{q}$ and d/dt can be thought of as operators that are to be applied to L and $\partial L/\partial \dot{q}$, respectively. If the position of a point moving on a straight path is described by its distance s from a fixed point on the path, then ds/dt is the velocity of the point. In general, dy/dx, the derivative of y with respect to x, can be applied to any function of x such as, for example, $y = x^n$, in which n is any number such as 0.5, 1, 2, and so on, but the physical interpretation of the derivative depends on the nature of the function.

There are simple algebraic rules for taking derivatives; for the power function $y = x^n$, the derivative $dy/dx = nx^{n-1}$ so that, if $y = x^2$, then $dy/dx = 2x$. Since dy/dx is also a function of x, the derivative of dy/dx with respect to x, namely, d^2y/dx^2, is equal to 2; this is the second derivative of y with respect to x.

As an example of the application of Lagrange's equation, consider a freely falling body of weight $\boldsymbol{W}$ and mass m at a height y above the ground, and assume that the differential equation of motion is required; *i.e.*, an equation containing the second derivative of y with respect to time—namely, $\ddot{y}$.

The velocity of the body is given by $\dot{y}$ and the kinetic energy T by $\frac{1}{2}m\dot{y}^2$; the potential energy V is $\boldsymbol{W}y$ and $L = T - V = \frac{1}{2}m\dot{y}^2 - \boldsymbol{W}y$. The obvious generalized coordinate is $q = y$, $\partial L/\partial y = -\boldsymbol{W}$, $\partial L/\partial \dot{y} = m\dot{y}$, and $d(m\dot{y})/dt = m\ddot{y}$. Substituting these values in Lagrange's equation gives $m\ddot{y} + \boldsymbol{W} = 0$, from which $\ddot{y} = -\boldsymbol{W}/m$, and since $m = \boldsymbol{W}/\boldsymbol{g}$ this reduces to $\ddot{y} = -\boldsymbol{g}$. This is the differential equation of motion for a falling body, which may be written as $d^2y/dt^2 = -\boldsymbol{g}$ and expresses the acceleration of the body in terms in the second derivative of y with respect to time. The velocity equation, $\boldsymbol{v} = dy/dt$, is the function of time that when differentiated will yield the acceleration. This process is known as integration, and it is the inverse of differentiation. It is easy to see that $\boldsymbol{v} = dy/dt = -\boldsymbol{g}t + \boldsymbol{v}_0$, in which $\boldsymbol{v}_0$ is the velocity of the body when $t = 0$; because $\boldsymbol{v}_0$ is a constant, its derivative is zero. Integrating the velocity equation yields $y = -\frac{1}{2}\boldsymbol{g}t^2 + \boldsymbol{v}_0 t + y_0$, in which y_0 is the position of the body when $t = 0$. This is the complete equation for a falling body, which was obtained previously by more elementary methods; differentiating it twice yields $d^2y/dt^2 = -\boldsymbol{g}$.

When using Lagrange's equations the differential equations of motion are derived in the same way for each generalized coordinate. Because kinetic energy involves velocities only, accelerations do not have to be considered. In addition, the velocity terms are squared so that the kinetic energy is always positive, and difficulties with algebraic signs are avoided.

Degrees of freedom

Lagrange's equations are particularly useful when the number of generalized coordinates is two or more or, in other words, when the number of degrees of freedom—*i.e.,* the ways in which the system can move—is two or more. The differential equations of motion for vibrating systems with multiple masses and springs are most easily obtained by Lagrange's equations.

The equations of Hamilton contain another generalized quantity—namely, generalized momentum $\boldsymbol{p}$—and the Hamiltonian function H. The generalized momentum $\boldsymbol{p}$ associated with the generalized coordinate q is defined as $\boldsymbol{p} = \partial L/\partial \dot{\boldsymbol{q}}$. If q is a length, then $\boldsymbol{p} = m\dot{\boldsymbol{q}}$, which has the dimensions of linear momentum, $m\boldsymbol{v}$; if q is not a length, it does not have this dimension, and the generalized momentum is strictly a mathematical definition. The function H is defined as $H = \Sigma \boldsymbol{p}\dot{\boldsymbol{q}} - L$, in which $\Sigma \boldsymbol{p}\dot{\boldsymbol{q}}$ is the sum of the $\boldsymbol{p}\dot{\boldsymbol{q}}$ terms for all generalized coordinates in the system and L is the Lagrangian function. For each generalized coordinate, Hamilton's canonical equations are

$$\dot{\boldsymbol{q}} = \frac{\partial H}{\partial \boldsymbol{p}} \tag{19}$$

and

$$\dot{\boldsymbol{p}} = \frac{\partial H}{\partial q}. \tag{20}$$

Because Hamilton's equations utilize both generalized coordinates and momenta, whereas Lagrange's equations use only generalized coordinates, there is more freedom in choosing the variable when Hamilton's equations are used.

It should be noted that, although Lagrange's equations are useful for finding the differential equation of motion of a dynamic system, Newton's equations are still unrivaled when the forces acting on the system are required. (A.Co.)

MECHANICS OF NONRIGID BODIES

Impact, or collision

When two moving bodies collide, the magnitude of the forces created at the region of contact and the resulting changes in the velocities of the bodies depend on the shapes of the bodies, their masses, their elastic properties, and their initial velocities. The fundamental theory of impact can be developed by means of the two spheres of the same material in Figure 14D, of masses m_1 and m_2 and moving in a straight line with velocities $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

If $\boldsymbol{v}_1$ is greater than $\boldsymbol{v}_2$, the spheres will collide; and, after collision, $\boldsymbol{v}_2'$, the final velocity of sphere 2, will be greater than $\boldsymbol{v}_1'$, the final velocity of sphere 1.

During the short period of contact the forces on the spheres will be equal and opposite, and, because they act for the same length of time, their impulses will produce equal and opposite changes in the momentum of the spheres. Hence, the law of conservation of momentum applies, and the total momentum after impact equals the total momentum before impact, or

$$m_1\boldsymbol{v}_1' + m_2\boldsymbol{v}_2' = m_1\boldsymbol{v}_1 + m_2\boldsymbol{v}_2. \tag{21}$$

In order to determine the velocities after impact, the energies of the spheres must be considered. At the moment of impact the energy is all kinetic; during impact some of this energy is lost as a result of permanent deformation of the spheres and the generation of vibrations and sound waves. If there are no losses, it can be shown that the velocity difference of approach $(\boldsymbol{v}_1 - \boldsymbol{v}_2)$ is equal to the velocity difference of separation $(\boldsymbol{v}_2' - \boldsymbol{v}_1')$. By experiment, Newton discovered that $(\boldsymbol{v}_2' - \boldsymbol{v}_1')$ is always less than $(\boldsymbol{v}_1 - \boldsymbol{v}_2)$ and that the relation between these velocity differences is given by

$$\boldsymbol{v}_2' - \boldsymbol{v}_1' = e(\boldsymbol{v}_1 - \boldsymbol{v}_2), \tag{22}$$

The coefficient of restitution

in which the numerical factor e, the coefficient of restitution, has values between zero and unity that depend solely on the material of the bodies and not on their masses or their velocities. If $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and e are known, the final velocities $\boldsymbol{v}_1'$ and $\boldsymbol{v}_2'$ can be calculated from equations (21) and (22).

If the spheres consist of lead, putty, or some other soft material, e would be almost zero, and from equation (22), $\boldsymbol{v}_2' = \boldsymbol{v}_1'$, which means that after impact the spheres would move on together with a common speed that can be calculated by equation (21). This phenomenon is known as inelastic impact.

For the special case of equal masses and purely elastic impact—*i.e.,* $e = 1$—it can be shown, by solving equations (21) and (22) simultaneously for $\boldsymbol{v}_1'$ and $\boldsymbol{v}_2'$, that $\boldsymbol{v}_1' = \boldsymbol{v}_2$ and $\boldsymbol{v}_2' = \boldsymbol{v}_1$, which means that the spheres will exchange velocities after impact. If sphere 2 is stationary when sphere 1 strikes it—*i.e.,* if $\boldsymbol{v}_2 = 0$—then $\boldsymbol{v}_1' = 0$, and sphere 1 will stop dead on impact, while sphere 2 will move on with the speed of sphere 1 before impact. This effect can be demonstrated by suspending the spheres on cords, as shown in Figure 14E. If sphere 1 is moved through the angle θ and let go, it will stop dead when it collides with sphere 2, which instantaneously will acquire the velocity of sphere 1 and swing up through an angle slightly less than θ on account of energy losses. Newton is reported to have used this apparatus with cords three metres long to confirm his third law.

If a small ball is dropped on the horizontal surface of a large mass of the same material, the coefficient of restitution can be calculated by comparing the height h to which the ball bounces with the height H from which it is dropped. Assuming that the large mass is stationary before and after impact, then $\boldsymbol{v}_2$ and $\boldsymbol{v}_2'$ are both zero, and from equation (22) $-\boldsymbol{v}_1' = e\boldsymbol{v}_1$, or $e = -\boldsymbol{v}_1'/\boldsymbol{v}_1$.

From equation (16) $\boldsymbol{v}_1 = +\sqrt{2\boldsymbol{g}H}$ in a downward (assumed positive) direction, and $\boldsymbol{v}_1' = -\sqrt{2\boldsymbol{g}H}$ in an upward (assumed negative) direction; then $e = \sqrt{h/H}$.

Mechanics of elastic bodies

There are several reasons—even in a treatment of classical mechanics, which deals largely with so-called rigid bodies—for considering the deformations of these bodies. It has already been pointed out that the number of unknown forces in the members of a machine or structure that can be determined by the equations of rigid-body statics is limited and that the elastic properties of the members must be taken into account if the forces in redundant members or supports are to be found. (A.Co.)

Elastic behaviour

In the broadest sense, a body has elasticity when it has the power of recovery. Thus, a china cup is not considered elastic when it falls to the floor and breaks. A straight strip of tempered steel will spring back from a bent position, even when it has been held so for a long time, whereas a strip of lead will not; the steel is considered to be elastic, the lead is not. The word elasticity was coined in the 17th century when some appreciation was being gained of the fact that all substances are more or less elastic, whether inanimate or part of living organisms, though elastic behaviour terminates rather early with some substances because of their plasticity or their brittleness. Even a gas may be said to spring back to its original volume when it has been compressed, and in 1662 the Anglo-Irish chemist Robert Boyle noted the "elastical power" of air. In 1678 Robert Hooke, an English physicist, maintained that in any "springy body" displacement is proportional to load. Not until much later were stress and strain given precise meanings. This lack of definition, however, did not deter the mathematicians of the 18th century from examining problems such as flexure of beams; the essential steps to-

ward a sound theory of elasticity were taken between 1820 and 1830 by French mathematicians, notably Augustin-Louis Cauchy.

THE PHENOMENON OF ELASTICITY

In 1807 the English physicist and physician Thomas Young proposed an elastic modulus, or a constant, relating unidirectional stress and strain. If a bar of uniform cross-sectional area A is stretched by a force of magnitude T so that the stress (σ) denotes the force per unit area, then stress is expressed by $\sigma = T/A$. Under this stress the bar, of original length L, will stretch by a length indicated by δL to give a total length of $L + \delta L$. Unidirectional strain e is the fractional elongation; *i.e.*, the increase in length divided by the original length so that $e = \delta L/L$. Young's modulus can now be defined as the constant of proportionality between stress and strain: $E = \sigma/e$. It expresses the inherent stiffness of the material under longitudinal stretching regardless of the size of the bar.

Young's modulus

Although Young had some knowledge of numerical modulus values, it was somewhat later that testing machines came into regular use for examining constructional materials. Sensitive extensometers for measuring extensions of the order of $1/100{,}000$ inch (10^{-5} inch) were also needed for accurate determination of elastic behaviour. Notable advances in the techniques of testing materials, particularly in the use of optical extensometers, were made in German research schools from about 1870.

At present, dynamic methods are often used for measuring Young's modulus. A bar of uniform section may be vibrated along its length (in its fundamental longitudinal mode) using a piezoelectric vibrator. The periodic time is the time required for a stress pulse to travel from one end of the bar and back, or twice the bar length L; however, the speed v of the propagation of this stress pulse turns out to be such that $v^2 = E/\rho$. Hence E, the modulus, may be calculated from the equation that relates E to the density ρ and the square of observed resonant frequency, f^2, that is, $E = 4\rho L^2 f^2$.

The idea that torsional stiffness (resistance to twisting) depends on a shear modulus was accepted before the theory of torsion was put on a firm general basis by the French engineer A.J.C. Barré de Saint-Venant in 1853. A cylindrical shell with length L and radius r may be thought of as fixed at one end. If the other end of the shell is twisted through an angle theta (θ), the engineering shear strain phi (φ) is defined as the angle θ multiplied by the ratio of radius r to length L; that is, $\varphi = \theta r/L$. Shear modulus G (also called the rigidity modulus) is the constant that relates shear strain phi (φ) to shear stress tau (τ) in an elastic body: shear stress equals the shear modulus times the shear strain, so that $\tau = G\varphi$. By considering a solid bar of outside radius R to be made up of a series of such shells, it is possible to calculate the total torsional moment M as equal to the product $\pi R^4 G\theta/2L$. Hence, from a graph of measured M and θ values, G can be found.

Shear modulus

Elasticity in viscous and crystalline bodies. Although matter in a condensed state is inherently elastic, by reason of bonding forces between molecules, distortional elasticity can become evident only when relative sliding between molecules is inhibited and the resulting resistance to flow is relatively high. In materials made up of crystals, the atoms within each crystal are tightly packed in a regular array. Other materials, such as glass and viscous liquids, are amorphous (*i.e.*, they consist of randomly assembled atoms). On an atomic scale, flow in an amorphous substance depends on the existence within the material of holes sufficiently large to allow the transfer of an atom from one position to another. Because there is thermal agitation of atoms within all materials at normal temperatures, the frequency with which an atom will gain enough energy to accomplish a transfer may be predicted theoretically by the application of certain equations of the Maxwell-Boltzmann type derived from kinetic molecular theory. Such methods of statistical mechanics have been used to explain the linear relation between stress and strain rate in a viscous liquid (which is then termed a Newtonian liquid) and also the effect of temperature and free volume on the coefficient of viscosity (equal to shearing stress divided by the rate of shear). It is a familiar observation that amorphous materials, such as wax or pitch, will soften progressively when heated but will behave elastically when cold or when subjected to stress applied rapidly.

When a liquid contains a moderate concentration of solid particles in suspension, its apparent viscosity is increased. Many concentrated solutions of organic substances, however, tend to form a gel (jellylike matter having elastic properties) when allowed to rest, through the growth of a network of solute molecules. Gelatinous material, for example, develops a definite yield strength, or resistance to stress, on setting, which enables it to oscillate when it is shaken.

Polymers, substances of high molecular weight made up of a large number of identical (or practically identical) smaller molecules, can exist as fluids or gels, or they may be glassy or crystalline. A polymer like rubber contains long molecules twisted around each other, and the extension by pulling or twisting of such a substance involves straightening or crumpling of molecular chains. Distortion involves viscous shearing as parts of the molecular chains of rubber move relative to each other. On release of stress the chains resume their former convoluted shape and the strain due to elasticity is gradually recovered. Fillers such as carbon modify the open structure of rubber, increasing its elastic modulus. Internal structure obviously has an important effect on the elastic modulus of fibre-reinforced plastics; it is at a maximum when stress is applied in the direction of the fibres and when the fibres are kept straight during molding.

Polymers

A piece of metal generally consists of an aggregate of large numbers of crystals formed during solidification, and these crystals persist through subsequent heating or forging. Although creep (a slow change in shape) can be due partly to movements of atoms around vacant lattice sites, the plastic deformation of metals takes place chiefly through movement of dislocations. These defects in the stacking arrangement of atoms within crystals have the effect of reducing the yield strength to about one-hundredth part of the strength of a perfect crystal. Studies of dislocations, carried out extensively since about 1945, have succeeded in explaining many aspects of the mechanical behaviour of metals.

Plastic deformation of metals

If, for example, a specimen of low-carbon steel is stretched in a testing machine, stress (σ) and strain (e) may be plotted as in Figure 20. The first application of any stress produces strain indicated by the straight line OA, its slope being equal to Young's modulus. An upper yield point A is reached, and plastic extension, as strain is increased, occurs along BC at a slightly lower stress; this reduction is associated with the spread of what are called Lüders bands, lines or markings on the formerly smooth surface, through the specimen. After C, strain hardening occurs along CD. When stress is released at point D, recovery along DE is substantially elastic. Application of compressive stress at point E, however, is not found to give an elastic compression EE′, but instead a curved inelastic line EF is followed. This is called the Bauschinger

Bauschinger effect

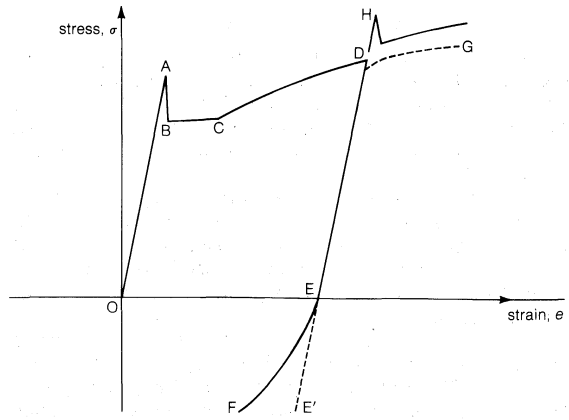


Figure 20: Behaviour of low-carbon steel during stretching in a testing machine (see text).

effect, and it is associated with an easy reverse flow of dislocations that have become piled up against obstacles such as locked groups of dislocations and grain boundaries. An immediate reapplication of tensile stress gives flow DG, as if loading had not been interrupted. If the specimen is rested for a few days, a new yield point H is encountered.

Yield points in steel can be associated with the presence of interstitial atoms of carbon and nitrogen. These diffuse through the iron lattice and, by gathering around dislocations, prevent their movement until an upper yield stress is applied. The numerous free dislocations present in a plastically strained material are thus gradually immobilized. The effect, called Cottrell–Bilby locking, is accelerated by heating to about 200° C (392° F). This treatment restores elasticity and eliminates the Bauschinger effect.

When a piece of steel is subjected to continuous rapid cycles of stress that do not exceed the static yield stress, a thin hysteresis loop develops (a graph of the resistance to a cycle of applied force due to a former state of the system). If the hysteresis is attributed to a limited number of atoms breaking away, it can be understood why the loop is widened by overstraining that provides more free dislocations and why it is reduced in width by heat treatment.

It is well known that a bar of steel emits a ringing sound when it is struck, but a bar of cast iron does not because of the higher internal damping (resistance to oscillations) of cast iron. As an idealization for use in studying inelastic response of materials, a constitutive equation has been proposed that includes rates of stressing and straining with respect to time. Constitutive equations give the relationship between deformation and force (see below *Deformation and flow*).

Properties of strain and stress

Three-dimensional strain and stress. The general properties of strain are concerned with the shape of an object, whereas those of stress derive from equilibrium of forces; therefore, these properties are independent of whether or not the body is elastic. If the position of a particle in the unstrained body is described by coordinate values x_1, x_2, and x_3, and if the amount of distortion is small, the position of this particle can be described with sufficient accuracy by the same coordinate values, even though small displacements have occurred. Then, if these displacements are u_1 in the x_1 direction, u_2 in the x_2 direction, and u_3 in the x_3 direction, normal strain will be equal to fractional elongation. Taking an element of length dx_1 lying along the x_1 axis of the coordinate system in the body, the displacement of the outer end will exceed that of the inner end by the partial derivative of the displacement in the x_1 direction with respect to x_1 multiplied by dx_1: $(\partial u_1/\partial x_1)dx_1$.

This is the rate of change of u_1 with respect to x_1, multiplied by the incremental distance. Hence the component of strain becomes $e_{11} = \partial u_1/\partial x_1$, in which e_{11} represents normal strain along the x_1 coordinate. The other normal strains follow by simply replacing the coordinate x_1 with either coordinate x_2 or x_3 and replacing the displacement u_1 with either displacement u_2 or u_3.

A rectangular element, Figure 21, may be considered next, having sides initially parallel to the x_1 and x_2 axes. Shear strain in the theory of elasticity, according to the mathematical definition, is equal to half the decrease in angle of the corner that is outermost with respect to the positive coordinate axes. (The angle A becomes B under the stress.) Small changes in inclination of the outer sides

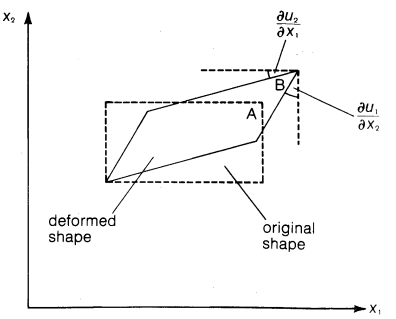


Figure 21: Calculation of shear strain in two dimensions, using angular changes (see text).

can be written in terms of displacement derivatives (Figure 21), leading to an expression for shear strain in which e_{12} represents strain in two dimensions:

$$e_{12} = \frac{1}{2}\left(\frac{\partial u_2}{\partial x_1} + \frac{\partial u_1}{\partial x_2}\right).$$

Strain-displacement equations

Two further components are obtained by replacing symbols x_1, x_2 with x_2, x_3 or x_3, x_1 and by replacing symbols u_1, u_2 with u_2, u_3 or u_3, u_1. These equations of strain displacement give six components of strain in terms of three displacement components, and thus continuity of displacements in a solid body imposes compatibility conditions on any permissible strain distribution.

A stress component in the theory of elasticity is specified by means of two subscripts. Normal stress, σ_{11}, means the stress having a line of action parallel to the x_1 axis and acting on an element of area perpendicular to the x_1 direction. It is a positive value when tensile stress is being measured. Shear stress, σ_{12}, means a stress acting along a line parallel to the x_1 direction on an element of area perpendicular to the x_2 direction. In the x_1, x_2 plane, this stress is positive if it acts in the positive x_1 direction on the side of an element situated at the larger positive x_2 value, or if it acts in the negative x_1 direction on the opposite side of the element. Equilibrium of moments (*e.g.*, forces that can rotate a body) of shear stresses acting on the edges of rectangular elements leads to equivalences between the following shear stresses: $\sigma_{12} = \sigma_{21}$, $\sigma_{32} = \sigma_{23}$, and $\sigma_{13} = \sigma_{31}$, or, in general, $\sigma_{ij} = \sigma_{ji}$ for $i, j = 1, 2, 3$. Hence, there are three independent shear stresses that, with the three normal stresses, make a total of six components acting at any point. These components can be transformed so as to give the components referred to any alternative set of axes. The particular set of three axes on which shear stresses are zero are known as the principal axes.

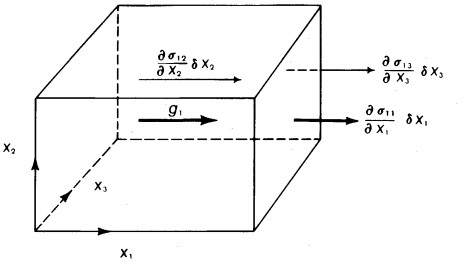


Figure 22: Calculation of shear stresses acting in three dimensions.

Stresses acting on any element must be in equilibrium. If stresses, σ_{11}, acting on opposite sides of the element are equal, they will exert no resultant force. It is assumed that this normal stress may increase across the element so as to have an excess value $(\partial\sigma_{11}/\partial x_1)\delta x_1$ on the outside face (Figure 22), which has an area represented by the product of the two deltas, representing the elements of length, δx_2 δx_3. Similar increases are permitted in shear stresses acting in the x_1 direction, and these excess values of stress are shown in Figure 22. Also, a body-force vector such as might be produced by gravity or inertia is supposed to have component g_1 acting in the x_1 direction, this being a force per unit volume. By equating to zero the sum of all net forces acting in the x_1 direction and removing the common factor equal to the volume of the element, an equation is obtained that may be written:

$$0 = g_1 + \frac{\partial\sigma_{11}}{\partial x_1} + \frac{\partial\sigma_{12}}{\partial x_2} + \frac{\partial\sigma_{13}}{\partial x_3}.$$

Stress equilibrium equations

Two more stress equilibrium equations representing equilibrium in the x_2 and x_3 directions can be obtained by interchange of coordinate values.

These equations define the restrictions to permissible distribution of stress, and they form the starting point in the analysis of stress in any deformable medium.

Linear elasticity and elastic constants. Arguments in favour of the existence of two independent elastic moduli were clarified by George Gabriel Stokes, a British physicist and mathematician, in 1845. He maintained that a

change in volume and a change in shape are physically distinct processes. A bulk modulus K may be defined by the statement that average stress, σ, produces a fractional change in volume. If e is the average linear strain, the law of volume elasticity is represented by an equation in which average stress—*i.e.*, the average of the three stress components—equals three times the product of the bulk modulus and average linear strain: $\sigma = 3Ke$.

Isotropic material

In an isotropic material (one having equal properties in all directions), a single constant can be used for relating reduced stresses and strains because this constant is a measure of rigidity, or of resistance to change of shape. General components of stress and the linear relation can be expressed in an equation that contains the shear modulus. The proportionality between stresses and small strains is calculated without reference to the property of springing back on removal of stress, though this property would need to be assumed if one wished to predict the final strain produced by a succession of stresses having dissimilar signs.

It should also be noted that, although a structure may consist of parts made from linear elastic material, there may be reasons due to its shape why the structure gives a nonlinear variation of deflection with applied load. Examples are struts loaded near the point of elastic instability and thin plates having large transverse deflections imposed.

It is found that Young's modulus for an alloy is not much affected by mechanical and thermal treatments, and it is given approximately by the (volumetric) average of values (of the modulus) for each of the constituent metals. For example, all steels with a content of alloying elements less than about 10 percent can be assumed for most purposes to have the same elastic constants as mild steel. Progressive cold working of metals, however, may lead to varying orientations of the crystallographic axes of individual grains, with development of elastic anisotropy—*i.e.*, of a variation in modulus depending on the direction of the specimen axis relative to the direction of cold working.

The bulk modulus increases with pressure because of a decrease in the space between molecules. Water at 20° C, for example, has a bulk modulus value $K = 320{,}000$ pounds per square inch (psi). According to experimental results, this should be multiplied by a factor containing pressure. (The factor is $[1 + 20 \times 10^{-6}p]$, in which p is pressure in psi.) The increase in K with pressure is in part caused by a decrease in size of the molecules themselves. Because this decrease is relatively small, it might be expected that solids having closely packed molecules would show only a small variation of K with pressure; experiment proves this to be true. For similar reasons, the increase in shear modulus G is small for most metals. (The correction factor here is approximately $[1 + 0.15 \times 10^{-6}p]$.) The elastic moduli of metals usually decrease with temperature due to the atoms moving farther apart, and there is a rapid fall at higher temperatures. This change in modulus may be nonlinear and sometimes irregular, but for many metals a decrease of 2 to 6 percent per 100° C may be expected over moderate ranges of temperature on either side of room temperature.

Molecular packing

MATHEMATICAL THEORY OF ELASTICITY

It has been seen that three sets of equations must be satisfied at all interior points in a linear-elastic body: the strain–displacement equations, the stress–equilibrium equations, and the stress–strain relations. A typical problem posed in the theory of elasticity is to find stresses and displacements at all interior points, given the external shape of the body and the stresses or displacements imposed over the boundary of the body or part of the boundary. Occasionally the basic equations can be used as they stand by inserting a simple system of stresses or displacements into the first-order differential equations.

Various idealized stress systems can be formulated, such as that appropriate for symmetry on planes, or about an axis, or for flexure of plates. The procedure is to look for solutions of the differential equations that can be made to fit the specified boundary conditions. Rectangular coordinates cannot be conveniently used for describing stresses and displacements on curved boundaries. In such problems, the differential equations need to be recast in curvilinear coordinates, which may call for many specialized techniques of applied mathematics.

It is often possible to verify theoretical solutions by means of experimental methods, as, for example, by using photoelastic models or electrical resistance strain gauges. The extensive range of techniques in modern experimental stress analysis can also be used in situations where no theoretical methods can be applied—*e.g.*, in the measurement of internal stresses set up in components during manufacturing processes.

Equations referred to earlier, as well as invariant functions, strain energy, and yield criteria, are treated below.

The constitutive equation of an anelastic solid is expressed

$$\sigma = t_1 \frac{d\sigma}{dt} = E\left(e + t_2 \frac{de}{dt}\right)$$

in which t_1 and t_2 are time constants. For very slow straining, the rate terms can be dropped, leaving the static elasticity relation $\sigma = E_e$. If straining is extremely rapid so that the rate terms predominate, however, the apparent modulus is given by $E' = Et_2/t_1$. Response to periodic stressing at frequency $\omega/2\pi$ can be found by inserting $\sigma = \sigma_0 \sin \omega t$ and $e = e_0 \sin(\omega t - \varphi)$. The angle of lag φ is then obtained as

$$\tan \varphi = \omega(t_2 - t_1)/(1 + \omega^2 t_1 t_2).$$

This predicts a maximum angle of lag at imposed frequency $\omega = 1/\sqrt{t_1 t_2}$, and it is found that many materials do exhibit such a maximum value.

It is usually convenient to use total stresses and strains rather than reduced values. New elastic constants, ν, called Poisson's ratio, and E may be introduced to obtain equations relating normal stress and strain of the type

Poisson's ratio

$$e_{11} = \frac{1}{E}[\sigma_{11} - \nu(\sigma_{22} + \sigma_{33})],$$

provided the new constants are related to the previous ones by

$$E = \frac{9KG}{3K + G} \quad \text{and} \quad \nu = \frac{3K - 2G}{2(3K + G)}.$$

These expressions lead to the useful expressions for Young's modulus $E = 2(1 + \nu)G = 3(1 - 2\nu)K$. To illustrate the significance of Poisson's ratio ν, the axis of a tensile specimen is made to coincide with the z direction, and the longitudinal stress is set equal to $\sigma_{zz} = T$, with transverse stresses $\sigma_{xx} = \sigma_{yy} = 0$. The strains are then $e_{zz} = T/E$ and $e_{xx} = e_{yy} = -\nu T/E$. This means that the fractional transverse contraction is ν times the fractional longitudinal extension.

An invariant function of components of stress or strain at a point is a function having the same numerical value for any choice of coordinate axes. There are three independent invariants of stress. The first is average stress σ, defined as one-third of the sum of the three normal components of stress. The second is conveniently written in terms of reduced stresses in which the reduced values of stress are distinguished by a prime, $\sigma'_{11} = \sigma_{11} - \sigma$, $\sigma'_{12} = \sigma_{12}$, and similarly for the other components;

Invariants, energy, and yield criteria

$$2\bar{\sigma}^2 = \sigma'^2_{11} + \sigma'^2_{22} + \sigma'^2_{33} + 2\sigma^2_{12} + 2\sigma^2_{23} + 2\sigma^2_{31}.$$

It is seen that when only a single shear stress σ_{12} is present, the other components being zero, then $\bar{\sigma} = \sigma_{12}$, so that this

Table 1: Approximate Values of Elastic Constants
(E in units of 10^6 psi)

material	E	ν
Nylon	0.4–0.5	0.4
Wood (along grain)	1.3–1.9	
Concrete	2	
Glass	8–10	0.25
Aluminum alloys	10	0.33
Titanium alloys	17–18	
Steels	29–31	0.29
Carbon fibre composites	20–50	
Alumina (sintered)	50–55	
Tungsten carbide (sintered, 5% cobalt)	85–90	0.21

invariant is sometimes called the equivalent shear stress. The third invariant is essentially a third-order combination of stress components that may be shown to have a one-to-one relationship with Lode's variable, $\mu = 3\sigma'_2/(\sigma_1' - \sigma_3')$. These reduced principal stresses are designated in such a way that $\sigma'_1 > \sigma'_2 > \sigma'_3$. This variable, as defined here, can have values ranging from minus one for uniaxial tensile stress to plus one for compressive stress.

Strain energy U per unit volume is essentially the summation of products of internal forces and displacements, and for any linear material this can be written

$$U = \frac{1}{2}(\sigma_{11}e_{11} + \sigma_{22}e_{22} + \sigma_{33}e_{33} + 2\sigma_{12}e_{12} + 2\sigma_{23}e_{23} + 2\sigma_{31}e_{31}).$$

These stresses and strains may be separated into isotropic and reduced parts so that the basic stress–strain relations for an isotropic solid can be used. As might be expected, the result can be expressed in terms of stress invariants, $U = \frac{1}{2}(\sigma^2/K + \bar{\sigma}^2/G)$. Alternatively, total stresses may be used together with elastic constants E and ν to obtain the strain energy per unit volume,

$$U = \frac{1}{2E}[\sigma_{11}^2 + \sigma_{22}^2 + \sigma_{33}^2 - 2\nu(\sigma_{11}\sigma_{22} + \sigma_{22}\sigma_{33} + \sigma_{33}\sigma_{11}) + 2(1+\nu)(\sigma_{12}^2 + \sigma_{23}^2 + \sigma_{31}^2)].$$

A similar expression can be obtained, if required, in terms of strain components alone.

The stress at which plastic flow commences in crystalline materials is not affected by superposition of moderate values of average stress. This fact has been demonstrated by carrying out tensile and torsion tests with specimens enclosed within a pressure vessel and subjected to high fluid pressures. Hence, the criterion of plastic yielding in an isotropic solid can be written in terms of the remaining stress invariants, in the form $\bar{\sigma}(\mu) = k$, in which k is some fixed numerical value. This function of μ is called a yield locus and can be represented in various geometric ways. For the simplest kind of yield locus, $\bar{\sigma} = k$, which is known as the Mises criterion. An alternative, the Tresca criterion, states that yielding occurs when the maximum shear stress acting within the material reaches a certain value, and, as this shear stress is equal to half the difference between the greatest and least of the principal stresses, this criterion becomes $(\sigma_1 - \sigma_3) = 2k$. (D.S.D.)

Deformation and flow

Deformation is the motion of one part of a body relative to another part so that the body changes size or shape (*i.e.*, a change of distances occurs between points in the body), and flow is a continuous change of deformation with time. The area of mechanics that deals with such motions is called continuum mechanics, or the mechanics of deformable media. Many examples of deformations are observed in everyday experience: the motion of a liquid being poured, of water in a flowing river, of a flag fluttering in the wind, of water boiling, of a rubber band being stretched, and of a violin string when it is plucked.

Continuum mechanics and its applications in technology

Continuum mechanics is widely applied in technology. Structures must be designed and construction materials chosen so that neither fracture nor too much deformation will occur in use. In airplane design, the force of the air on the wings must be calculated for maximum lifting power and efficiency. Paints must be formulated so that they will flow from the brush but will not drip afterward. Pipelines must be designed with appropriate diameters so that fluids can be pumped at optimum rates.

The function of many biological fluids depends on their flow characteristics. The viscosity of blood plasma increases when the organism is diseased. The synovial fluid present in healthy joints has more complicated flow properties than that in rheumatoid joints.

When forces are applied to some materials, they assume a deformed shape in equilibrium and return to their original shapes after the forces are removed. Such materials are called solids. Other materials can maintain an equilibrium shape only when subjected to hydrostatic pressures (the force is perpendicular to the surface). These materials, both liquid and gaseous, are known as fluids. Under other types of forces, they deform indefinitely as long as the forces are applied and do not return to their original form when these forces are removed. Such an irrecoverable deformation is called flow.

An important part of continuum mechanics is the formulation of constitutive equations—*i.e.*, those that give the relationships between force and deformation—that describe the mechanical behaviour of a given material. Rheology is the part of continuum mechanics that deals with constitutive equations. Such equations provide an approximate description of behaviour over a certain range of circumstances. For example, a metal under a light load might be considered rigid; under a heavier load or with more accurate measurement of length, a linearly elastic solid; under a very large load, a plastic solid; and under a small oscillatory motion, a linearly viscoelastic solid. These three types of behaviour will be discussed below.

Rheology and constitutive equations

Two of the basic concepts of continuum mechanics are those of strain—the measure of the amount of deformation—and of stress, in units of force per unit area, as a measure of the contact force (*i.e.*, the force exerted by one part of the body on a neighbouring part).

HISTORY

Fracture, or breaking, was the chief subject in early work on the mechanics of solids. The notebooks of Leonardo da Vinci contain a discussion of the strength of columns and beams and a sketch of a device for measuring the tensile strength of a wire to which a basket was attached, using sand as a variable load. Galileo noted in 1638 that the strength of a bar in simple tension is independent of its length and is proportional to its cross-sectional area; he was, thus, apparently the first to describe the property of a material in terms of a force per unit area.

In 1678 the aforementioned English physicist Robert Hooke concluded, on the basis of a number of different experiments involving a wide variety of materials, that the elongation of elastic materials is proportional to the force applied.

In a mathematical treatment of the shape of a loaded elastic bar, the Swiss mathematician Jakob Bernoulli in 1705 was the first to state that the elastic law should be formulated as a relation between the force per unit area and the ratio of elongation to the original length. Although another Swiss mathematician, Leonhard Euler, was the first to assume (1727) that this relation was linear, the proportionality constant is called Young's modulus after Thomas Young, who introduced it in 1804 in connection with the elongation of an elastic bar. Euler also explicitly set forth the concept of the internal pressure in a fluid some years earlier (1752). Internal tension in solids (force acting perpendicular to the cross section) was recognized and used by the German mathematician and physicist Gottfried Wilhelm Leibniz in 1684 and by Jakob Bernoulli in 1691, while shear stresses (forces acting parallel to the cross sections) were recognized by the French mathematician Antoine Parent in 1713 and explicitly by Charles-Augustin Coulomb in 1773. The general concept that contact forces may be represented by the stress, an idea fundamental to present-day understanding, was introduced by Augustin-Louis Cauchy in 1822, as were the stress tensor (see below), the differential equations representing the balance of force, and the application of the balance of moment of momentum. Balance of force had been recognized by another member of the Bernoulli family, Daniel Bernoulli, in 1727, and both balance laws by Euler in 1775.

Introduction of stress tensor

Euler in 1775 proposed the first constitutive equation, that for a perfect fluid, based on the assumption that the force exerted by one part of the fluid on the neighbouring part is always perpendicular to the surface separating the two parts. The linearly viscous fluid is sometimes called the Newtonian fluid because Newton proposed a simple case of such a material in his *Principia* (1687). The equations governing the motion of this fluid are called the Navier–Stokes equations after the French engineer Claude-Louis-Marie Navier (1821) and the British physicist George Gabriel Stokes (1845). Two Frenchmen,

Siméon-Denis Poisson (1830) and Barré de Saint-Venant (1843), also made important contributions to the theory. The rate of efflux through narrow tubes, calculated on the basis of the Navier–Stokes equations, was found to be in good agreement with the experimental results on water obtained during the next decade or so.

On the basis of a molecular hypothesis, Navier arrived in 1827 at a theory of elasticity of isotropic solids (solids in which elasticity is uniform in all directions) that contained only one elastic constant. Using some phenomenological concepts, Cauchy (1823), Stokes (1845), and their contemporaries produced the modern theory that involves two elastic constants, the shear modulus and bulk modulus, to be discussed later. Careful experiments confirming the modern theory were performed shortly thereafter.

In 1835 Wilhelm Weber, a German physicist, found that a load suddenly applied to a silk thread produced an immediate extension that was followed by a further lengthening with time, a behaviour called creep. Further experimental investigations of this time-dependent behaviour were performed; and in 1874 Ludwig Boltzmann, an Austrian physicist, formulated the theory of linear viscoelasticity to explain the phenomenon. That some fluids show a similar viscoelastic behaviour was exhibited on fluids of high consistency (hot glass, pitch) in 1905 and on more mobile fluids (gelatin solutions) in 1915.

Theory of linear visco-elasticity

Experiments around the beginning of the 20th century led to the conclusion that colloidal solutions (solutions in which one component consists of particles invisible under ordinary microscopes but thousands of times larger than ordinary molecules) do not act like Newtonian fluids in steady flow, but that their viscosity depends on the rate of deformation. The method of calculating this viscosity function from efflux experiments in the 1920s greatly advanced this field of study.

The transition from the empirical approach to the development of appropriate constitutive equations has gone through several stages. Nonlinear theories proposed in the 1940s and '50s have been superseded by theories that assume that stress depends on the history of the strain that has been applied but not in a simple linear relationship. These theories explain not only nonconstant viscosity in steady flow but also time-dependent behaviour and peculiar distributions of tractions perpendicular to the bounding surfaces (the so-called normal stress effects observed in many situations). Experimental work on these effects began during World War II and is still continuing.

CONTINUUM MECHANICS

In continuum mechanics, which is used to give theoretical insight into deformation and flow, kinematics deals mainly with motions that are deformations, and dynamics, with the stress (*i.e.,* with the forces exerted on one part of the body by a neighbouring part). The detailed discrete nature of the atomic and molecular structure is ignored; it is assumed that the net effect of the atomic and molecular forces and motions can be adequately represented by the stress, the strain or rate of strain, and an appropriate constitutive relation. A thorough understanding of continuum mechanics requires the use of mathematics, particularly vector and tensor analysis. In the sections below, the basic concepts are first presented without the application of mathematics, followed by a more quantitative discussion, though the use of vectors and tensors is restricted to an elementary level.

Kinematics. Before presenting the general notion of strain, some simple deformations that are commonly discussed in many branches of continuum mechanics will be defined.

1. In uniform expansion, all volume elements of the body are transformed to geometrically similar elements of greater dimensions. Such a deformation occurs when a body is uniformly deformed in all directions so that any arbitrary particle originally having the coordinates (x_1, x_2, x_3) moves to a point whose coordinates are (ax_1, ax_2, ax_3), in which a is a multiplicative factor (Figure 23). The displacement of the particle may be represented by the vector $\boldsymbol{u}$ whose components (u_1, u_2, u_3) are $[(a-1)x_i]$, $i = 1, 2, 3$. Any physical quantity that has direction, such as displacement, velocity, force, or electric field, can be represented in magnitude and direction by a vector. The vector $\boldsymbol{u}$, for instance, can be represented by an arrow that points in the direction of the motion of the particle with the length representing the displacement or distance moved. Each component is the length of the axis intersected by lines drawn perpendicular to it from the head and tail of the vector. For example, the tail intersects the X_1-axis in the Cartesian coordinate system (X_1, X_2, and X_3 coordinates mutually perpendicular) at x_1 and the head intersects at ax_1, so that the displacement along the X_1-axis is a component u_1 of length $ax_1 - x_1$, or $(a-1)x_1$.

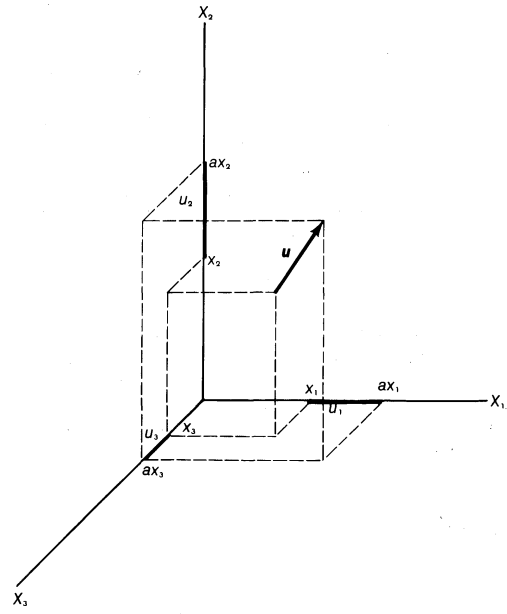


Figure 23: The vector $\boldsymbol{u}$ and its components in a Cartesian coordinate system X_1, X_2, X_3 (see text).

2. Simple shear is usually envisaged as taking place in a material between large parallel plates and may be thought of as sheets of material sliding parallel to one another. Simple shear can be described more precisely by taking a Cartesian coordinate system with the origin O on one of the plates, say the fixed plate, the X_2-axis perpendicular to the plates and the X_1-axis parallel to the direction of motion (Figure 24). In simple shear, an arbitrary particle in the plane ABCD moves in the X_1-direction to a point in the plane ABC′D′ by an amount proportional to its distance x_2 from the fixed plate, sx_2, in which s is a proportionality factor; *i.e.,* it undergoes a vector displacement with components (sx_2, 0,0). In simple shearing flow, each particle of the material moves with a constant velocity v in the X_1-direction, the direction of motion of the sliding

Simple shear

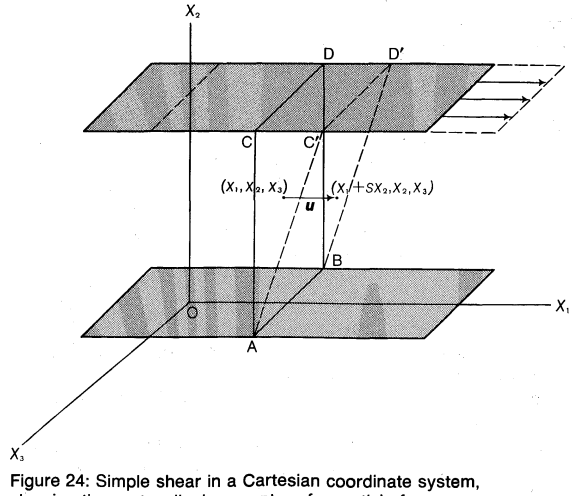


Figure 24: Simple shear in a Cartesian coordinate system, showing the vector displacement $\boldsymbol{u}$ of a particle from x_1, x_2, x_3 to $x_1 + sx_2$, x_2, x_3.

plate. The components of the velocity are $(\kappa x_2, 0,0)$ in which κ is a constant called the rate of shear.

3. General infinitesimal strain is involved in more complicated deformations where the strain is more difficult to specify. The components (u_1, u_2, u_3) represent the displacement $\boldsymbol{u}$ of a particle in the body in the deformed state from its position (x_1, x_2, x_3) in the undeformed state. If the spatial derivatives (a spatial derivative is the change of a component with distance along any one of the three axes) of the displacement are small—strictly speaking, infinitesimally small—the complete description of the strain at a given point in the body may be given in terms of the six new components made up of these spatial derivatives:

Components of strain tensor

$$e_{11} = \partial u_1/\partial x_1;$$
$$e_{22} = \partial u_2/\partial x_2;$$
$$e_{33} = \partial u_3/\partial x_3;$$
$$e_{21} = e_{12} = \tfrac{1}{2}(\partial u_1/\partial x_2 + \partial u_2/\partial x_1);$$
$$e_{13} = e_{31} = \tfrac{1}{2}(\partial u_1/\partial x_3 + \partial u_3/\partial x_1);$$
$$e_{23} = e_{32} = \tfrac{1}{2}(\partial u_2/\partial x_3 + \partial u_3/\partial x_2);$$

or, in general,

$$e_{ij} = e_{ji} = \tfrac{1}{2}(\partial u_i/\partial x_j + \partial u_j/\partial x_i),$$

for $i = 1, 2, 3$ and $j = 1, 2, 3$. The components e_{11}, e_{22}, e_{33} in the above equations represent relative, but not actual, elongations parallel to the X_1-, X_2-, and X_3-axes, respectively; the other components, called shear components, are measures of changes of angle. It is of great importance to the theory that these components are the Cartesian components of a tensor called the infinitesimal strain tensor, or sometimes simply the strain. As a result, the powerful theorems of tensors, a well-developed area of mathematics dealing with such objects, may be exploited in continuum mechanics.

Applying the foregoing definitions to the components (u_1, u_2, u_3) of the vector $\boldsymbol{u}$ for uniform expansion, one finds, if $(a - 1)$ is small, that only three components are not zero (namely, $e_{11} = e_{22} = e_{33} = a - 1$, or $e_{ii} = a - 1$ for $i = 1, 2, 3$) and that the shear components are all zero. When the spatial derivative equations are applied to components $(sx_2, 0,0)$ of the vector $\boldsymbol{u}$ for simple shear with $s << 1$ (s vanishingly small), all components of strain vanish except $e_{12} = e_{21}$, which is equal to $s/2$.

Finite strains

4. For finite strains, the infinitesimal strain is no longer a valid measure of deformation (for example, a zero value does not necessarily mean that the material is undeformed). Many kinds of tensor relations are in current use, all of which include terms that are quadratic (squared) in the spatial derivatives of the displacement.

5. The flow of fluids is discussed in terms of the symmetric rate of deformation tensor **D**, with components defined as $D_{ij} = \frac{1}{2}(\partial v_i/\partial x_j + \partial v_j/\partial x_i) = D_{ji}$, for $i = 1, 2, 3$ and $j = 1, 2, 3$—that is $D_{11} = \partial v_1/\partial x_1$; $D_{22} = \partial v_2/\partial x_2$; $D_{33} = \partial v_3/\partial x_3$; $D_{12} = D_{21} = \frac{1}{2}(\partial v_1/\partial x_2 + \partial v_2/\partial x_1)$; $D_{13} = D_{31} = \frac{1}{2}(\partial v_3/\partial x_1 + \partial v_1/\partial x_3)$; $D_{23} = D_{32} = \frac{1}{2}(\partial v_2/\partial x_3 + \partial v_3\partial x_2)$ in which v_1, v_2, and v_3 are the components of the velocity of an arbitrary particle. Applying these equations for the components of **D** to the simple shearing flow mentioned earlier, in which $v_1 = \kappa x_2$, the only ones that do not vanish are the components $D_{21} = D_{12} = \kappa/2$.

Dynamics. In continuum mechanics it is assumed that the net effect of all forces between atoms may be considered the sum of two parts: (1) body forces due to gravity, magnetism, and electricity; and (2) contact forces exerted on one part of the body by a neighbouring portion. In Figure 25, B is a body that is deformed and A is an arbitrary surface dividing B into two parts, with dA representing an infinitesimal portion of this area at the point Q. It is assumed that the force $d\boldsymbol{F}$ exerted by the matter on one side of the surface on that on the other is given by the equation $d\boldsymbol{F} = \boldsymbol{T}dA$, in which $\boldsymbol{T}$ is known as the stress vector. The existence of such a stress vector is one of the basic assumptions of continuum mechanics. Generally, $\boldsymbol{T}$ depends on both the position in the body and the orientation of the surface dA at that point. The stress vector may be resolved into the sum of a normal (in this sense, perpendicular) stress vector that is directed at right angles to that surface and a shear stress vector that is parallel to that surface.

Stress vector

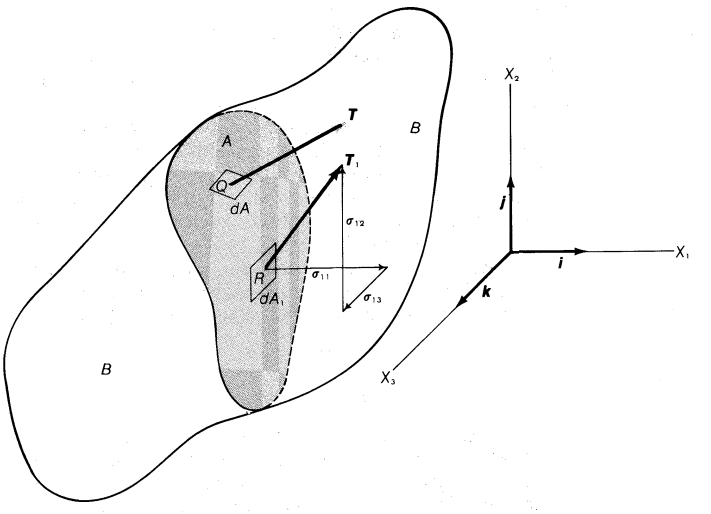


Figure 25: Stress vector $\boldsymbol{T}$ acting on an arbitrary surface A that divides the deformed body B into two parts. A vector component $\boldsymbol{T}_1$ can be decomposed along three coordinate directions, X_1, X_2, and X_3.

It is common to consider the stress vector across surfaces parallel to the orthogonal planes of a chosen Cartesian coordinate system (Figure 25). A plane normal or perpendicular to the X_1-direction is called an X_1-plane. At a given point R within such a plane, the stress vector $\boldsymbol{T}_1$ (not necessarily acting in the X_1-direction) may be decomposed into components $(\sigma_{11}, \sigma_{12}, \sigma_{13})$, one in each of the coordinate directions; that is, the stress vector $\boldsymbol{T}_1$ can be represented as the sum $\boldsymbol{i}\sigma_{11} + \boldsymbol{j}\sigma_{12} + \boldsymbol{k}\sigma_{13}$, in which $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are unit vectors (vectors of given directions and unit length) in the three coordinate directions. The first number in the subscript indicates the direction of the normal or perpendicular to the plane across which the contact force acts, and the second number indicates the direction of the component of the force. A similar set of three components $(\sigma_{21}, \sigma_{22}, \sigma_{23})$ may be obtained from the stress vector $\boldsymbol{T}_2$, which acts across the X_2-plane through R, and $(\sigma_{31}, \sigma_{32}, \sigma_{33})$ from the stress vector $\boldsymbol{T}_3$, which acts through the X_3-plane. The stress components with repeated index (*i.e.*, σ_{11}, σ_{22}, σ_{33}) are the normal stress components; those with different indices (say, σ_{12}) are the shear stresses, or σ_{ij} for $i = j$ represent normal stress components and for $i \neq j$ represent shear stresses.

It can be shown that the set of these nine components

$$\begin{matrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{matrix}$$

constitutes the components of a tensor, called the stress tensor, or simply the stress σ_{ij} for $i = 1, 2, 3$, $j = 1, 2, 3$, at the point with reference to the chosen Cartesian coordinate system. The stress tensor gives a complete characterization of the stress at a point. It can also be shown, as a consequence of the laws of mechanics, that the stress tensor is symmetric (*i.e.*, $\sigma_{12} = \sigma_{21}$, $\sigma_{13} = \sigma_{31}$, and $\sigma_{23} = \sigma_{32}$, or $\sigma_{ij} = \sigma_{ji}$, for $i = 1, 2, 3$, $j = 1, 2, 3$). Thus, only six components are required to specify completely the stress tensor.

Stress tensor

CONSTITUTIVE EQUATIONS

Perfect fluid. In many cases, the motion of fluids such as air and water can be described to a good approximation by the equations of the perfect fluid. The concept of perfect fluid, therefore, forms the basis of much of aerodynamics, hydrodynamics, and fluid mechanics.

By definition, no shear stresses can exist in the model of perfect fluid. As a result, the stress components are given by

$$\sigma_{ii} = \sigma_{11} = \sigma_{22} = \sigma_{33} = -p;$$

and, for $i \neq j$,

$$\sigma_{ij} = \sigma_{12} = \sigma_{13} = \sigma_{23} = 0,$$

Isotropic stress

which are independent of the orientation of the Cartesian coordinate system. Such a stress is called an isotropic stress, and p, in the equation, is the hydrostatic pressure.

Linearly viscous fluid. The model of the linearly viscous (also called a Newtonian or a viscous) fluid describes the flow behaviour of fluids consisting of small molecules under all conditions in which the concept of a perfect fluid is not adequate (*e.g.*, near solid walls) except under extreme circumstances (*e.g.*, ultrasonic waves). It also describes the behaviour of more complex fluids, such as molten plastics at very low rates of deformation.

If a linearly viscous fluid is undergoing simple shearing flow (see above *Kinematics*), the components of the stress with respect to the coordinate system defined above are

$$\sigma_{ii} = \sigma_{11} = \sigma_{22} = \sigma_{33} = -p$$
$$\sigma_{12} = \eta\kappa$$
$$\sigma_{13} = \sigma_{23} = 0,$$

in which η is a property known as the viscosity.

The general equation is based on the assumption that the stress is isotropic when the fluid is at rest and is a linear function of the velocity gradients, or the partial derivatives of the velocity components with respect to the coordinates $\partial v_1/\partial x_1$, $\partial v_1/\partial x_2$, etc., when it is in motion. Expressing these concepts mathematically in such a way that the constitutive equation is independent of the observer, it is found that the stress components may be expressed, for incompressible fluids, in terms of the rate of deformation tensor, D_{ij}, as

$$\sigma_{11} = -p + 2\eta D_{11},$$
$$\sigma_{12} = 2\eta D_{12},$$

for an arbitrary motion and arbitrary choice of Cartesian coordinate systems. The additional equations for σ_{22} and σ_{33} are analogous to that for σ_{11} and those for σ_{13} and σ_{23} to that for σ_{12}. The flow of a viscous fluid is an energy-dissipative process.

The viscosity of a fluid is commonly determined by measuring the rate of flow through a tube of small diameter when a pressure difference is applied. In the metric system the unit of η is called poise in honour of a French physiologist, J.-L.-M. Poiseuille. In the neighbourhood of 20° C, the viscosity of water is 0.01 poise, of air 1.8×10^{-4} poise, of mercury 0.016 poise, of pure glycerin 15 poise.

Linearly elastic solid. The linearly elastic (also called Hookean) solid provides a good model for the behaviour of many solids; *e.g.*, metals, salts, organic crystals, glasses, and rubber at very small strains under static conditions. The model becomes inadequate for some of these materials at elevated temperatures or for some dynamic experiments where linear viscoelasticity may prove a better description.

If an isotropic linearly elastic solid undergoes simple shear, the only nonzero components of the stress are $\sigma_{12} = \sigma_{21} = \mu$, where μ is a property of the material known as the shear modulus, or the rigidity.

Rigidity

If a solid obeying Hooke's law is subjected to a small uniform expansion [*i.e.*, $(a-1) << 1$] the stress is isotropic:

$$\sigma_{11} = \sigma_{22} = \sigma_{33} = -p = 3K(a-1),$$

in which $3(a-1)$ is the fractional change of volume and K is known as the bulk modulus.

The general constitutive relation between stress and strain for the isotropic linearly elastic solid is the set of equations of the form

$$\sigma_{11} = \tfrac{1}{2}\mu e_{11} + (K - \tfrac{2}{3}\mu)(e_{11} + e_{22} + e_{33})$$

for the normal stresses and of the form $\sigma_{12} = {}^1/_2\mu e_{12}$ for the shear stresses. Thus, two constants, the shear modulus μ and the bulk modulus K, specify the properties of an isotropic linearly elastic solid. If the material is anisotropic (not uniform in all directions), as, for example, a crystal, up to 21 constants may be required to specify its properties.

When a stress is applied to an elastic solid, it deforms; when the stress is removed, the solid returns to its original shape. The deformation of an elastic solid is an energy-storing process.

The shear modulus μ may be determined by measuring the torque required to twist a cylinder of the material. Sometimes an elastic modulus may be found from the speed of propagation of sound or from the natural frequency of oscillation of a mechanical system containing the material. The constants μ and K for steel have values of 8×10^{11} and 2×10^{12}, respectively; for lead 6×10^{10} and 4×10^{11}; for rubber 7×10^{6} and 2×10^{10}; all in units of dynes/cm^2.

Linear viscoelastic materials. This model provides a good description for the time-dependent behaviour of many materials under low stresses: for example, attenuation of ultrasonic waves in gases and liquids; oscillations, creep, and stress relaxation in metals, glass, rubber, plastics, and fluids containing long molecules (*e.g.*, molten plastics and protein solutions).

If a shear stress T_0 is suddenly applied to a viscoelastic material in simple shear and kept constant thereafter, the deformation does not increase and attain its steady-state (unchanging with time) condition at once. This type of behaviour is known as creep. The variation of the shear strain $s(t)$ with time t is determined by the property creep compliance $J(t)$ and is represented by the equation $s(t) = T_0 J(t)$. For a viscoelastic solid, the steady state is a constant strain; for the fluid, it is a constant rate of strain.

Creep

If a shear strain S_0 is suddenly imposed on a material in simple shear and kept constant thereafter, the shear stress $T(t)$ initially is very large and then decreases to a constant value, called a steady-state value, which is zero for a fluid. Such an experiment is known as a stress–relaxation experiment. This decay of stress is determined by a property of the material, the relaxation modulus $G(t)$, according to the equation $T(t) = S_0 G(t)$.

The special feature of viscoelasticity is that the stress at a given time depends on all previous values of the strain, with more recent strains having greater influence.

Finite elasticity. There are materials, notably various types of rubber, that can undergo very large strains and return to their original shape after the force is removed. In the theory of isotropic nonlinear elasticity, which provides a good description of the equilibrium behaviour of these materials, the stress is a nonlinear function of the deformation.

Effects arise in nonlinear theories that do not occur in linear ones. According to linear elasticity theory, torsion of a circular cylinder requires only the application of a torque. The theory of finite elasticity demands, in addition, the imposition of normal forces on the plane ends of the cylinder; otherwise the cylinder will change in length.

Nonlinear viscoelastic fluids. High polymers, when in solution or a melted state, show many types of deviations from the behaviour predicted by the theory of Newtonian fluids. They exhibit time-dependent behaviour—for example, creep. In steady flow, the rate of efflux from a tube of small diameter is not proportional to the pressure drop; it appears that the viscosity decreases at the higher rates of flow. If such a fluid undergoes steady shear in the annulus between a stationary plate and a cone rotating at constant angular speed, a greater normal force (per unit area) is exerted by the fluid on the plate near the axis of rotation than at the outside. Based on the assumption that the stress is a nonlinear function of the previous history of the deformation, the theory of the nonlinear viscoelastic fluid is able to explain all of these phenomena.

Fluids exhibiting these phenomena show other technologically important deviations from Newtonian behaviour. The diameter of a stream of fluid may be considerably larger than the orifice from which it was extruded. A rotor in such a fluid may cause a completely different flow pattern from that seen in a Newtonian fluid.

Plasticity. When a rod composed of some solids, such as aluminum and certain steels at ordinary temperature, is subjected to increasing tension, it behaves as a linear elastic solid. As extension increases, the force–elongation curve, as shown in Figure 26, deviates from proportionality at a point Y when the tensile force exceeds a value, f_Y. The ratio f_Y/A_0, in which A_0 is the original cross section, is called the yield stress. The curve then goes through a maximum at M before the sample finally breaks at point B. When the rod is extended beyond Y, some irrecoverable

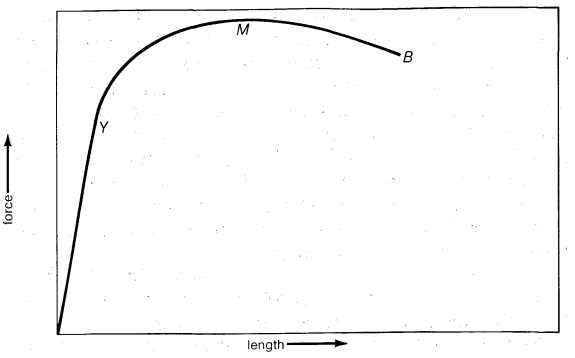


Figure 26: Force–elongation curve.

Plastic flow

deformation, called plastic flow, takes place; the rod does not return to its original length when the force is removed.

One of the most important considerations of the engineer is that structures must be designed and materials of construction chosen so that plastic deformation does not occur to any appreciable extent. On the other hand, many fabrication processes (for example, machining, forging, and drawing) depend on the behaviour in the plastic regime.

FRACTURE

It is customary to classify fracture of solids into two types: brittle and ductile. Generally speaking, brittle fracture occurs when the sample behaves as a Hookean elastic solid until a critical load is reached, at which time it breaks. In ductile fracture, plastic flow precedes fracture. A given material may change its mode of fracture with varying temperature or conditions of the deformation. Thus, many metals and polymers have a brittle–ductile transition temperature that is a function of the rate of straining.

Tensile strength

If an increasing tensile force is applied to a rod of material until it breaks in a brittle manner, the measure of the strength of the material is the tensile strength, defined as ratio of the force at break to the original cross section of the rod. Glass, cast iron, stone, and plastics under ordinary conditions exhibit this type of fracture. If the fracture took place by the simultaneous rupture of all molecular bonds in the surface of separation, the tensile strength would be 10 or 100 times greater than the generally observed values for the usual brittle materials. The discrepancy is due to microscopic sharp long cracks that exist in the material.

Fracture can occur in a tensile experiment even if the stress always remains below the tensile strength. If the metal is subjected to alternating cycles of imposition and removal of stress, it may break if the number of cycles exceeds a critical number. Fatigue is the term used for this type of failure.

When a bar of some transparent glassy plastic (*e.g.*, polystyrene or polymethyl methacrylate) is subjected to a tensile force, microscopic, highly reflective, planar defects form in planes normal (at right angles) to the applied tension. These defects are called crazes, which consist of microscopic voids and polymeric material oriented in the direction of the applied force.

MOLECULAR THEORIES

Gases. Newton's law of viscous flow has been derived from the kinetic theory of gases. If the gas is dilute and the molecules are considered to be rigid spheres that do not attract one another, the viscosity is found to be independent of pressure. A more detailed kinetic theory calculation gives the viscosity in terms of forces between molecules. Viscosity data are now used to estimate the forces between molecules. A rigorous kinetic theory developed in 1956 shows that a rarefied gas is not strictly a Newtonian fluid; normal stress effects are to be expected.

At ultrasonic frequencies of the order of 10^6 cycles per second, viscoelastic effects are indicated by changes in the velocity of propagation of the sound waves and the attenuation of their amplitude. These effects have been attributed to distortions in the distribution of velocities of the gas molecules and in the distribution of molecules among vibrational and rotational states.

Crystalline solids. Viscoelastic effects in crystalline materials usually have been attributed to imperfections in the crystal, such as the motion of interstitial solute atoms or the slip of one crystal grain past a neighbour along a grain boundary. Plastic behaviour is understood in terms of the motion of dislocations.

Liquids. On the molecular level, the liquid is pictured as having neighbouring molecules in contact, as does the crystalline solid, but not packed together in a systematic order; hence there are many "holes" in the structure. In the liquid at rest, a molecule can move from its location into a neighbouring hole over an energy barrier if its thermal energy is high enough and the hole large enough. A shear stress induces a preferred direction for such motions and thus leads to flow.

Polymer behaviour

Polymer molecules in the liquid state (as a melt, rubber, or solute) are loose coils that can change their conformations readily; a barrel full of long, excited snakes has been suggested as a useful analogy. Because the molecules must move through tortuous paths to become displaced relative to one another, the viscosity is high and sensitive to the length of the chain. Viscoelastic effects occur because it takes time for the long entwined molecules to adjust to the applied force. If the temperature is lowered to the point at which large enough holes are rare, the motion becomes so restricted that the polymer acts like a glassy solid.

If a reaction is performed on a polymer (as in the vulcanization of rubber) such that the polymer chains are connected by covalent bonds and therefore cannot move past one another, flow is impossible. Because the polymer chains on the average have a coiled conformation, however, great extensibility of a rubber sample is possible since the molecule can, with little change of internal energy, assume a greatly extended conformation. The deformation of rubber chiefly involves a change of entropy (a change to a more ordered state of about the same energy), while the deformation of crystals and glassy solids entails a change of internal energy. (He.M.)

Tides and related phenomena

Sea-level oscillations of approximately daily, or diurnal, and twice daily, or semidiurnal, period are a worldwide phenomenon observed at continental coasts and islands. According to Newton, this motion is caused by the difference of the gravitational attraction between celestial bodies and the centrifugal acceleration of their rotation. These differences of accelerations are called tide-generating accelerations, and the motion itself is called tides.

Because these tide-generating accelerations act on the particles in the atmosphere and in the Earth as well as on the water masses of the ocean, there are atmospheric tides, Earth tides, and ocean tides. Tidal oscillations are periodic because they are related to the motion of the Earth, the Sun, and the Moon. The tides on the Earth are not influenced by other celestial bodies. It is customary in the English language to refer sometimes to certain phenomena that do not arise from tide-generating forces as tides—for instance, meteorological tides and tsunamis (great sea waves produced by submarine earthquakes or volcanic eruptions).

For centuries efforts have been directed toward understanding the physics and hydrodynamics of the tides in the sea. Amplitudes of tidal oscillations are largest there; atmospheric and Earth tides are less significant. At most coastal or island locations the interval between successive high waters is about 12 hours and 25 minutes, which is half the time of the Moon's apparent revolution around the Earth. In some areas, such as the China Sea, the interval is approximately 24 hours. The first are called semidiurnal tides, the second diurnal tides. If there is a prevailing semidiurnal tide, the difference between the high-water time and the time the Moon crosses the meridian of the location is approximately constant, and this difference is known as the high-water interval for the particular location.

In addition to semidiurnal and diurnal tides, there are also long-period tides. Periods of 14 days to 19 years exist as a result of the motion of the Sun, Moon, and Earth.

A fourth group of tides, which are not of direct astronomic origin, consists of the shallow-water tides that result from the interaction of the semidiurnal and diurnal and other tides.

THE CAUSES OF TIDES

Tide-generating forces

The gravitational attraction of two bodies such as the Earth and the Moon is proportional to the product of their masses and inversely proportional to the square of the distance between their centres. Their gravitational attraction tends to force them to approach one another, but this is inhibited by the centrifugal force of rotation around their joint centre of gravity. If it is assumed that the mass of each body is concentrated in its centre, then the two forces balance each other. There is an important difference between these two forces, gravitational attraction and the centrifugal force of rotation. The first depends on distance and consequently varies for different particles on the Earth; the second is constant for all points of the Earth. If a single water particle of the ocean is considered, there is no balance; the gravitational attraction of the Moon may be different from the centrifugal force of rotation.

These differences are the tide-generating forces. In the same way, tide-generating forces are caused by the gravitational attraction of the Sun; they are determined by the motions of the Sun, the Earth and the Moon, the distance, the mass, and the volume. The tide-generating forces are vectors; in short, they are physical quantities that have both magnitude and direction. One component is directed to the centre of the Earth and coincides in direction with the gravitational force of the Earth itself, altering this only in a very small amount (the ratio is 1 : 9,000,000). The other components are situated on a tangential plane at the sea surface and drive the water masses in the direction of the Moon and Sun, respectively.

Figure 27 shows the distribution of the tide-generating forces on the Earth. The lengths of the arrows are proportional to the magnitudes of the forces. M is the direction of the Moon; the water masses tend to build a deviation toward the Moon (Point V), as well as in the opposite direction (Point I). At point D there is a lowering of the sea level. A deformation of the sea level is similarly caused by the tide-generating forces of the Sun. The tide potential V of the Moon at the sea surface is expressed as follows:

$$V = 26{,}206\left[\left(\frac{c}{r}\right)^3\left(\cos 2\delta + \frac{1}{3}\right) + 0.002598\left(\frac{c}{r}\right)^3(5\cos 3\delta + 3\cos\delta)\right]\frac{\text{cm}^2}{\text{sec}^2},$$

in which c is the average distance between the centre of the Moon and the point on the Earth's surface; r is the distance from the centre of the Moon to the centre of the Earth; and δ is the angle between the line joining those centres and the line from the Earth's centre to the point on the planet's surface under consideration. A similar equation is valid for the tide potential of the Sun.

The differential quotients of the tide potential yield the tide-generating acceleration. The dimension of the potential V is centimetres squared per second squared. If V is divided by the acceleration of gravity (which has the dimension centimetres per second squared), a term remains, ξ, which has the dimension centimetres. The surfaces of equal acceleration of gravity are disturbed by this value ξ. Figure 27 indicates a maximum elevation at points V and I and a maximum depression at point D. This shape of the sea surface would occur in an ocean covering the whole Earth, if the planet were completely rigid and the tide potential remained constant. This type of tide, called equilibrium tide, is fictitious, but it accurately represents tide-generating forces at the surface of and within the Earth. In such a case the slope of the surface of the sea would be balanced by the tide-generating acceleration, and there would be no tidal currents.

The equilibrium tide

In investigating the dynamics of the tides, the tide-generating acceleration may be replaced by the acceleration caused by the slope of the equilibrium tide.

The maximum elevation of the equilibrium tide of the Moon is 35.6 centimetres (14 inches), and the greatest depression would be 17.8 centimetres (7 inches). The natural tidal range in the sea is quite different from these values of the equilibrium tide; accordingly, the conclusion that there are no tidal currents in the sea is completely unrealistic.

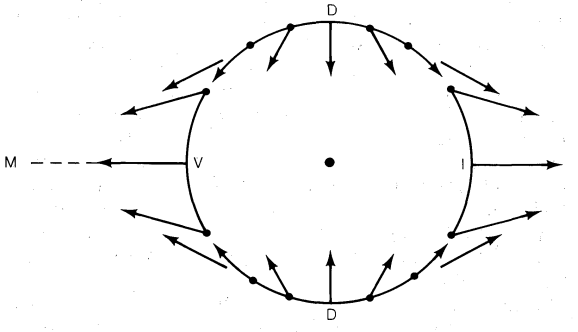


Figure 27: Tide-generating forces (see text).

TIDAL ANALYSIS

Tidal harmonics. The tide potential of the Moon and the Sun are functions of the following variables: the Earth's speed of rotation, γ; the mean motion of the Moon, σ; the mean motion of the Sun, η; the speed of revolution of the Moon's nodes, N; the mean motion of the Moon at its closest approach—the lunar perigee, ω; and the time. These speeds are incommensurable (*i.e.*, they are not divisible, in whole numbers, by each other), and they may serve as independent variables in representing the tide potential by means of a Fourier series (a mathematical series of terms consisting of the trigonometric functions, sine and cosine), each term in which is called a tide or a tidal constituent; the most important are given in Table 2.

Use of the Fourier series to represent tide potential

The coefficient is proportional to the equilibrium amplitude within the given species and indicates the magnitude of a constituent.

Table 2: Tidal Constituents

name	symbol	speed (° per hour)	coefficient	speed
Diurnal species				
Lunisolar	K_1	15.041	0.265	γ
Larger lunar	O_1	13.943	0.189	$\gamma - 2\sigma$
Larger solar	P_1	14.959	0.088	$\gamma - 2\eta$
Semidiurnal species				
Principal lunar	M_2	28.984	0.454	$2\gamma - \sigma$
Principal solar	S_2	30.000	0.211	2γ
Larger lunar elliptic	N_2	28.440	0.088	$2\gamma - 3\sigma + \omega$
Lunisolar	K_2	30.082	0.058	2γ

Besides the diurnal and semidiurnal constituents there are long- and short-period constituents. Tidal oscillations of high frequency may be important in shallow-water areas.

Tidal constituents. As long as tidal motions are linear (*i.e.*, no interaction occurs between different constituents), the constituents of the Fourier series may be investigated separately. Table 2 shows that the principal lunar (M_2) tide is the predominant one. The nature of the tide potential gives rise to a similar development of the observed sea level into harmonic series—a series of terms whose reciprocals form an arithmetic progression, 1, 1/2, 1/3, . . . , 1/n. The frequencies are taken from the tide potential, and, for a given place, measurements of the sea level within a time interval are used to determine amplitudes and phases by various mathematical methods. After amplitudes and phases are known, the measurements may be reproduced. Furthermore, it becomes possible to synthesize the tide for any time, past, present, or future. This harmonic method may be used to predict tides and the components of tidal currents, but only for places for which sufficient long-term records of sea-level elevations are available. Harmonic-tide predictions are fairly good for deep-water ports; in shallow-water areas, especially in tidal estuaries, special corrections are needed in order to get precise results, or nonharmonic methods must be applied.

Special machines have been developed to synthesize tidal

Tide-predicting machines

constituents, especially to determine high- and low-water height and time. The first to design such a tide-predicting machine was the Scottish physicist William Thomson (Baron Kelvin) in 1872. In 1930 H. Rauschelbach constructed a machine containing 62 tidal constituents. Such analogue machines have been replaced by high-speed digital computers. Predicted high- and low-water time and water heights for important ports all over the world are published in tide tables prepared by hydrographic offices in the different countries.

In a similar way it is possible to predict tidal currents, but normally no such predictions are available.

Tidal measurements. A large number of tide gauges are in permanent operation along the coasts of continents and islands. The harmonic analyses of these sea-level records are collected and published by the International Hydrographic Bureau (IHB). Measurements of tidal currents are available for only a few places. The same is true of sea-level observations in the open sea. Modern instruments have made it possible to measure the vertical tidal components in the deep sea.

In the North Sea the tides are semidiurnal; in the South China Sea they are diurnal. Considering tidal currents, the amplitude and phase may vary from the surface to the bottom. Figure 28 represents tidal currents as a function of time for different depths, as revealed by records made west of the entrance to the English Channel.

From M.N. Hill (ed.), *The Sea*, vol. 1 (1962); John Wiley & Sons, Inc.

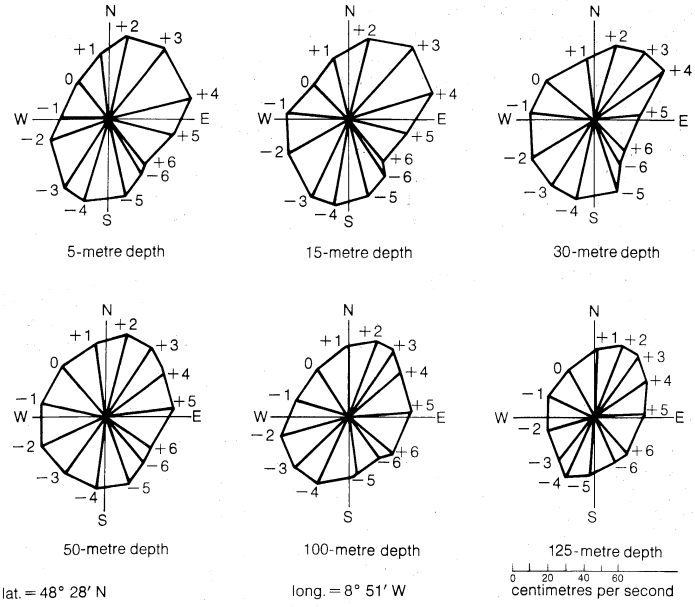


Figure 28: Observed tidal currents at the entrance to the English Channel, and their variation with water depth. Positive and negative values shown indicate the number of hours after and before the Moon's transit of Greenwich.

The elements of tidal current ellipses, including the amounts, direction, and time of maximum velocity, may be shown on maps. The large axes of the tidal current ellipses in Figure 29 are proportional to the maximum speed of the tidal current, and their orientation coincides with the direction of the maximum tidal current in the North Sea. Small axes refer to the minimum speed of the tidal currents. Full lines indicate the time of maximum speed of tidal currents referred to Moon's transit in Greenwich. Maps of this type provide an idea of the geographic distribution of the tidal currents in the sea. The time of the occurrence of high water and the tidal range and their geographic distribution are also of interest. Yet, because tidal sea-level records originate primarily from harbours and other coastal locations and because no measurements from the open sea are available, it is difficult to draw maps containing cotidal and corange lines (along a cotidal line the time of high water is constant, and along a corange line the tidal range is constant). In several cases such lines in the open sea have been constructed by means of an interpolation of coastal values. If no measurements are available, then the distribution of cotidal lines in the open sea remains uncertain. Corange lines have not been drawn. In order to overcome these difficulties, theoretical methods have been used.

TIDAL DYNAMICS AND PREDICTION

Mathematical-physical laws governing tidal motion

Tides and tidal currents are best explained by the mathematical-physical laws that govern the tidal motion in the sea. These laws are expressed as hydrodynamic differential equations, a system that consists of the equations of motion and the continuity equation. The equations of motion express the several forces that produce tidal accelerations. The continuity equation expresses the fact that the volume of any body of water must remain constant despite its motion. The density of the seawater is assumed to be constant.

Under certain conditions the hydrodynamic equations make it possible to determine the tides and tidal currents in an estuary, in a sea, or even in the world ocean. These conditions are: (1) the tide-generating accelerations are given quantitatively as functions of time and space; (2) the distribution of the depth and the shape of the coastline of the sea are considered to be known; (3) boundary conditions must be defined, especially those which relate to the coastline: no flow normal (perpendicular) to the coast.

Condition (1) is fulfilled because the tide potential is known very precisely. Condition (2) is normally met by existing pilot charts. With respect to condition (3), there are two types of boundaries. The first is represented by a coastline, and the condition is: no water transport into or out of the land; consequently, the normal component of the tidal current must be zero. The world ocean is an example of a sea totally surrounded by land. The advantage of this coastal boundary condition is that the tides are uniquely determined without using any tidal measurements. The second type is the open boundary. If the sea level is prescribed along a line around the open sea, then the tides within the area bounded by this line are uniquely determined. In natural seas both of these boundary conditions normally occur. The second condition can be applied in cases of neighbouring seas.

Tidal prediction using analytic methods. In the past, mathematicians and physicists developed analytic solutions of the hydrodynamic equations, partly in Cartesian, partly in spherical coordinates. For this purpose it was necessary to simplify the equations, especially with regard to the depth and the shape of the water body considered. In recent years tides in oceans bounded by meridians have been investigated by two British oceanographers, Joseph Proudman and Arthur Doodson, but it has not been possible to reproduce the natural tide quantitatively by means of these analytic solutions.

A very simple solution is obtained by assuming the tidal

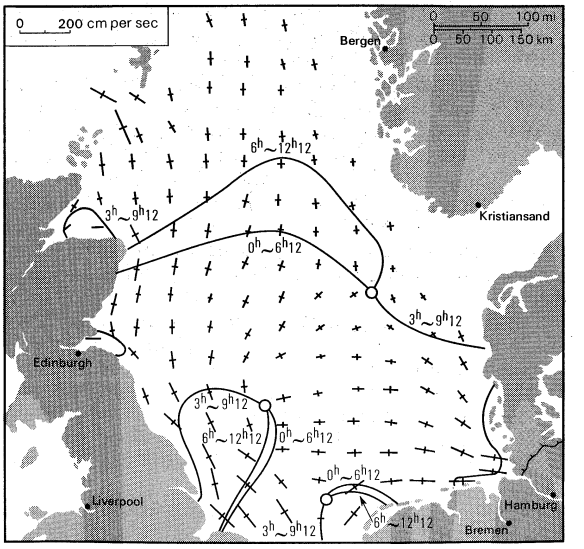


Figure 29: Major and minor axes of tidal current ellipses in the North Sea (see text).

currents to be zero. In this case the tide-generating accelerations are balanced by the shape of the sea surface, and the equilibrium tide is reproduced. There is no real possibility of reproducing natural tides in such a simple way.

There exists a large number of solutions concerned with tides in areas of very simple geometry. Amplitudes and phases depend on the ratio of the frequency of the tide-generating acceleration to the eigenfrequency (characteristic frequency) of the basin. The amplitudes increase if this ratio approaches one. Friction inhibits the unlimited increase of the amplitude if frequencies are equal to each other. The eigenfrequencies depend on the geometry of the basin.

Eigenfrequencies

A tide-generating constituent in a narrow canal that is closed on both ends causes simple harmonic oscillations to occur. These tides are called independent tides. In a canal closed on one side and open on the other, tidal oscillations may be caused by co-oscillation with the water masses beyond the open end of the canal. These tides are called co-oscillating tides. In many adjacent and marginal seas the co-oscillating tide is much more important than the independent tide.

The distinction between completely and partly closed seas is important in the view of tide reproduction. In the case of a completely closed basin, tides and tidal currents can be determined quantitatively without using any measurements. In a sea with partly open boundaries, such as the North Sea or the Gulf of St. Lawrence, it is necessary to prescribe the sea level as a function of time on or along the open boundary. Completely closed seas, then, are well suited to checking the results derived from hydrodynamic equations against measurements from the sea under consideration.

Although the world ocean fits these conditions, it is too complicated to use in such an analytical investigation. There are no other completely closed seas on Earth, and the tides of lakes, such as Lake Baikal, are small. Other seas that are almost completely closed have, like lakes, only very small tides: the Black Sea, the Baltic Sea, and even the Mediterranean Sea are examples.

Therefore, in order to check hydrodynamic theories, it is necessary to restrict the investigations to seas with boundaries that are not too open. The tides within such an area should be appreciable. Although a full check of the quality of hydrodynamic theories is not possible in such cases, because the solution of these equations is influenced by the prescribed sea level, the computed tides in the interior of the area may be compared with the measurements in corresponding locations, and in this way the quality of the solution may be tested.

Tidal prediction using hydrodynamic-numerical methods. *Tides in natural seas of small width.* Solutions of the hydrodynamic equations, taking into account arbitrary depth and coastline distribution, are developed as follows: time is eliminated by assuming that the sea level is a pure harmonic function of time. The derivatives in space direction are then replaced by finite differences. Finally, a system of linear equations remains to be solved. The tides in canallike seas have been investigated using this difference method in one dimension. As a boundary condition, the tidal currents are assumed to be zero at the closed end, and the amplitude and phase are prescribed at the open end. The numerical computation is a simple one, and no mathematical difficulties appear. The Austrian oceanographer Albert Defant and others have applied these or similar methods to several canallike adjacent areas, among them the Adriatic Sea, the Red Sea, and the Persian Gulf. Figure 30 represents Defant's results for the Red Sea. The quality of the reproduction of the tidal measurements is remarkable; furthermore, these investigations yield an understanding of tidal processes in elongated canals. The way in which the resulting amplitudes and phases are related to the geometry of canallike seas is clear.

Finite difference method

Defant also considered the North Sea, and even the Atlantic, as an elongated canal in order to apply the one-dimensional method, but it was not possible to reproduce the measurements quantitatively. Numerical methods for solving tidal problems in two dimensions were developed. First, the boundary-value method was performed, the time

Boundary-value method

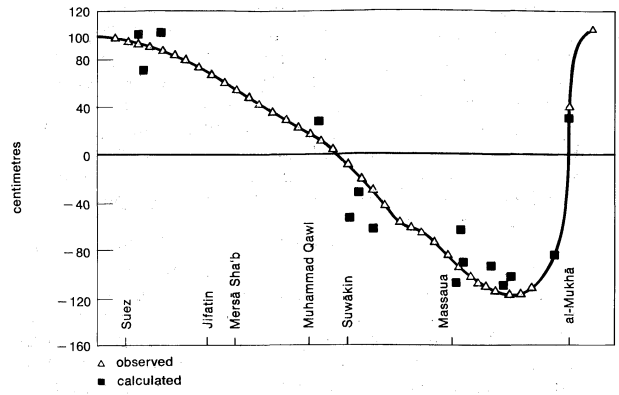


Figure 30: Range of observed and calculated values of the M_2 tidal constituent in the Red Sea.

dependency again assumed to be harmonic. After eliminating the components of tidal currents, for the sea-level variation there remains a partial differential equation that is replaced by a system of linear equations that relate sea level to the points of a grid placed upon the sea under investigation. The tide is determined if the sea level is defined by the grid points on the boundary. It is also possible to transfer the hydrodynamic equation, after introducing a time factor, into a system of linear equations without eliminating the components of the tidal currents.

In water that is not too shallow this boundary-value method yields a sufficient reproduction of measured tides. There are, however, a number of disadvantages. Different tidal constituents must be computed separately. The resulting sea level is a pure harmonic function of time. Because the equations are linear, no nonlinear interaction can be taken into account. Especially in shallow-water areas, such as tidal estuaries, tides and tidal currents may differ strongly from a harmonic function of time. Further-

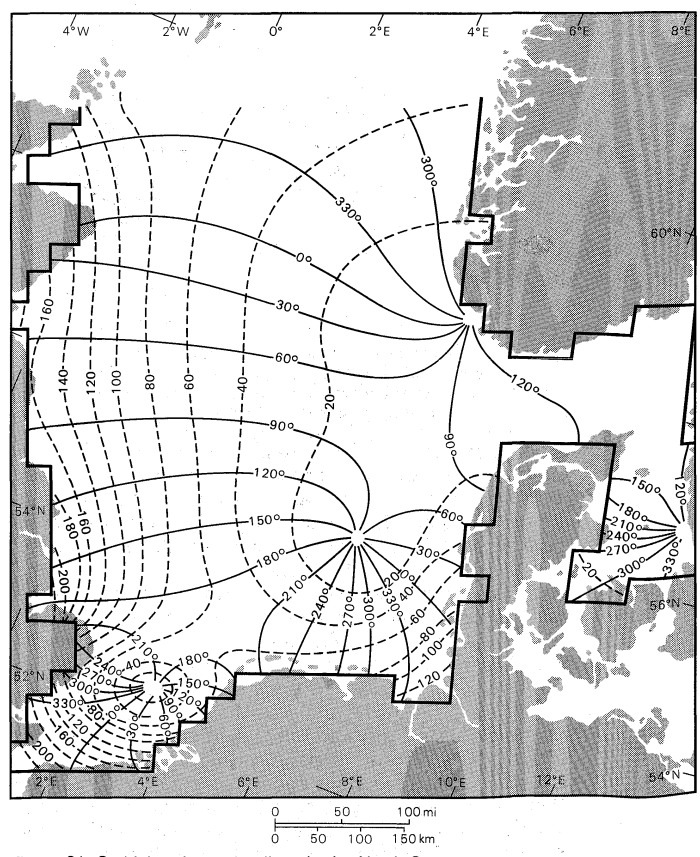


Figure 31: Cotidal and corange lines in the North Sea according to HN computation. Measured range and time of high water in coastal locations. Zero time equal to Moon's transit in Greenwich. Amplitudes in centimetres.

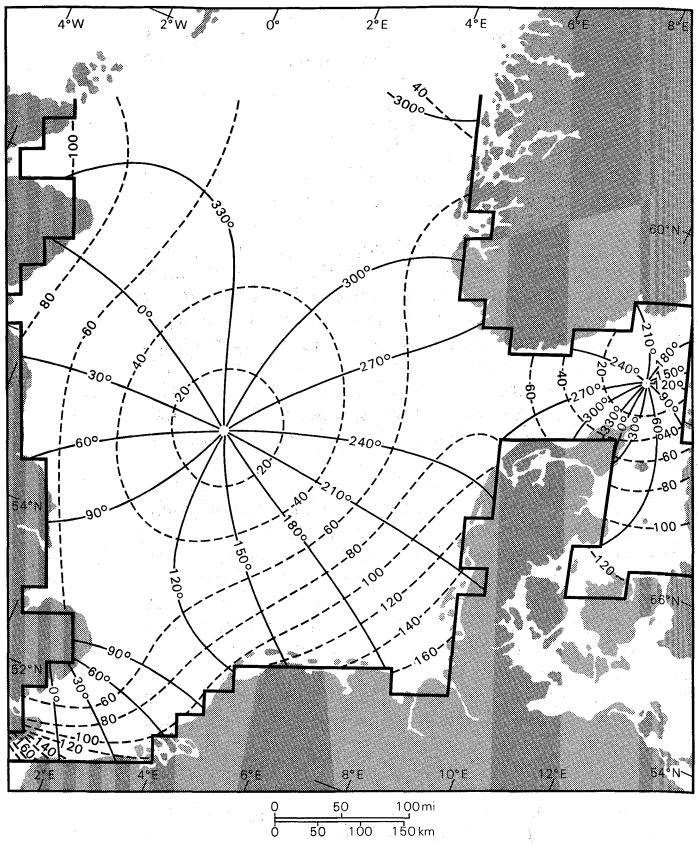


Figure 32: Tides in a North Sea basin of constant depth (80 metres), with the Earth's rotation taken into account. Solid lines are cotidal lines referred to Moon's transit in Greenwich. Dashed lines are amplitudes in centimetres.

more, it is impossible to investigate the interaction between different tidal constituents and among tides, surges, and gradient currents.

General prediction of tides. Attempts to design more general numerical methods, as a continuation of the boundary-value method but avoiding its disadvantages, have been made since 1945. These efforts resulted in the hydrodynamic-numerical (HN) methods, which also may be called initial boundary-value methods.

The method is fairly general and is applicable to all kinds of processes of motion in the deep ocean and in shallow waters. Nonlinearity and nonharmonic time dependency can be taken into account, and the method yields not only tidal motion but any motion in the sea, including turbulence.

Computers were not initially available for computation work with hydrodynamic-numerical models, and it was not possible to start investigations with the world ocean. Tidal estuaries, adjacent, and marginal seas were therefore treated first.

Computing tides in the North Sea

As an example, the tide in the North Sea has been computed by means of a hydrodynamic-numerical model. In this area, tides along the coast and tidal currents in the open sea are fairly well known, and a comparison between computed and measured values is possible.

Figure 31 shows the corange and cotidal lines of the principal lunar (M_2) tide of the North Sea, obtained by hydrodynamic-numerical-model computation. The correlation of observed and computed sea levels is excellent.

The influence of the depth distribution and the rotation of the Earth is indicated in the results of two HN models that accounted for these assumptions. Figure 32 shows the cotidal and corange lines of the tide in the North Sea with a constant depth of 80 metres; Figure 33 shows the cotidal and corange lines with the depth constant (80 metres, the mean depth of the North Sea), but the Coriolis acceleration (that due to rotation of the Earth) assumed to be zero—hence the effect of the rotation of the Earth is neglected. The results show that the high tidal range along the British coast is not caused by the rotation of the Earth, as had been assumed in earlier investigations. HN models make it clear that the depth of the North Sea is predominantly responsible for observed distribution of the range.

Corange and cotidal lines that are drawn for the tides in a rectangular basin of constant depth, which, as a very rough first approximation, represents the North Sea, reveal amplitudes and phases of the principal lunar (M_2) tide that are quite different from the tides discussed immediately above. This demonstrates in a distinctive way the important influence of the coastal configuration, and that it is not possible to reproduce the tides in natural seas by investigating basins of simple geometry.

Before considering the world ocean, it is useful to examine tidal prediction in a body of water of intermediate size like the Atlantic Ocean. Figure 34 shows the computed sea level at the time of the Moon's transit in Greenwich and three hours six minutes later, compared with the measured sea level along the coasts of the Atlantic Ocean. Again, the hydrodynamic-numerical method can be considered successful.

Tides in the world ocean

Figure 35 represents the world ocean including amphidromic points of the principal lunar (M_2) tide delivered by a 6° model. Amphidromic points, characterized by zero tidal range, are points at which the sea level remains unchanged by the tides. Dashed lines are referred to zero sea level at the time of the Moon's transit in Greenwich, and full lines correspond to zero sea level at three hours six minutes after the time of the Moon's transit in Greenwich. Because the hydrodynamic-numerical model does not require any tidal measurement, the results are independent of sea-level records. On the other hand the grid size of 6° in this model is relatively large, and, consequently, not all the fine structure of the ocean's depth and coast is sufficiently accounted for.

Accuracy of hydrodynamic-numerical methods. Using hydrodynamic-numerical models, it is possible to perform tidal predictions of sea level and tidal currents for an entire

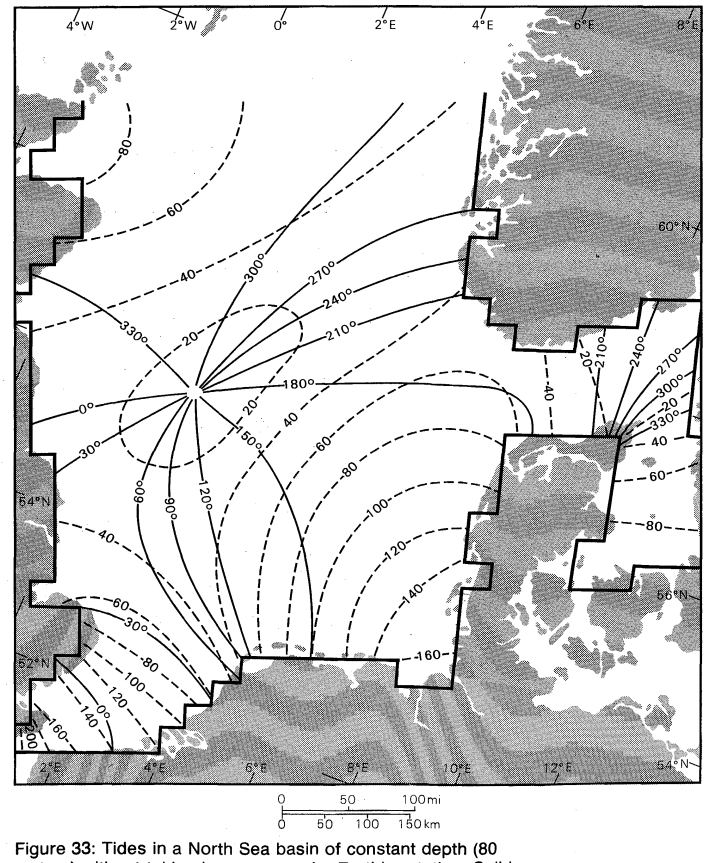


Figure 33: Tides in a North Sea basin of constant depth (80 metres) without taking into account the Earth's rotation. Solid lines, cotidal lines, refer to Moon's transit in Greenwich. Dashed lines are amplitudes in centimetres.

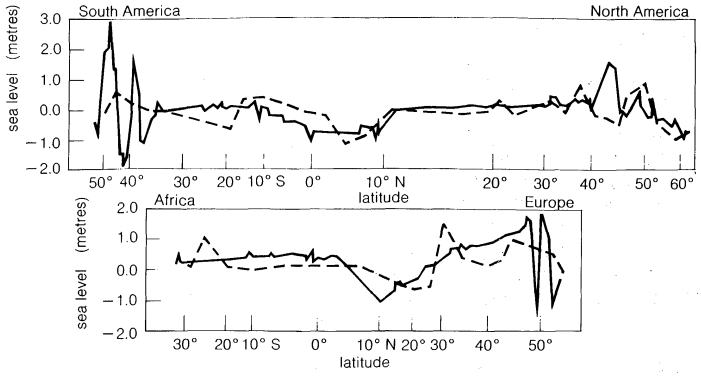


Figure 34: M_2 constituents along the Atlantic coast, showing computed (broken lines) and measured (solid lines) sea level.

sea with only one computation. At the open boundary, the tidal constituents given in Table 2 (at least M_2, S_2, N_2, K_2, K_1, O_1, P_1) are prescribed by their amplitudes and phases. By means of these constituents the tides on the boundary are predicted. The model determines tides and tidal currents as a function of time at all grid points in the interior of the basin. This prediction method has been tried for several areas, and Figure 36 shows the model computation compared with tidal prediction for selected points around the North Sea. Tidal currents are predicted by the same computation. Because only measurements on the boundary are needed, hydrodynamic-numerical-model tidal prediction provides a maximum of information with a minimum of measurements.

Effect of shallow water. In shallow-water areas the nonlinear terms in the hydrodynamical equations become increasingly important. Different tidal constituents, such as M_2 and S_2, cannot be treated separately, and the sea level is no longer a simple harmonic function of time. For tidal currents in the harmonic case, a water particle returns to its original location within a tidal period, thus describing a tidal ellipse. In the nonlinear case, however, the water masses do not return to their origin within one tidal period. Consequently, a periodic process on the open boundary of the sea results in a nonperiodic motion like a permanent transport in the interior of the sea under consideration.

Tidal currents. Tidal current measurements show the variation of velocity as a function of depth (see Figure 28). Attempts have been made to determine the tidal current as a function of depth. In estuaries and in narrow straits with alternating tidal currents, it was found that the true velocity is approximated by the formula

$$U = U_{ob}\left(\frac{Z}{H}\right)^{a},$$

in which U_{ob} is the surface velocity, Z is depth, H is the total depth, and a is a coefficient that varies between $1/5$ and $1/7$.

If the water in the sea is stratified, there are layers, each with its own density, which may oscillate at their interfaces. Amplitudes may become remarkable and can be expected to be on the order of tens of metres. This type of motion is referred to as internal waves; the oscillations are irregular, but there are periods of a day or of half a day. Hydrodynamic-numerical models are specifically designed for stratified water masses; however, not enough measurements are available for comparing computations and measurements.

Practical aspects. The knowledge of tides and tidal currents can serve important purposes. Navigators approaching a coast and steaming into a river need information on the depth in these areas; moreover, it is advantageous for a ship to sail in the direction of the tidal current flow. Coastal engineering work is also intensively concerned with tides. The regulation of a tidal estuary, the damming of a tidal river, and the construction of a harbour are problems that require knowledge of tides and tidal currents; such work may change the natural condition to such a degree that unwanted effects arise. As an example, the damming of a tidal river may cause flood current of very short duration and high velocity and, consequently, an ebb current of long duration and low velocity. A river in a sand-bed channel allows transport of such bottom material landward during each tidal period, but if the ebb current is extremely small, this material will not return to the sea within the same tidal period. The result is that the depth of such a tidal river decreases within a short time, and ultimately the river cannot be used as a waterway for ships. Information on the behaviour of the tides in a river that will result from artificial construction is therefore of great practical value, and hydrodynamic-numerical models may be useful in such cases. Other practical aspects of tide information are related to efforts to convert the energy of tides into electric energy and to the problems that arise from the pollution of water in coastal areas and in the open sea.

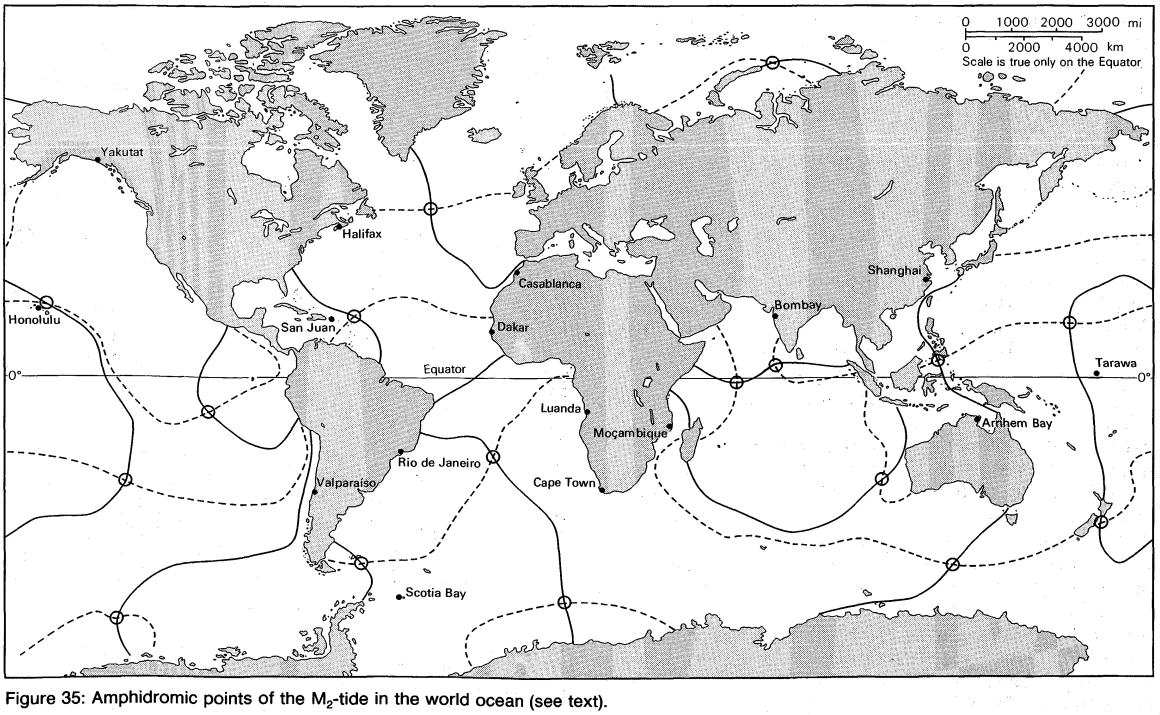


Figure 35: Amphidromic points of the M_2-tide in the world ocean (see text).

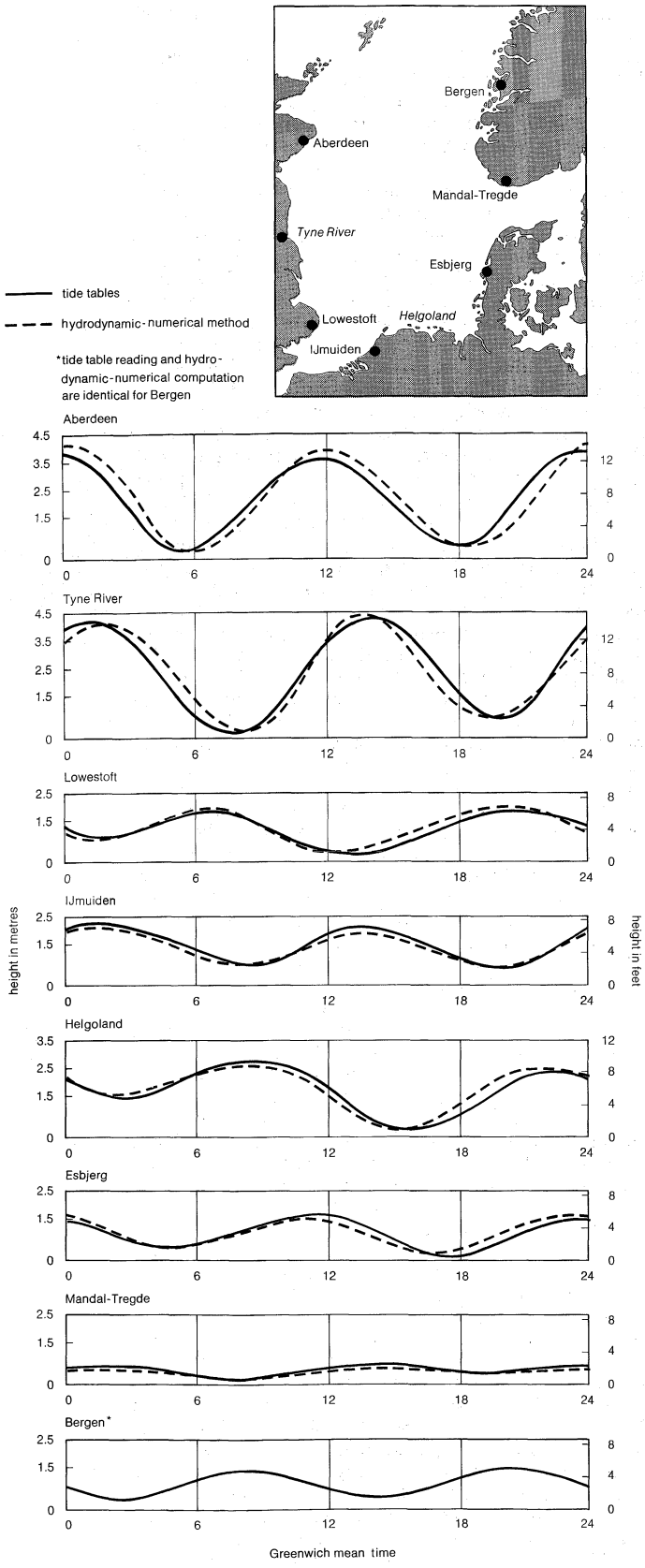


Figure 36: Comparison of high- and low-water time and height in the North Sea based on computation by the hydrodynamic-numerical method and on tide tables.
By courtesy of the Institut fur Meereskunde der Universitat Hamburg

TIDES IN THE ATMOSPHERE AND SOLID EARTH

Atmospheric tides. Like the water, the atmosphere is under the influence of the tide-generating acceleration, and air tides are to be expected. In the layers of air near the surface, pressure oscillations of the principal lunar (M_2) as well as principal solar (S_2) frequencies have been determined for a number of places. Because of the lesser density of the air, the tides in the atmosphere are much smaller than those in the ocean. Analysis yields larger amplitudes for S_2 than for M_2 atmospheric tides, which is explained by the fact that daily heating and cooling is responsible for these oscillations. Because the frequency is precisely the same as the frequency of the tidal-generating oscillation, it is not possible to separate the effects. The M_2 amplitude of atmospheric tides is of the order of magnitude of 0.06 millimetre (0.002 inch) mercury.

Atmospheric oscillations

In the ionsophere there are different layers that oscillate with lunar or solar frequencies; the amplitudes may be of the order of several hundred metres.

Earth tides. The tidal potential also acts on the solid body of the Earth. Whereas it is possible in the case of ocean tides to assume that the depth of the ocean is small compared with its horizontal dimensions, the density variation and other properties of the solid Earth make the determination of bodily tides more difficult than determination of ocean tides.

Difficulty of calculating bodily tides

Knowledge of the precise nature of the solid body of the Earth as a whole is limited, but it is known to be influenced by the tide-generating acceleration in two different ways. First, these accelerations act on the particles of the solid body of the Earth, which behaves as an elastic body, so that tides are developed within it. In addition, the tides of the ocean, especially the water masses transported by tidal currents, cause a periodic loading and unloading, a process that results in indirect tides. If the solid body of the Earth were completely rigid, no bodily tides would occur, and, at least in principle, the tide-generating forces could be measured; in this case they would be in agreement with the equilibrium tide. Measurements indicate that the theoretical values are not reached and that the elasticity is comparable to the elasticity of a steel sphere the same size as the Earth.

Tide-generating accelerations are a spacewide phenomenon acting on the particles of celestial bodies. If there are oceans on other planets, it is possible, in principle, to determine the tides in such oceans, and, if all the astronomical parameters are known, the bodily tides of the planets also can be calculated. (W.Hn./Ed.)

ASTRONOMICAL TIDES

Tidal effects on the planets and other bodies in the solar system, like those on the Earth, are caused by differences in the pull of gravity exerted by one body on various parts of another. The gravity difference can cause an elongation through stress if the subject body is deformable; if it is not, the stress set up may cause some adjustment of the distribution of material. In extreme cases, tidal force caused by an external gravity difference may be greater than the internal gravity of a body combined with any cohesive forces that may be present in it. In such cases, the body will be disrupted by the tidal force, and it will break up. In the solar system, the most interesting tidal effects are found in the Earth–Moon system, in the Mars satellites, and in the rings of Saturn. Tides also influence stars, star clusters, and galaxies.

A tidally produced elongation of an astronomical body may be permanent or temporary, according to the nature of the material of which it is composed. Also, in the absence of drag (tidal friction), the elongation may be in the direction of the gravity difference (equilibrium tide), or, when the body is rapidly rotating, the tidal bulge may be carried forward around it by an amount that depends on the physical conditions.

Just as the Moon raises tides on the Earth, the Earth, with its far greater mass, raises tides on the Moon. Moreover, because the surfaces of stars are gaseous, the individual stars of many close pairs in binary star systems (two stars that revolve around a common centre of gravity) are elongated by tidal forces and often show evidence of streams of gas passing between the two stars.

Effect of the Earth on the Moon

Lunar tides. Tides raised on the Moon by the Earth are much simpler to depict than those raised on the Earth by the Moon. Because there is no surface water on the

Moon, it is only necessary to consider the tides in the solid material—namely, the bodily tides. Solar tides can be neglected for the Moon. The solar tide-raising potential is about the same for both Earth and Moon, but tide-raising potential does not need to be taken into account when considering lunar tides; as the Earth is much closer to the Moon, the solar tide is overwhelmed by the Earth's tidal effect on the Moon. On the average, the solar tide-raising potential on the Moon is only 1.5 percent of the Earth's and is never greater than 3 percent.

The long-term effect of tidal friction on the Moon has been to slow its rate of rotation until it is synchronous with its period of revolution around the Earth. (The same effect on the Earth is increasing the average length of the day by about two milliseconds per year.) The Moon rotates once each month and completes one revolution of the Earth in approximately the same time. For this reason, the same face of the Moon is always turned toward the Earth. Because the Moon's orbit around the Earth is irregular, it alternately approaches and recedes from the Earth. While rotating at constant speed, which is so slow that tidal friction is very small, the Moon moves forward fastest in its orbit at the point where it is closest to the Earth (perigee) and slowest at the point at which it is farthest away (apogee). Because of these variations in the Moon's rate of revolution around the Earth and because of the inclination of the Moon's axis of rotation, small areas near the Moon's edge appear and disappear from time to time. As a result of this oscillatory phenomenon, which is called libration, slightly more than half of the Moon's surface is seen as it revolves. At the same time relative distances between the Earth and individual areas on the Moon's surface are altered. Thus, tidal forces on the Moon due to the Earth are greatest at perigee and least at apogee. Moreover, because the amount of change in these forces varies with the libration position, they may differ considerably from one time to another at the same place on the lunar surface.

Evidence suggests that two types of phenomena on the Moon are tidally related. Temporary changes in the appearance of small areas occasionally have been noted. The sighting of glows of light, colour changes, and obscuration of detail in such areas have been reported since the late 18th century, and similar observations were made even before the invention of the telescope. These lunar transient events seem to occur unpredictably, but statistically they are observed more frequently at perigee, when the overall tidal effect of the Earth on the Moon is greatest.

Tidally related lunar events

In 1970 the discovery of seismic signals, which measure interior disturbances in the Moon, heightened interest in lunar tides. One type of signal, called type A, occurred at each perigee and was found to originate at very nearly the same place each time. Because it was extremely unlikely that meteorites would fall precisely at each perigee in the same place, it was concluded that the disturbances that caused the A-type signals were moonquakes, triggered by extreme tidal conditions. A second group, type B, were believed to be tidally related because they were found to occur sometimes at perigee and also at apogee.

The relative frequencies of the two types are so similar that it now appears certain that the lunar events are related to moonquakes of a kind similar to those with which the seismic signals were associated; the glows and obscurations may represent the appearance of small quantities of gas and dust, tidally released through cracks in the Moon's surface. This idea is also supported by the small amounts of gases that have been detected on the Moon.

The Roche limit. Although a tidal force on a planet or satellite in the solar system is essentially disruptive, it amounts to only a small fraction of that body's own gravitational forces. If, however, such forces are sufficiently large, they may overcome the combined effect of internal gravity and the cohesive forces that hold a body together. The body will then begin to break up.

If a natural satellite moves closer to its planet through time, the effect of the planet's tide-generating force increases rapidly, and the spherical shape of the satellite becomes distorted. If the satellite is sufficiently deformable, it will become more and more elongated. Finally, if it comes close enough to the planet, the satellite will break up completely.

A French scientist, Édouard Roche, calculated the shape that a small satellite would assume if it were free to adapt itself to tidal stresses under conditions similar to those just described. Roche also determined how close such a satellite could approach its planet without breaking up. In order to change its shape freely the satellite would have to be a fluid, composed either of a gas or a liquid. Although this situation is quite artificial, it can be used as a first approach to find the absolute limits in some real examples.

Roche found that the limiting figure, or shape, of such a liquid or gaseous satellite is a prolate spheroid—a nearly egg-shaped body in which the longest axis is slightly more than twice the length of either of the shorter ones. The satellite is least likely to break up when it moves with its long axis pointing directly toward the centre of the planet. Roche also proved that when the satellite and the planet have the same density, the satellite will break up, in the absence of any internal cohesive forces, at a distance of 2.44 times the planet's radius, or 1.44 times the radius above the planet's surface. In other words, for a planet and satellite of the same density, the satellite cannot exist permanently inside a sphere drawn around the planet at 2.44 times its radius, or 1.44 times the radius above its surface. The limiting distances, or Roche limit, in any such situation depends on the relative densities of the planet and satellite as well as the planet's radius; it would be farther away from the planet for higher planetary density and closer for a satellite that is denser than its planet.

Tidal stability limit

Although the Moon is well outside the Roche limit of the Earth, many small artificial satellites have been put into orbits much closer, and some of them have lasted for several years. Eventually, though, the satellites lose energy through atmospheric drag and spiral toward the Earth. If, however, the parts of the satellite were not rigid and rigidly held together by rivets, screws, and the like, the satellite's internal gravity would not be sufficient to prevent the parts from gradually separating. The case of a nose cone and a space vehicle, or of two space vehicles traveling together, perhaps illustrates the point. Unless the units are physically joined by a clamping device, continual small adjustments in speed must be made to keep them together, or else they will slowly drift apart.

Saturn has 20 known satellites, three of which orbit within the Roche limit, and the rest outside of it. The planet also has around it a remarkable series of broad flat rings, the nature of which remained a puzzle for some 200 years following their discovery in the 17th century. The rings are inclined to Saturn's orbital plane and also to the orbital plane of the Earth (the ecliptic). Most of the time the rings appear as concentric ovals that change in shape; occasionally, they disappear from view. Because this happens whenever the Sun or the Earth passes through the plane of the rings, it indicates that the material of the rings is confined very closely to a plane.

The rings of Saturn

Because the whole system of rings is inside the equal-density Roche limit, the existence of the rings can be directly attributed to tidal forces. This was recognized by Roche in 1848 and was explained in mathematical terms in 1859 by the Scottish physicist James Clerk Maxwell. Saturn has such extremely low density that a piece of it would float in water. If, on the other hand, the ring material is solid, it must be heavier than the material of which Saturn is composed and the actual Roche limit must be beyond the rings. It appears that the rings are indeed composed chiefly of small, separate particles of water ice revolving in free orbits and held together by internal cohesion only. The particles probably collide occasionally and thus erode very slowly into smaller and smaller particles. Even if such a flat system as Saturn's rings could exist temporarily as a single continuous sheet, it would be unstable, and the tidal force of the planet would break it up.

Mars has two satellites, Deimos and Phobos. Deimos is relatively far removed from the planet, but the orbit of Phobos is close to Mars, at a mean distance of 2.75 Martian radii. The density of Mars is known to be 3.95; if the densities of Phobos and Mars are in about the same proportion as those of the Moon and the Earth, the Roche

The Mars satellites

limit will be roughly 2.86 times the planet's radius. Thus, Phobos revolves at just about the limit of stability.

In 1970 the image of the little satellite was recorded by chance by a spacecraft circling Mars. The spacecraft, which was engaged in photographing the planet itself, was more distant than Phobos from the planet. Silhouetted against the body of the planet, Phobos was seen to be elliptic. It measures about 23 kilometres in diameter along the orbital and approximately 19 kilometres in the perpendicular direction. Eventually, it may be broken up by the tidal action of Mars.

Tides in stars and galaxies. Tidal disturbances are present in many binary star systems. Observations show that stars in such systems are usually ellipsoidal rather than spherical in shape. The amount of distortion can be measured in terms of the ellipticity of the star; in general, the shorter the period of rotation of one star about the other, the greater the ellipticity. Because the period, in turn, depends on the size of the orbit and because the square of the period is proportional to the cube of the mean distance between the stars, the ellipticity depends strongly on the distance. This effect is at least partly tidal, though in some binary star systems—particularly those containing very hot (young) stars—another factor, rapid rotation, may also be active. In such cases, the two possible causes of ellipticity would reinforce each other.

In many binary systems, spectral measurements show that streams of gas pass from one star to the other. The best-known systems with gas streams of this kind are Beta Lyrae (the second brightest star in the constellation Lyra, the Lyre) and W Ursa Majoris (the variable, W, in the constellation of Ursa Major, the Great Bear). In Beta Lyrae, only one star can be seen; the other is too faint. Nevertheless, the gas streams that have been detected (as bright mission lines in the spectrum) are flowing from the bright, early-type star to the fainter, which is believed therefore to be the more massive, one. In addition to showing tidal distortion, the bright star also rotates rapidly, and this increases the instability of the system.

Disintegration of star clusters

The stars in a cluster are in continuous motion and interact gravitationally with one another. Close chance encounter between the stars in a cluster, however, may change any star's motion sufficiently for it to escape from the group. The differential rotation of a galaxy and its tidal action on the cluster both act to separate the innermost and outermost parts of the cluster; in time, the less massive stars are lost, and the cluster progressively disintegrates. For the Pleiades, the time that is needed for disintegration is about 10^9 years; for a looser group, or stellar association, the so-called O-B class (hot and bright) of stars, very much less time is needed (see also STARS AND STAR CLUSTERS).

Some external galaxies occur in pairs or groups. When the pairs are quite close in the sky, it is sometimes found that jets and streams of bright material extend from one to the other. The inference in these cases is that the pairs are actually close in space and are interacting tidally. In spiral galaxies with companions, radio observations of neutral hydrogen have shown some asymmetries that are also believed to be due to tidal distortions. One such example is known as M101 (NGC 5457). Similar but rather smaller asymmetries have been found by radio measurement of the hydrogen gas in the galaxy called the Andromeda Nebula and its nearby companions; others are believed to exist in the Milky Way system, which has as close companions the Magellanic Clouds. On the other hand, bright galaxies without close companions are apparently undistorted at radio wavelengths. Optical studies of bright galaxies have revealed that about 20 percent have close companions and exhibit distorted optical features. (B.M.M./Ed.)

WAVE MOTION

The ripples on the surface of a pond and the great tsunamis on the ocean, sometimes mistakenly called tidal waves, have something in common; they are both examples of wave motion. A disturbance—a pebble dropping into the pond or an undersea earthquake—excites water waves. Radiating outward from the disturbance, the waves propagate the signal to the distant reaches of the pond or ocean. Examination of the motion of a floating leaf near the disturbance shows that it bobs up and down and to and fro about its original position. That is to say, the leaf does not undergo any net displacement as a result of the wave motion.

The nature of waves

Another more distant leaf repeats the motion at a somewhat later time. The motion of the second leaf is a duplicate of the motion of the first leaf, except that perhaps the magnitude of the motion is not as great. Thus, the wave transmits information. Furthermore, because the leaves are set into motion, even moving against gravity, the wave transmits energy. Because the ripples travel over the surface of the pond, the waves have a definite velocity. Although the example of leaves helps to indicate the wave motion on the surface of the water, the waves actually propagate through water itself—*i.e.,* the water is the medium for the waves. By the mechanism of waves, energy is transmitted through the medium, although the medium itself experiences no net motion.

Water waves are just one example of a large class of similar physical phenomena. More generally, wave motion may be defined as the propagation of a disturbance through any medium. Such a wave motion can be described by specifying the positions of all points of the medium as time progresses. This space–time description can be stated mathematically in the form of a wave equation (see below *The wave equation*).

Other examples of wave motion are the transmission of musical sounds in the air of a concert hall, the arrival of light from a distant star, the passage of X rays through the body of a patient, and the sending of a commercial message via television; in other words, sound, light, X rays, and television signals propagate by wave motion. There is an essential similarity in the way these signals travel through their respective media.

DEVELOPMENT OF WAVE CONCEPTS

The history of the knowledge of waves is bound up with the entire development of science, for waves are encountered in many fundamental fields of physics. As noted earlier, by the 3rd century BC, the Greeks had developed some of the laws of mechanics. In the early 17th century, Galileo introduced the modern method of science—namely, the testing of theories by experiment. The ensuing burst of scientific activity during the rest of the century was capped by Newton's formulation of the laws of mechanics. Newton's laws, together with the laws of electricity and magnetism formulated by James Clerk Maxwell some 200 years later, form a comprehensive, quantitative basis for the understanding of classical physics, including basic wave phenomena.

The study of vibrating strings played an important role in the development of understanding of mechanical waves. The Greek philosopher Pythagoras, active in the 6th century BC, is said to have noted that pleasing sounds were made by a stretched string when it was divided into two segments with a simple numerical ratio between the lengths. Galileo measured the effects of mass and tension on the frequency of vibration of a string, as did a French scientist of his time, Marin Mersenne, independently. Johann Bernoulli and his son Daniel analyzed the vibration of a set of connected masses during the 17th and 18th centuries. In 1753 Daniel Bernoulli stated the first version of the superposition principle (see below *Superposition of waves*), which formed the basis of the later mathematical analysis of wave motion by another mathematician, Joseph Fourier of France, in the 19th century. The first formulation of a wave equation was made in the 18th century by Jean Le Rond d'Alembert, who applied it to the vibrating string.

Early formulation of the superposition principle

Newton derived the first formula for the velocity of sound

in air. In the 18th century Pierre-Simon Laplace modified Newton's formula to take into account the fact that the compression of the air associated with sound waves raises the temperature, affecting the sound velocity.

The early development of knowledge of electromagnetic waves, those waves involving the combined actions of electric and magnetic fields, such as light waves, overlapped that of mechanical waves. Galileo attempted unsuccessfully to measure the speed of light. Much later, in the mid-19th century, a French physicist, Armand-Hippolyte-Louis Fizeau, refined Galileo's method and succeeded. The first measurement, however, was made in 1676 by a Danish astronomer, Ole Rømer, who used observations of the eclipses of Jupiter's moons.

An Italian Jesuit, Francesco Maria Grimaldi of Bologna, discovered prior to 1665 the process now known as diffraction, whereby light spreads out from an aperture. Christiaan Huygens, a Dutch scientist, suggested as early as 1678 that light behaves like waves, and he invented a graphic method of constructing wave fronts that is still in use. Newton observed some interference phenomena, such as rings produced by light incident on small air gaps and the colours of light reflected from oil films, but he persisted in trying to explain his observations with his particle theory of light. An elegant experiment was performed in 1801 by Thomas Young, who passed light through two slits to observe the interference pattern. Young's work convinced scientists that light has essential wave characteristics in apparent contradiction to Newton's corpuscular (particle) theory.

The conflicting particle–wave aspects of light were reconciled much later with the development of the quantum theory. Meanwhile, a French physicist, Augustin-Jean Fresnel, showed that light waves consist of transverse vibrations, much like the vibrations of a violin string. Finally, in one of the greatest achievements of theoretical physics, Maxwell derived the properties of electromagnetic radiation from his general equations of electromagnetism. Maxwell's analysis unifies the understanding of the various kinds of electromagnetic radiation, including light, radio waves, X rays, and others.

Maxwell's equation of electromagnetic radiation

A prediction that particles of matter would show wave properties was made by the French physicist Louis-Victor de Broglie in 1923. Experimental verification was achieved by the American physicists Clinton Joseph Davisson and Lester Halbert Germer in 1927. The Austrian physicist Erwin Schrödinger formulated the quantum theory mathematically in the form of a wave equation for physical systems having atomic dimensions. The density of the Schrödinger wave is a measure of probability. The quantum theory, together with Einstein's theory of relativity, forms the basis of the modern understanding of radiation and matter (see below *Quantum mechanics*).

TYPES OF WAVES

Classification by medium. Waves can be classified in several ways. One of the simplest is by the medium supporting the transmission of the wave motion. For example, water waves denote waves that travel on water. Sound waves commonly travel in air but also move through any material, be it gas, liquid, or solid.

Light, X rays, radio waves, and infrared heat rays are all forms of electromagnetic radiation. Such radiation does not require a material medium—electromagnetic waves can travel through empty space. The light from distant stars travels through the void of interstellar space. Early scientists felt so keenly the need to postulate a medium for light that they invented a substance called ether. A famous experiment by two scientists in the United States, A.A. Michelson and Edward Williams Morley, in 1887, demonstrated the lack of any physical evidence for the ether. The ether concept has since been replaced by the concept of the electromagnetic field. The field propagates with the speed of light, even across a vacuum. That the field exists can be directly verified by placing an electric charge (*i.e.*, an electron) at some point in the field. The charge will be set into motion by electromagnetic forces. The propagation of an electromagnetic field and the associated energy through space constitutes an electromagnetic wave. The quantity corresponding to the mechanical displacement in a water wave is the strength of the electromagnetic field.

The ether concept

Motion of the particles in a wave. Waves may also be classified according to how the particles of the medium move. This classification has the advantage of emphasizing the similarity between waves, even though they have quite different media. For example, a wave can be set up by plucking a guitar string. The resulting wave is a transverse wave; *i.e.*, particles of the medium (the string) vibrate at right angles to the direction of propagation of the wave (along the string; see Figure 37). Another example of a

Transverse waves

From F.W. Sears and M.W. Zemanski, *University Physics*, 4th ed. (1970); Addison-Wesley, Reading, Mass.

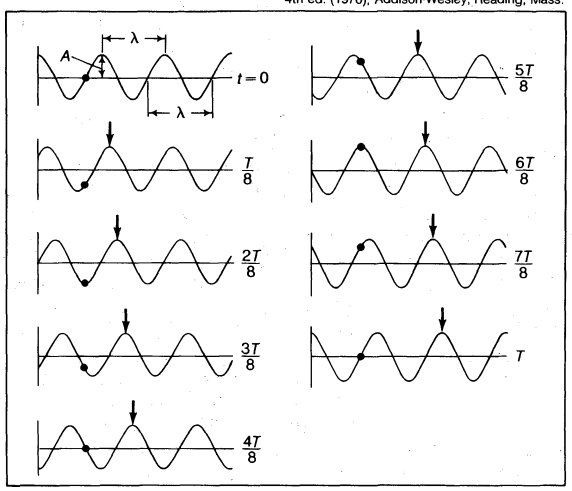


Figure 37: A sinusoidal transverse wave traveling toward the right, shown at intervals of one-eighth of the period *T*, beginning with time $t = 0$. *A*, maximum amplitude; λ, wavelength; dot, particle vibrating up and down with simple harmonic motion; arrow, traveling crest.

transverse wave is an electromagnetic wave, such as a radio wave. Here, the particles are replaced by the electric or magnetic fields. The direction of the electric (or magnetic) field is 90° to the direction of propagation of the radio wave. At a point in space, the magnitude of the electric field varies in time as the wave passes. An electron in an atom located in the field will be set into vibration with the electric field—*i.e.*, perpendicularly to the direction of propagation of the wave (see Figure 38).

From J.R. Pierce, *Electrons and Waves* (copyright © 1964); Doubleday and Company, Inc., reprinted by permission

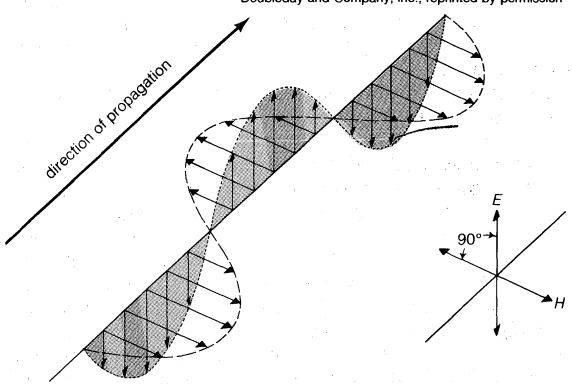


Figure 38: An electromagnetic wave. *E*, electric field components; *H*, magnetic field components. Dotted and dashed lines are sinusoidal (see text).

If the vibrations of a transverse wave are confined to a single direction in space, the wave is said to be plane polarized. A wave that is traveling in an *x, y, z* coordinate system in the *z*-direction could be polarized in the *x*-direction or in the *y*-direction, for example.

In a sound wave, the particles of the medium vibrate in the direction of propagation, in contrast to the case described above for transverse waves. Sound waves are examples of longitudinal waves. The longitudinal displacements of the air molecules result in a series of high-pressure and low-pressure regions called, respectively, condensations and

Longitudinal waves

rarefactions. These regularly spaced pressure zones travel through the air with a certain velocity. Such sound or pressure waves propagate through any material medium, including solids.

The ease with which a given type of wave is transmitted depends on the nature of the medium. For example, liquids transmit longitudinal waves readily but cannot support transverse waves because the liquid has no elasticity for transverse vibrations. The role of the medium in transmission has made it possible to study the physical constitution of the interior of the Earth. Earthquakes excite both longitudinal and transverse waves in the Earth. By observing the transmission of both kinds of waves through the centre of the planet, geophysicists have determined that the core is molten.

Torsional waves

Yet another kind of wave is a torsional wave. If the end of a metal rod is twisted, the twist travels down the rod. The particles in the rod undergo angular displacements about the centre of the rod in a plane perpendicular to the axis of the rod. If the end of the rod is twisted back and forth periodically, a periodic wave motion can be set up.

Surface water waves exhibit a unique property—the water molecules move in a circular motion. The circular motion results from the fact that water is nearly incompressible. Thus, the fluid cannot support a purely transverse vibration involving fairly large displacements but must simultaneously move longitudinally. A circular motion can be thought of as a combination of two linear vibratory motions at right angles to each other. In the water wave, the circular motion is in a plane perpendicular to the surface of the water containing a line in the direction of propagation. The sense of rotation is such that the water at the top of the circle moves forward in the direction of propagation, whereas the water at the bottom of the circle moves backward. The waves described here are traveling on the surface of deep water. Over shallow water, the motion becomes elliptical.

Bow waves and shock waves. Thus far, the kinds of waves that have been described are those generated by stationary sources or by sources moving at speeds less than the speed of the wave motion in the medium. If the source moves faster than the speed of the wave motion, a bow wave is formed. A simple form of the bow wave is produced by a ship. If the ship is moving faster than the surface wave ordinarily generated by its bow, the crest of the wave slants back from the bow in a straight line. Viewed from above, the bow wave is an inverted V, with the ship's bow at the apex. On the other hand, the bow wave generated by a supersonic aircraft is a three-dimensional wave. The crest of the wave is a cone, the axis of which is along the line of flight with the aircraft at the apex.

Source speed and wave speed

If a charged particle moves faster than the speed of light in a transparent medium such as glass or water, an optical bow wave is formed. The associated light, called Cherenkov radiation after the Soviet discoverer, Pavel Alekseyevich Cherenkov, is bluish in colour. This effect has been applied by nuclear physicists to the detection of fast-charged particles.

In all of the above examples of bow waves, a simple formula can be given for the angle the wave front makes with the line of motion of the source. If the speed of the source is u and the wave speed is v, then the angle, symbolized by the Greek letter alpha (α), is given by the following equation: sine of the angle is equal to wave speed divided by source speed, or, $\sin \alpha = v/u$. The reciprocal of the ratio v/u—that is, u/v—called the Mach number, after the Austrian physicist Ernst Mach, must be greater than 1 for the bow wave to form. Thus, the quantity $\sin \alpha$ will be less than 1, and the angle α is defined. The equation states that as the source speed, u, increases, the angle, α, decreases, in accordance with experience.

Shock waves

A bow wave from a supersonic object such as an aircraft also becomes a shock wave because of the properties of air. During the passage of the initial wave crest, the air is compressed. This compression heats the air, causing the subsequent waves from the trailing portions of the aircraft to travel faster, because the speed of sound increases with air temperature. Thus, the later waves catch up with the first crest, piling up to form an intense wave front, which causes the sonic boom. The pressure buildup also causes a curvature of the wave front because sound travels faster at higher pressure. At the apex of the shock wave, the crest makes a larger angle with the direction of travel than it does farther back.

THE THEORY OF WAVES

Vibratory motion. Wave motion consists of a vibratory motion of the particles of the medium. When a particle in an elastic material is displaced from its ordinary equilibrium position, a force is set up that tends to restore the particle to its original position. When the particle is displaced and then released, it will reach its original equilibrium position and be carried beyond by the particle's momentum. The restoring force will then reverse its direction. As a result, a vibratory, or periodic, motion will ensue.

One way of describing a certain kind of periodic behaviour is to employ the trigonometric functions sine or cosine. The sine of theta (θ), for instance, varies in magnitude between 0, +1, 0, −1, and 0 as θ, called the argument, changes in angle from zero through $\pi/2$, π, $3\pi/2$, to 2π radians (2π radians is equivalent to 360°). The cosine of θ will vary in the same way, except it will be out of phase by $\pi/2$, or 90°; that is, for a magnitude of $\cos \theta$ equal to 0, angle θ is equal to $\pi/2$; for $\cos \theta$ equal to −1, angle θ is equal to π; and so on.

An easily visualized example of vibratory motion is a mass suspended by a spring. In most elastic substances, if the strain (distortion) is not too great, the restoring force is proportional to the displacement but opposite in direction. This relation, known as Hooke's law, states that the restoring force (F) is equal to minus a proportionality constant (k) times the linear displacement (x)—*i.e.*, in equation form, $F = -kx$. The constant k is sometimes called the spring constant. If the displacement is doubled, the force is doubled, etc. The minus sign means that the force is in the opposite direction to the displacement.

Hooke's law

When a mass is suspended by a spring and the mass is pulled down and then released, it will bob up and down in vibratory motion. It can be shown that, if the spring obeys Hooke's law, the displacement (denoted by y) of the mass from its equilibrium (rest) position will be proportional to the sine of an angle equal to a multiple of 2π, the multiple being the product of the number of vibrations per unit time (called the frequency, f) and the time (t). In equation form it is

$$y = A \sin (2\pi ft). \qquad (23)$$

The factor A in the equation is a constant (independent of time or position), called the amplitude. Because the maximum value of the sine function is 1, the maximum displacement is A. A complete vibration involves the motion from the extreme downward position to the extreme upward position and then back to the starting point. The time elapsing during a single complete vibration is called the period. Because the amplitude and frequency are constants in this equation, the displacement at any instant is determined by the length of time (t) after initiating the vibration.

The frequency of the vibration is given by a numerical factor $(1/2\pi)$ times the square root of the ratio of the spring constant (k) to the vibrating mass (m), or $f = 1/2\pi(k/m)^{1/2}$. For larger spring constants (stiffer springs), the frequency is higher, whereas larger masses result in lower frequencies.

Simple harmonic motion

Motion according to equation (23) is called simple harmonic motion. As an example, a spot of paint on the rim of a rotating wheel, when viewed from the side, will appear to move up and down. In fact, if the wheel rotates at a constant speed, the spot executes simple harmonic motion because, if the wheel's radius is A, the vertical displacement of the spot from the centre of the wheel will be given by equation (23). The frequency, f, will be the number of revolutions of the wheel per unit time.

Simple harmonic motion is one of the most important kinds of vibratory motion. For small displacements, the particles of most elastic materials vibrate with simple har-

monic motion. It can also be shown that any vibration ordinarily encountered can be constructed by adding together a number of component vibrations, each of which is harmonic. The mathematical technique generally employed for harmonic analysis is called Fourier analysis and will be described below in the section *Pulses and wave trains.*

Most vibrating physical systems deviate at least slightly from pure harmonic motion. Such deviations, for example, contribute to the quality of musical sounds from instruments. Motion that is not harmonic is called anharmonic. Occasionally, a system will be encountered that is not harmonic even for small displacements. Such cases, however, are exceptional.

A vibrating system contains energy. At the extreme of the displacement, the energy is stored in the elastic material (*i.e.,* the spring) as potential energy. When the particle is passing through the equilibrium point, the material is not strained, and the energy is entirely in the form of kinetic energy of motion. Thus, the vibration involves a constant interchange between potential and kinetic energy. For simple harmonic motion, the energy (E) of the vibrator is directly proportional to the spring constant (k), but proportional to the square of the amplitude (A), and can be expressed by the equation:

$$E = \tfrac{1}{2}kA^2. \tag{24}$$

Doubling the amplitude will quadruple the energy.

Simple vibrators such as the mass on a spring are also called oscillators. A more complicated vibrating system can be built up by coupling together a number of oscillators. The coupling is an elastic connection that communicates the vibration of one oscillator to the adjacent oscillators. An example of a large number of coupled oscillators is the array of atoms in a crystal. Each atom can vibrate about its equilibrium position in the crystal lattice. The interatomic forces couple the atomic oscillators together to give rise to lattice vibrations. In this manner, waves can travel through the crystal.

Traveling waves

Velocity of waves. One of the simplest kinds of traveling wave has a profile at any instant of time given by a sine function. This kind of wave can be produced on a stretched string. If the string is stretched along the x-axis, the displacement (also called the amplitude) of a particle of the string at position, x, and time, t, is given by an equation in which the multiplying factor, ft, in the argument of the sine function is diminished by the ratio x/λ—*i.e.,*

$$y = A \sin 2\pi\left(ft - \frac{x}{\lambda}\right). \tag{25}$$

Equation (25) applies to an infinitely long string undergoing transverse vibrations in the y-direction. The wavelength, symbolized by λ, is the distance between adjacent crests (or troughs) of the wave. The maximum displacement y of the string is A (see Figure 37).

At the position $x = 0$, equation (25) reduces exactly to equation (23); in other words, the particles of the string at $x = 0$ vibrate up and down with simple harmonic motion. Equation (25) shows that other particles of the string also vibrate with simple harmonic motion but for progressively later times in the positive x-direction. In fact, the wave is traveling in the positive x-direction with a speed (v) given by the product of frequency and wavelength, expressed

$$v = f\lambda. \tag{26}$$

To see this, attention can be fixed on a point on the wave form that has no displacement—*i.e.,* $y = 0$. Such a point is called a node. The condition $y = 0$ can be met by requiring the argument of the sine function to be zero, as the magnitude of sin (0) is zero: $ft - x/\lambda = 0$. Solving for x, $x = (f\lambda)t = vt$. Because distance (x) is equal to velocity (v) times time (t), the factor $f\lambda$ can be identified with v, and equation (26) is derived. The velocity is the speed with which a given value of the sine function travels, and therefore v is called the phase velocity. The phase velocity is the speed with which a node or any other point on the wave form moves down the string.

Phase velocity

For a string, one can calculate the velocity from the formula that velocity (v) is equal to the square root of the tension (T), the stretching force applied to the ends of the string, divided by the mass per unit length, symbolized by μ, of the string, or $v = \sqrt{T/\mu}$. Increasing the tension results in increased velocity. The wave travels slower on a heavy string than on a light one. This is how, intuitively, a string would be expected to respond to variations of the conditions.

The functional form of the above equation for the phase velocity is typical for mechanical waves in that velocity depends on the square root of a measure of the strain of the material divided by the inertial mass of a unit amount of the material. The strain determines the strength of the restoring force, and the inertia, the kinetic response of the vibrator. The same functional dependence is evident in the expression for the speed of sound in a gas: $v = \sqrt{\gamma p/\rho}$. The pressure of the gas is represented by p and the density of the gas by ρ (the Greek letter rho). The constant, symbolized by γ, depends on the gas; for air, γ is equal to 1.40 at STP (standard temperature and pressure).

In the simple examples just given, the velocity is not dependent on the frequency of the wave, and the medium is said to be nondispersive. On the other hand, there are media in which the velocity depends on the frequency, and these media are called dispersive.

The wave velocities described above are the velocities of propagation of the waves with respect to the medium or with respect to an observer stationary with respect to the medium. If the observer is in motion with respect to the medium, the velocity of the wave propagation relative to the observer will be different. A remarkable exception is encountered in the case of electromagnetic waves such as light. One of the postulates of Einstein's special theory of relativity states that the speed of light in free space (*vacuo*) is a constant. This postulate means that the speed of light in empty space is the same with respect to any observer, irrespective of the observer's motion. This constancy is quite contrary to ordinary experience that is concerned with objects moving much slower than the speed of light. The experimental observations of Michelson and Morley, however, are consistent with Einstein's postulate. Subsequent work has provided overwhelming confirmation of Einstein's theory.

Speed of light in a transparent medium

In a transparent material, the speed of light is less than it is in free space. This decrease is the result of the interaction of the light wave with the material of the medium. The atoms of the material contain bound electrons that can be regarded as oscillators. The electromagnetic field associated with the light wave acts on the electric charges of the atoms, setting them into oscillation. These vibrating charged particles reradiate their own electromagnetic waves that combine with the original wave to modify the overall speed of the light in the material. If the index of refraction of the material is denoted by n and the speed of light in free space by c, then the speed of light, v, in the material is given by the ratio of the speed c to the refractive index, $v = c/n$.

Ordinary transparent materials such as water or glass have indexes greater than one; hence the speed of light in these materials is less than the speed of light in free space. This difference makes it possible for a charged particle to travel in a transparent medium at a speed greater than v but of course less than c. The result is a bow wave of light (Cherenkov radiation). The same interaction of the electronic oscillators in the material with the electromagnetic wave gives rise to optical dispersion; *i.e.,* the index of refraction depends on the frequency of the light, and hence the wave velocity also depends on the frequency. Optical dispersion can be seen in the separation of white light into a rainbowlike spectrum of colours by a prism.

Optical dispersion

The wave equation. The space–time description of wave motion in one dimension is represented by a differential equation stating that the second partial derivative of displacement (y) with respect to propagation distance (x) is equal to the reciprocal of the phase velocity (v) squared times the second partial derivative of the displacement (y) with respect to time (t), expressed

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y}{\partial t^2}. \tag{27}$$

The solutions to equation (27) are the sinusoidal traveling waves (the direction depends on the choice of sign), specified by the sum of a sine and cosine term $y = A \sin 2\pi (ft + x/\lambda) \pm B \cos 2\pi (ft \pm x/\lambda)$. Equation (27) is a second-order differential equation requiring two constants (A and B) in the solution. These are to be evaluated from the initial conditions. If the displacement $y = 0$ when time $t = 0$ and distance $x = 0$, then constant $B = 0$. Equation (27) is linear in y in that y appears raised only to the first power. An important result of this linearity is that the principle of superposition will be satisfied (see below *Superposition of waves*).

The wave equation (27) may be written in more than one dimension or in other coordinate systems, such as spherical coordinates. All of the mathematical treatment of wave motion stems from this equation. For a mechanical medium, Newton's laws are applied to derive the phase velocity v. In the case of electromagnetic waves, Maxwell's equations must be employed to determine the phase velocity c of electromagnetic radiation.

Transport of energy. As was pointed out above, when a simple sinusoidal wave propagates along a string, each particle in the string undergoes simple harmonic motion. Each vibrating particle represents an oscillator with a certain energy content (see equation [24]). Any given length of string that vibrates has therefore a certain amount of associated energy. The amount of energy for a length of string equal to one wavelength can be calculated by summing the energy content of all the particle oscillators. Energy (E) turns out to be given by the product of a numerical factor $2\pi^2$, the frequency (f) squared, the unit mass, symbolized by μ, the amplitude (A) squared, and the wavelength, symbolized by λ, or

$$E = 2\pi^2 f^2 \mu A^2 \lambda. \tag{28}$$

It should be noted that the energy is proportional to the square of the wave amplitude, a general result that applies to all types of waves.

When a sinusoidal wave is being generated by a source at some point on the string, the wave travels out along the string, carrying the energy E (of equation [28]) in each wavelength. The amount of energy passing any point of the string per unit time is the energy, E, multiplied by the number of cycles (complete vibrations) per second, f. Therefore, the power (P) passing a given point (and supplied by the source) is given by:

$$P = 2\pi^2 f^2 \mu A^2 \lambda f = 2\pi^2 f^2 \mu A^2 v. \tag{29}$$

Similar expressions can be derived for the power carried by other types of mechanical waves.

Energy transported by electromagnetic waves

Electromagnetic waves also transport energy. An electromagnetic wave contains an electric field and a magnetic field. Each of these fields can be represented by a vector, which is an arrow drawn in the direction of the field and having a length proportional to the strength of the field. The electric field vector and the magnetic field vector are perpendicular to each other and to the direction of propagation of the wave (see Figure 38). As the wave moves, the electric vector (and the magnetic vector) varies sinusoidally in magnitude. In other words, the electric field vibrates transversely about the direction of propagation just like the stretched string, with the electric field corresponding to the mechanical displacement. It is then not surprising that the energy content of the electromagnetic wave is proportional to the square of the electric field intensity because this is also the dependence on amplitude shown by mechanical waves.

The power carried by an electromagnetic wave across a unit area per unit time is proportional to the square of the electric field intensity averaged over a complete cycle. The energy is carried in free space with the speed of light. Free space is nondispersive because all frequencies travel at the same velocity. In a dispersive medium, the situation is more complicated. In general, energy cannot be carried by a single wave, but other waves with different frequencies must be present. This condition means that, in general, the energy is not carried with the phase velocity (see below *Pulses and wave trains*).

Momentum transport and radiation pressure. Because waves carry energy, it may be expected that they also transport momentum, which for a particle is the product of its mass and velocity. This is indeed the case for electromagnetic waves. If light is incident on a surface, the energy carried by the light wave can be absorbed. The light wave will also deliver momentum to the surface at a rate just equal to the rate of absorption of energy divided by the speed of light. The acquisition of momentum by the surface means that there is a pressure on it, called radiation pressure.

Radiation pressure

A simple explanation can be given for the origin of the radiation pressure. If the light wave is incident on an electron in an atom, the electron will start to oscillate with the electric field. The magnetic field that is also present in the light wave, however, will act on the moving electron charge, giving rise to a force in the direction of propagation. Absorption of light will therefore always be accompanied by radiation pressure.

For mechanical waves, the question of momentum transport is not as simple as for electromagnetic waves. Sound waves traveling in a gas do exert a pressure on a surface placed in the path approximately equal to the rate of energy flow divided by the wave speed, as in the case of electromagnetic waves. The exact pressure, however, depends on the properties of the gas. In the case of longitudinal waves in a bar of material obeying Hooke's law exactly, it can be concluded that these waves carry no momentum at all. Actual materials do not obey Hooke's law exactly, and the deviations from ideal behaviour can lead to momentum transport by the waves; however, the process is complicated.

SUPERPOSITION OF WAVES

The analysis of wave motion is not complete with a consideration of single waves only. Most of the rich wave phenomena in nature are the result of the interaction of many waves with each other. One of the most striking features of wave motion is that two waves can apparently pass through each other without being destroyed. Indeed, they emerge apparently unchanged. This phenomenon was noted by the 15th- and 16th-century artist and scientist Leonardo da Vinci for water waves. It is a particular case of a general principle called the principle of superposition. This principle states that the result of combining two physical quantities, such as the amplitudes of two vibrations, can be obtained by adding the two quantities as though they remained independent. The principle also implies that the combined effect can be reseparated into the two original effects. Systems for which the principle of superposition holds are said to be linear. A mechanical system obeying Hooke's law is a linear system. Linear systems are particularly amenable to mathematical analysis. Fortunately, a large fraction of the interesting physical systems are linear.

Physical systems amenable to mathematics

A pulse being sent down a string may be thought of as a wave consisting of only a few oscillations, perhaps only a single "hump." If another pulse is traveling down the string in the opposite direction, then during the time when the pulses are passing each other, the string at any instant will assume a shape that is a composite of the two overlapping pulses. The displacement of the string at any point will be the sum of the displacements that each of the two pulses would have produced in the absence of the other. This is the essence of the principle of superposition.

Standing waves. Another simple but important example of superposition occurs for two oppositely traveling waves on a stretched string. For simplicity, two transverse waves are taken such that they vibrate in the same plane (have the same polarization) and have the same frequency and maximum amplitude. Then the result of superimposing these two waves will be a standing wave—*i.e.*, a wave with stationary nodes. The wave form will not travel along the string even though (more rigorously, because) it is made up of two oppositely traveling waves.

This assertion can be demonstrated rather simply mathematically by considering the string as stretched along the x-axis and the vibrations to be in the y-direction. Then the sum of the two oppositely traveling waves can be written similarly to equation (25) as the sum of two terms:

$$Y = A \sin 2\pi\left(ft - \frac{x}{\lambda}\right) + A \sin 2\pi\left(ft + \frac{x}{\lambda}\right). \quad (30)$$

The two terms represent, respectively, waves traveling in the positive x-direction and the negative x-direction; the directions are different because of the different algebraic signs of the x/λ terms. The sine terms can be expanded by using the equalities:

$$\sin(a \pm b) = \sin a \cos b \pm \cos a \sin b. \quad (31)$$

Then by combining terms, the equation for the composite wave, consisting of the product of a sine term and a cosine term, follows, in which y is displacement, A is amplitude, x is position, λ is wavelength, f is frequency, and t is time: *i.e.*,

$$y = 2A \sin 2\pi ft \cos 2\pi\frac{x}{\lambda}. \quad (32)$$

A node is a position x on the string such that $\cos 2\pi(x/\lambda) = 0$. For these values of x, the displacement (y) will always be zero according to this equation (32). Therefore, the nodes do not move, and the wave appears to be stationary.

The standing wave described above can be produced by stretching a string between two walls of separation L (see Figure 39). Reflection at the ends of the string produces oppositely traveling waves. The string can be set into vibration such that the wavelength of the standing wave is just twice the length of the string. Then there is just one loop vibrating and the ends of the string are stationary.

Fundamental mode of vibration

This is called the fundamental mode of vibration. The fundamental mode has the longest wavelength of any of the standing waves that can be produced on the string. The physical condition for a standing wave on the string is just that the ends be stationary inasmuch as they are fastened to the walls. This restriction on the waves is called a boundary condition.

The boundary condition for the string can also be satisfied for a standing wave having a wavelength just equal to the length (L) of the string. Then there will be two loops and a node in the middle. By adding more loops and nodes, a series of standing waves can be constructed. These have wavelengths given by: $\lambda_1 = 2L$, $\lambda_2 = 2L/2$, $\lambda_3 = 2L/3$, etc., or the series can be written: $\lambda_1 = 2L$, $\lambda_2 = \lambda_1/2$, $\lambda_3 = \lambda_1/3$, etc. Because the wave velocity is the same for all these waves, the following relationships hold for the frequencies: $f_2 = 2f_1$, $f_3 = 3f_1$, $f_4 = 4f_1$, etc. The higher frequencies f_2, f_3, etc., are called harmonics of the fundamental frequency f_1. The standing waves comprising this sequence are also called the normal modes of the system. They are natural states of vibration, each involving the vibration of the entire string at the same frequency. Once the string is set into vibration in one of the normal modes, it will continue to vibrate in that mode. That the frequencies of the modes for the vibrating string follow a harmonic sequence is a special result of the assumption of a uniform, perfectly flexible string. The frequencies of the normal modes need not, and in general do not, have such a simple relationship.

Standing waves of tubes and drums

Another example of standing waves is afforded by longitudinal vibrations of the air in a tube. If one end of the tube is closed and one end open, then there are two different boundary conditions. For standing waves, the closed end requires a node, whereas the open end requires an antinode—*i.e.*, a point of maximum vibration. If the tube length is L, the first three modes will be given by $\lambda_1 = 4L$, $\lambda_2 = 4L/3$, $\lambda_3 = 4L/5$.

Standing waves may also be produced on membranes like the top of a kettledrum. Compared to the vibrating string, one more dimension has been added to the vibrator. The nodes on the membrane are lines that separate the vibrating zones. The boundary condition is that the edges be stationary. Again, a series of normal modes of oscillation can be identified. The fundamental involves the vibration of the entire membrane up and down, with the maximum amplitude in the centre and the edges being fixed.

Modulation of waves. Two waves having different frequencies can be superimposed to form a single wave. By a procedure similar to that applied in the derivation of equation (32), it can be shown that the composite wave

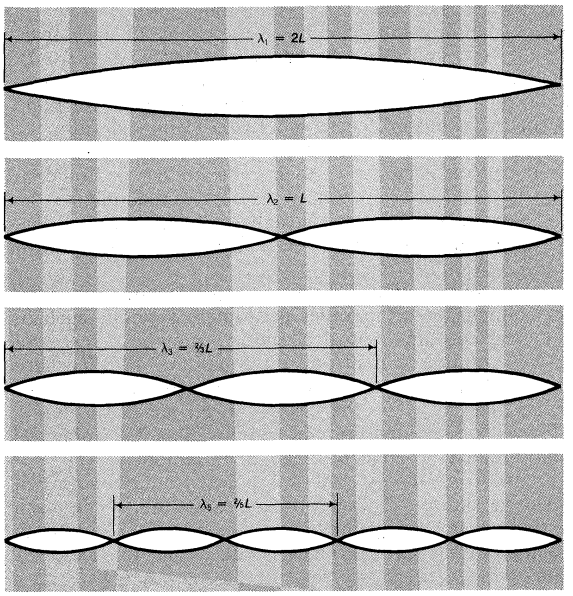


Figure 39: Standing waves on a string, showing simple modes. Wavelengths, λ, are expressed as multiples of the string length, L. The modes shown, from top to bottom, are the fundamental, 2nd, 3rd, and 5th.
From D.C. Miller, *The Science of Musical Sounds*

is represented by a product of two sinusoidal factors. One factor oscillates with the average of the two original frequencies. The other factor oscillates with half of the difference of the two frequencies. When two waves are combined that have almost the same frequency, the difference in frequency is small compared to the average, and the composite wave oscillates with nearly the average frequency except that the amplitude varies with the difference frequency. The latter is said to modulate the average frequency; *i.e.*, it determines the envelope (a curve that encloses the crests of all waves) of the oscillations. The relatively slow oscillations of the envelope are called beats (see Figure 40). Beats can be heard if two musicians attempt to sound the same musical note, but their instruments are tuned slightly differently.

From *Waves* (Berkeley Physics Course, vol. 3); McGraw-Hill Book Co., New York, 1968

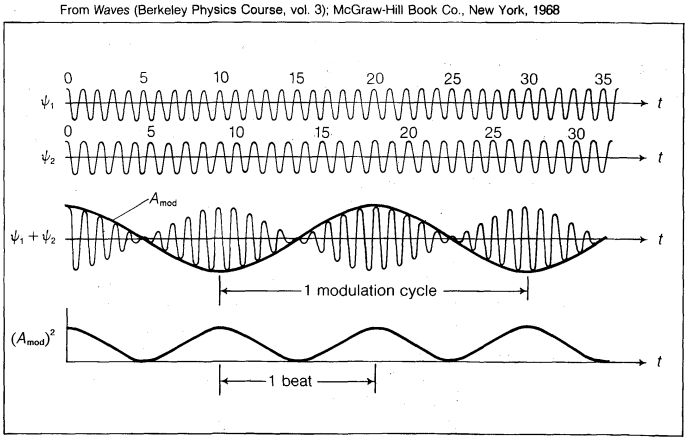


Figure 40: Beats produced by addition of waves ψ_1 and ψ_2 with frequency ratio $f_1/f_2 = 10/9$. The superposition $\psi_1 + \psi_2$ has a slowly varying amplitude A_{mod}. The intensity (loudness) of the beats has amplitude $(A_{mod})^2$, and the beat frequency is twice the modulation frequency (see text).

Pulses and wave trains. The waves encountered in nature always have a finite extension in space along the direction of propagation. The wave motion consists of a certain number of undulations called a wave train. If the wave train is short, it may be called a pulse. The pulse can consist of any disturbance that the medium can support; it need not be of any particular shape.

A pulse or wave train passes a given point in a finite time. It therefore cannot consist of a pure vibration of a

single frequency. Other vibrations of different frequencies must be superimposed to cause the pulse to begin at a certain time and to end at a certain time. A pulse consists, then, of a group of superimposed vibrations with different frequencies, amplitudes, and phases. The principle of superposition assures that any pulse can be constructed out of component vibrations. This idea was applied by Fourier to the analysis of periodic motion—*i.e.*, motion that repeats itself (see Figure 41). Any periodic motion can be represented by a Fourier series in which, aside from a constant, each term is a sine (or cosine) function of time. Each sine term varies with a single frequency that is either a fundamental frequency or integral multiples thereof. If a periodic function of time is symbolized $f(t)$, then the Fourier series for $f(t)$ may be written as the sum of sine and cosine terms: $f(t) = a_0 + a_1 \cos 2\pi(ft) + b_1 \sin 2\pi(ft) + a_2 \cos 2\pi(2ft) + b_2 \sin 2\pi(2ft) + a_3 \cos 2\pi(3ft) + b_3 \sin 2\pi(3ft) + \ldots$. The as and bs are constants (independent of time). They are the amplitudes of the oscillations and can be determined by a procedure due to Fourier. The terms of the series involve higher and higher frequencies. In practice, the highest frequency that must be included to approximate a pulse to a given accuracy depends on the fastest variation of the pulse itself. Considering the wave form in space, sharp corners mean that high frequencies are present.

Analysis of periodic motion

Another important question concerns the velocity with which a pulse or wave train travels through the medium. Because the pulse consists of a superposition of two or more waves of different frequency, the phase velocity of any one of these waves is not the velocity of the group of waves or pulse. An example of the addition of two traveling waves that have slightly different frequencies was described earlier in the section *Modulation of waves*. The result of superposition is a slow modulation of a nearly pure oscillation with the average frequency. The modulation produces beats; the envelope of the vibrations goes up and down with the beat frequency. Each of the beats can be considered to be a wave train. The velocity of the beat or wave train is commonly called the group velocity (see Figure 42).

Group velocity

At this point, another definition is required—namely, the wave number k, which is simply the reciprocal of the wavelength; *i.e.*, $k = 1/\lambda$. If the frequencies of the two waves being added are f_1 and f_2, and the wave numbers are, respectively, k_1 and k_2, then the group velocity (v_g) is given by the difference in frequency divided by the difference in wave number: $v_g = (f_2 - f_1)/(k_2 - k_1)$. This equation may be generalized by taking the limit as the difference in the fs (and the ks) approach zero. Then the group velocity is the slope of the plot, frequency versus wave number.

From *Waves* (Berkeley Physics Course, vol. 3); McGraw-Hill Book Co., New York, 1968

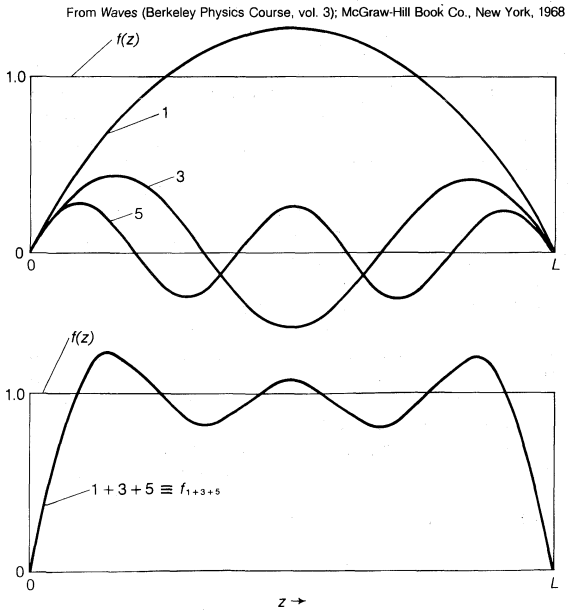


Figure 41: (Top) Fourier analysis of a square wave $f(z)$ showing the first three components, labeled 1, 3, 5. Even terms are zero in this example. (Bottom) Square wave $f(z)$ and the sum of the first three Fourier components.

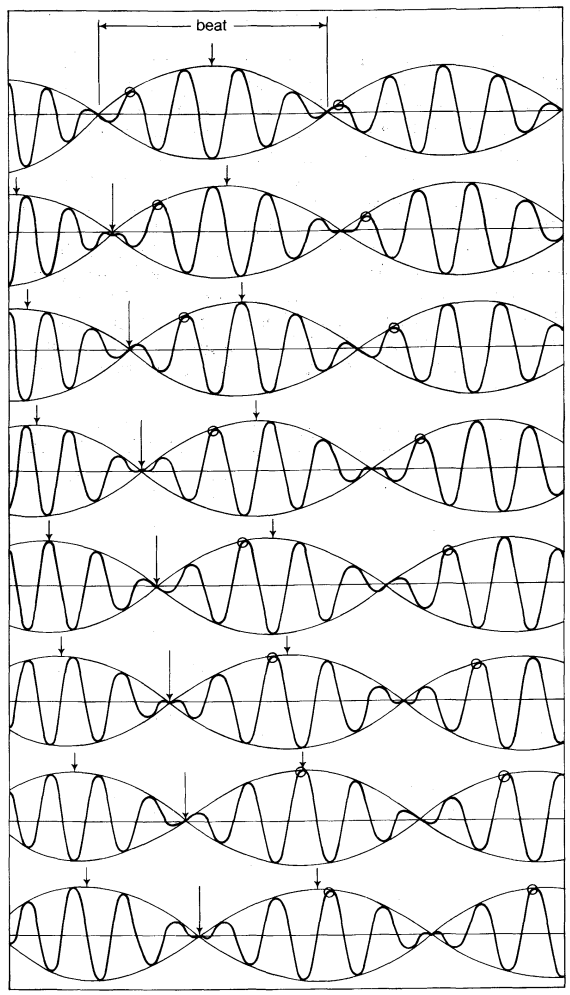


Figure 42: *Illustration of group velocity.* Arrows follow beats, which travel at group velocity. Circles follow wave crests, which travel at average phase velocity.
From *Waves* (Berkeley Physics Course, vol. 3); McGraw-Hill Book Co., New York, 1968

The significance of group velocity is that it gives the speed with which energy is transported through the medium. When an electromagnetic wave traverses certain media, the phase velocity can be apparently greater than the speed of light because of the phase shift that has been produced. The energy in the wave is transported, however, with the group velocity that is always less than the phase velocity and less than the velocity of light. Hence, no contradiction of Einstein's theory of relativity occurs.

Another consequence of the fact that pulses consist of the superposition of more than one wave should be noted. When a pulse traverses a dispersive medium, the component waves will travel at different speeds. As a result, the relative phases of the component waves will change, and the pulse will change shape as it travels.

THE BEHAVIOUR OF WAVES AT BOUNDARIES

Reflection and transmission. If a traveling wave encounters a boundary in a medium, part or all of the wave will be reflected. The boundary might be an interface between the medium and empty space or an interface between two media. The example of standing waves produced on a string stretched between two walls implicitly employs the reflection of the traveling waves at the walls. The walls represent an infinitely dense medium. In this case, the entire wave is reflected; *i.e.*, the direction of the wave is reversed at the interface. Another important change in the wave, however, takes place. If a portion of the wave involving upward displacement of the string encounters the wall, it will leave the wall traveling backward but also

with downward displacement. The wave is inverted as a result of the reflection.

Inversion of waves

The inversion of the wave can be thought of in the following manner. A second wave may be imagined as entering the string from within the wall. Furthermore, this wave arrives at the end of the string with exactly the opposite vertical displacement that the original wave has. Then the two waves will cancel each other at the boundary, producing the desired node where the string is fastened. The wave that came from within the wall will continue to travel down the string and will constitute the reflected wave (see Figure 43). If the traveling wave is sinusoidal, then the inversion at the boundary means a phase change of 180°, or π radians. The phase is the argument of the sine function. Adding 180° to the argument produces an inversion because of the trigonometric relationship $\sin(\theta + \pi) = -\sin\theta$.

From A.P. French, *Vibration and Waves*, © 1966, 1971 by MIT, and F.W. Sears and M.W. Zemansky, *University Physics*, 3rd ed., part 1 (1963); Addison-Wesley, Reading, Mass.

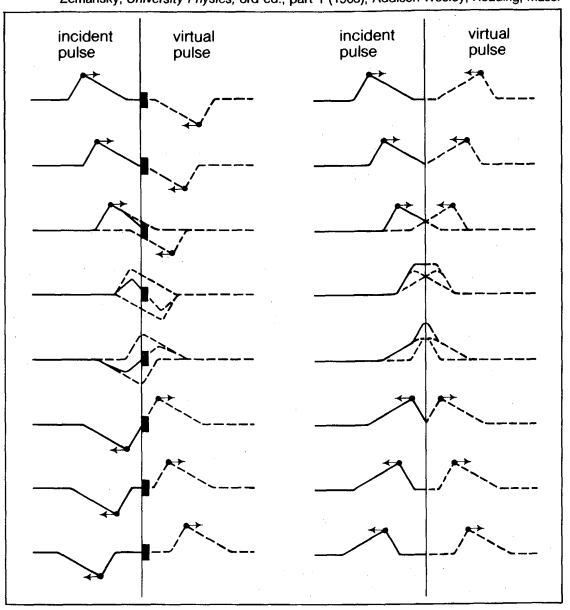


Figure 43: *Reflection of a pulse traveling on a string.* (Left) Reflection at a fixed end, resulting in inversion of pulse. The reflection can be constructed by imagining a virtual pulse coming from the opposite direction and superimposing the two pulses. (Right) Reflection at a free end, producing no inversion (see text).

The other extreme in boundary conditions for a string is a free end. Then the wave will also be reflected at the free-end boundary, but the displacement will not be inverted, and there will be no phase change. The free end requires an antinode. Again, another wave can be imagined arriving at the free end of the string from the opposite direction to the original wave, except that this time the waves arrive at the boundary with the same vertical displacement.

The general case involves a boundary in front of a second medium, which is neither infinitely dense nor of zero density but in between. In this situation, only part of the wave will be reflected; *i.e.*, the amplitude of the reflected wave will be less than that of the incident wave. If the wave travels on a string having a certain density (mass per unit length) and encounters a boundary marking a change of the string to greater density, then the reflected wave will be inverted, as in the case of a wall. On the other hand, if the second string beyond the boundary is less dense, there will be no inversion of the reflected wave.

If the wave is only partially reflected at the boundary, the conservation of energy requires that the remaining wave energy be transmitted, and therefore some of the wave travels into the second medium. The transmitted wave will not be inverted; the wavelength will be different, however, because the wave velocity will be different in a medium of different density.

Refraction. The conditions for reflection and transmission of waves at a boundary can be generalized from waves on a string to waves in extended media. In the case of sound waves or light waves, the waves must be described in three dimensions. A simple kind of such waves is a plane wave, in which the wave front is a plane or, equivalently, the wave is propagating in a single direction. When a plane wave encounters a boundary between two media, reflection and transmission will occur in a manner analogous to the case of a wave on a string. There is an additional consideration, however, because the plane wave may strike the interface at an angle.

For reflection of a plane wave from a plane interface, the angle between the direction of propagation of the incident wave and the interface is equal to the angle between the direction of propagation of the reflected wave and the interface. In other words, the component of the wave velocity perpendicular to the interface is reversed, whereas the component parallel to the interface is unchanged. The transmitted wave will be deviated in direction (refracted) from the direction of the incident wave. The refraction happens because the wave velocity is different in the two media. In the case of a wave leaving one medium and entering a second medium where the wave velocity is less, the wave slows down on entering the second medium and the wave front changes direction. The new wave direction is more nearly perpendicular to the interface. An analogous effect takes place when a row of marching soldiers leaves a paved surface and enters a muddy field obliquely.

Law of refraction

The law of refraction can be obtained from simple geometric considerations. If i is the angle between the direction of travel of the incident wave and the normal (perpendicular) to the interface, r is the angle between the direction of travel of the refracted wave and the normal to the interface, and v_1 is the wave velocity in the first medium and v_2 the wave velocity in the second medium, then the relation between these quantities states that the ratio of the sines of the angles of incidence and refraction is equal to the ratio of velocities in the respective media, or $\sin i/\sin r = v_1/v_2$.

In the special case of an electromagnetic wave going from free space ($v_1 = c$) into a medium such as glass, then the ratio v_1/v_2 is the index of refraction (see above *Velocity of waves*).

If a wave approaches an interface from within a slower medium, the wave is refracted away from the normal to the interface; *i.e.*, angle r (now in the faster medium) is greater than angle i. An interesting phenomenon occurs when the angle of incidence i has a certain critical value such that $\sin r = (v_2/v_1)\sin i = 1$, and, as a consequence, the refracted wave has a direction along the interface. For larger angles of incidence, the incident wave will undergo total reflection, and there will be no transmitted wave.

INTERFERENCE AND DIFFRACTION

When a wave impinges on a barrier with holes in it, the wave will emerge from the holes and spread out on the other side of the barrier. Each point on the wave front at a hole becomes a centre for an outgoing spherical wavelet. The holes provide secondary wave sources that are coherent. The condition applying to coherent sources is that they have a definite phase relationship to each other. If coherent sources are in phase at a particular instant, they will always be in phase. If they are out of phase by a certain amount, that amount will remain constant.

The process of interference

Interference is the process of addition of waves that occupy the same space. A simple interference pattern can be produced by allowing a plane light wave to strike a barrier with two long parallel slits in it (see Figure 44). The secondary waves from the slits will strike a screen placed parallel to the barrier and fall on top of one another. If the distance from one slit to any point on the screen is d_1, and from the other slit to the same point on the screen is d_2, and if their difference $(d_2 - d_1)$ is an integral multiple of the wavelength, then the waves from the two slits will add constructively, and the point on the screen will be bright. The condition for constructive interference can be written: $d_2 - d_1 = 0, \lambda, 2\lambda, 3\lambda, \ldots$. Conversely, the condition for destructive interference producing a dark spot on the screen is: $d_2 - d_1 = \lambda/2, 3\lambda/2, 5\lambda/2, \ldots$. The pattern on the screen will consist of alternately bright and dark lines parallel to the slits.

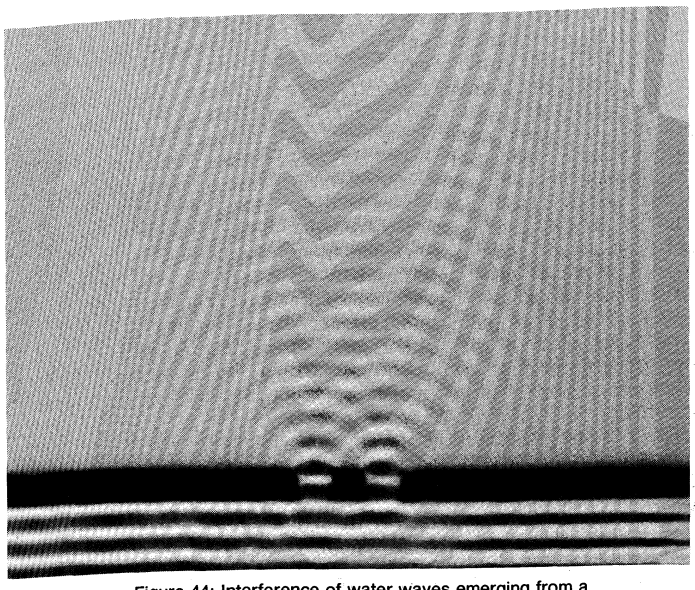


Figure 44: Interference of water waves emerging from a double slit. The wave fronts approach the slits from the bottom of the figure.
From *PSSC Physics*, D.C. Heath and Company, Lexington, Mass., 1965

The set of double slits is a typical example of coherent sources in that the light actually comes from a single original source. A notable exception is the laser, which produces coherent radiation from a tremendous number of different atomic sources. Coherence is accomplished by storing energy in the atoms of the laser and then triggering the emission of light by a large number of the atoms at the same time by a process called stimulated emission.

Special attention must be given to the intensity of the wave produced by two interfering waves. The amplitude of such a resultant wave is equal to the sum of the amplitudes of the two interfering waves. It was stated above that the intensity (energy per unit area per second) of a wave is given by the square of its amplitude. Thus, if two coherent waves that are in phase interfere and if the waves have equal amplitudes, then the intensity of the resultant wave is four times that of each of the original waves, not twice. If a large number of coherent waves interfere, then the amplitudes of the waves must be summed before squaring. Thus, for waves in phase, the resultant intensity is considerably larger than the mere sum of their intensities.

Coherent waves

Interference of waves from different parts of the same source is called diffraction. A simple example is a single slit in a barrier. Points on the wave front at the slit serve as sources of new wavelets. The overall effect is that the wave spreads out in a pattern beyond the slit. If the radiation is coherent, then an interference pattern (called a diffraction pattern in this context) will be produced on a screen placed beyond the slit. Because the slit has a finite width, it can be divided up into a number of imaginary narrow slits for the purpose of analysis. As in the case of the double slit, the differing distances from the imaginary slits to points on the screen produce light and dark regions—*i.e.*, the diffraction pattern. (W.J.)

WATER WAVES

There are two main categories of water waves. Very small waves, with wavelengths (distance from crest to crest) of less than a centimetre or two, are propagated by a force arising predominantly from surface tension. These are commonly called ripples. All waves with longer wavelengths owe their properties mainly to the force of gravity. In addition to this principal subdivision, gravity waves occur in two distinct regimes. Waves in deep water are different from those in shallow water, and different theories for the two cases have been developed.

Ripples and gravity waves

What is meant by shallow water in this context depends upon the wavelength. When the water is deeper than about half the wavelength of the waves, the regime that obtains is that for deep water. Shallow-water effects do not begin to become important until the depth is less than about one-quarter of the wavelength. An enormous range of wavelength is to be found in nature. The waves that constitute the tides (see above) over the open seas, for example, may be thousands of kilometres between crests and will therefore cross the deepest oceans as shallow-water waves. On the other hand, the small gravity waves that occur on reservoirs and lakes will travel as deep-water waves over water that is only 30 to 60 centimetres deep.

Shallow-water waves. The study of waves can best be considered by treating first the case of waves in shallow water. In shallow water, the motion of the water particles is predominantly backward and forward. The crests and troughs of the waves are formed because of differences in horizontal displacement; crests occur where the water particles are crowded together and troughs where they are more widely spaced. The theory of shallow-water waves is most easily understood in terms of the tidal bore.

If a system has an adverse current that is equal to the velocity of the bore (Figure 45), then the effect would be to bring the wave to rest. The problem is one of a steady state, in which water arrives at the wave on one side with velocity c and leaves on the other side with velocity $(c-u)$. Two equations are required for determination of c and u.

One of the required equations can be obtained from the condition of continuity—namely, that as much water must leave the wave as enters it. This means that the water depth multiplied by the velocity on arrival must equal the water depth multiplied by the velocity on departure. Therefore, from Figure 45:

$$Hc = (H+h)(c-u)$$

or

$$hc = u(H+h),$$

in which H and $(H+h)$ are the water depths and c and $(c-u)$ are the velocities.

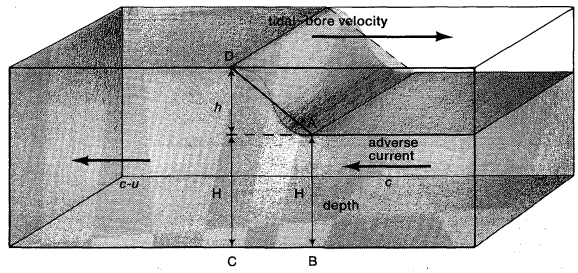


Figure 45: Relation of adverse current to the velocity of the tidal bore (see text).

The second equation can be obtained by considering the energy of the water particles on the surface as they mount the wave and pass on. The surface is a region of constant atmospheric pressure, and there are no pressure gradients to be taken into consideration to complicate matters. The kinetic energy lost by the particles as they mount the wave (AD) will simply be equal to the work done against gravity in rising to the height, h, of the wave. The particles of mass m arrive with a kinetic energy of $\frac{1}{2}mc^2$ and leave with $\frac{1}{2}m(c-u)^2$. Thus,

$$\tfrac{1}{2}mc^2 = \tfrac{1}{2}m(c-u)^2 + mgh$$

or

$$cu = gh + \frac{u^2}{2},$$

in which m is particle mass, h is height, c and $(c-u)$ are velocities, and g is the acceleration due to gravity. Substituting here the value of u determined from the continuity equation, namely, $u = hc/(H+h)$, and assuming that the height of the bore, h, is small compared with the depth of the water, H, the expression reduces to $c = \sqrt{gH}$, in which c is wave velocity, H is water depth, and g is the acceleration due to gravity.

This expression is a most important one, and it appears continuously in problems of hydrodynamics; the main characteristics of waves in shallow water may be deduced

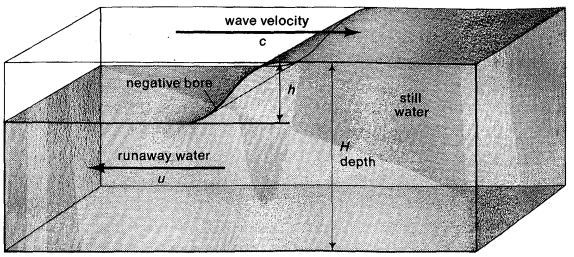


Figure 46: Negative tidal bore (see text).

from it. Instead of a tongue of water moving up a river, the case of a canal from which water is suddenly released by the collapse of a lock gate might be considered. This situation would create a negative bore. A wave would run up the canal separating the still water on one side from the moving water that is running away on the other side (Figure 46).

A calculation similar to that just performed for the positive bore shows that the wave would move up the canal with approximately the same velocity, the positive bore actually moving slightly faster.

Wave-tank studies

Solitary waves. If a single negative bore immediately follows an equal positive bore, a solitary wave is formed (Figure 47). The water is still before and after the passage of the wave, but it moves forward under the wave as it passes. If the positive bore follows the negative one instead of leading it, a solitary trough will traverse the water as shown in Figure 48. The water under the trough moves in the opposite direction to the wave.

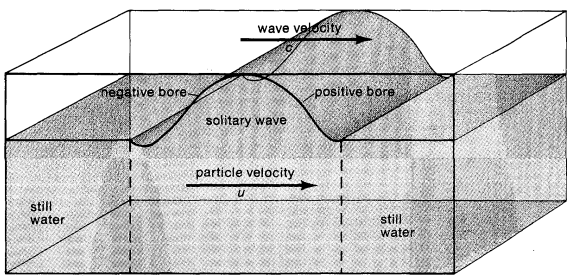


Figure 47: Negative bore following an equal positive bore, forming a solitary wave.

Solitary waves may be studied in a simple wave tank, consisting of a wooden box about 180 centimetres long and roughly 30 centimetres wide and deep, made watertight by lining it with a plastic sheet. A paddle that nearly fits the cross section of the box may be used to generate waves. It is easy to cause either solitary crests or troughs to pass along the tank. Dust particles on the bottom will be carried forward with the wave when a solitary crest passes over it and then will become stationary once again after the wave has passed. The wave will be reflected from the end of the tank, and the dust particles will be carried back to their initial positions as the reflected wave passes backward over them. The reverse happens when a solitary trough travels along the tank. The particle motion is opposite to the direction in which the wave travels as the trough passes over them, and they become stationary again when it has gone.

Effects of water depth. As waves pass into shallower water, their speed diminishes. The energy they contain becomes concentrated in a smaller area and, as a consequence, the height of the waves increases. A positive bore always travels faster than a negative bore of the same height. Thus a train of waves in shallow water, made up of both positive and negative elements, will not be propagated without change of shape. Crests will travel faster than troughs and ultimately will overtake them; the front of the wave will steepen, and finally the wave will break. The breaking of large waves on a steeply shelving beach provides some of the most spectacular—and some of the most destructive—effects produced by waves on the sea.

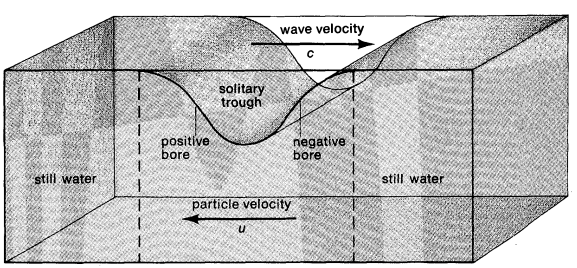


Figure 48: Positive bore following an equal negative bore, forming a solitary trough.

There are two kinds of breakers, spilling breakers and plunging breakers. Spilling breakers are formed on gently shelving beaches, where the overtaking of the troughs by the crests occurs slowly. The crests then slide down the forward slope. This type of breaker is comparatively harmless.

When the beach shelves steeply, the process occurs very rapidly and the crest plunges forward into the trough in front. A large quantity of air can be trapped within the wave, and, when the mass of water strikes an obstacle, its momentum causes the air to be compressed. It is this compressed air within the breaker, and the air within cracks and crevices in seawalls and breakwaters, that causes extensive damage.

There are two principal effects resulting from the decrease in the speed of waves as they enter shallow water. One of these, as previously mentioned, is the concentration of wave energy into a smaller area, and the resulting increase of wave height. The most striking effects occur with so-called tidal waves. These waves are not caused by the tides but rather by movement of the ocean floor. Being of seismic origin, they are properly called tsunamis, or seismic sea waves. These waves possess very long wavelengths and travel across the deepest oceans as shallow-water waves with velocities of more than 640 kilometres per hour.

Seismic sea waves

The other principal effect of the retardation of water waves as they enter shallow water is deflection from their normal path, or refraction. If a wave approaches a gently shelving beach obliquely, the shoreward end of the wave is slowed up and the wave swings around. It therefore tends to approach the shore nearly at right angles, whatever the direction of travel might have been at sea. Also, the wave energy tends to be concentrated upon capes and promontories, if they extend seaward as shallows.

The passage of atmospheric depressions can also lead to the generation of waves. If the atmospheric pressure is suddenly reduced over a certain area, the level of the sea rises to compensate, similar to the action of mercury in a barometer. Waves spread out from the rising region at a velocity equal to $\sqrt{gH}$. If the depression travels forward at this speed, there is a kind of resonance effect, and the height of the bulge on the surface continually increases. This is one cause of tidal surges, which from time to time lead to very high tides and inundate low-lying coastal areas. They are likely to occur in the North Sea, where flooding of the eastern counties of England and parts of the Netherland coast results. More or less inland waters, such as the North Sea, also have natural periods of oscillation that can be excited by atmospheric disturbances that are traveling at appropriate speeds; these effects also lead to tidal surges.

Deep-water waves. *Wave velocity.* With respect to the regime of waves on deep water, it is best to consider first a set of standing waves formed by the interference of two traveling waves going in opposite directions. As shown in Figure 49, where the wave motion is between reflecting walls AB and CD, the motion of water particles is up and down at the walls but is horizontal across the plane EF.

The effect of the oscillation is that the dotted region is temporarily void of water, while the crosshatched area is occupied. The increase in potential energy will be proportional to the wavelength and the square of the displacement, or $\lambda g x^2$, in which λ is wavelength, x is the displacement, and g is the acceleration due to gravity. (The mass of water raised will vary as λx, and the height it is lifted will vary directly with x.) The depth to which motion will be communicated to the water by the waves is unknown;

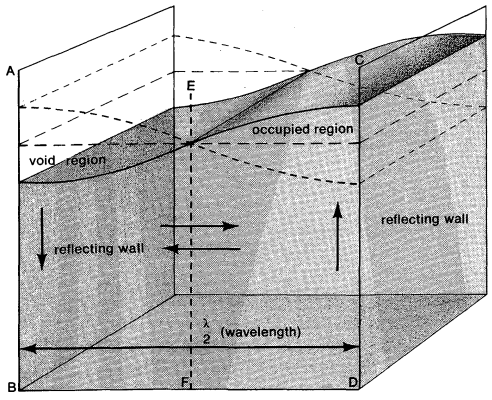


Figure 49: Standing waves formed by the interference of two traveling waves. Arrows indicate motion of water particles.

however, assuming similarity of figures, as the wavelength λ is varied, the depth to which the motion goes will vary as λ. The mass of water taking part in the oscillation will thus vary as λ^2, because the distance AB and CD will also vary as λ. Thus the periodic time of oscillation T will vary as $\sqrt{\lambda/g}$; and the velocity of the component traveling waves will vary as $\sqrt{\lambda g}$, being equal to λ/T.

This is probably the simplest statement of the fact that the velocity of waves on deep water varies as the square root of the wavelength, which fact is sufficient for all that follows.

Motion of water particles

The motions shown in Figure 49 would be reproduced if the water particles moved in circular or elliptical paths under the influence of one traveling wave only (Figure 50). Elliptical motion occurs in the intermediate regime between deep water and shallow water, when the presence of the bottom begins to restrict the vertical movement. If the device of the adverse current is again used, and such a traveling wave is considered to have been brought to rest, then it is clear that in the case of circular paths for the particles, the wave form traced out must be a trochoid, the path of a point on the spoke of a wheel that rolls along horizontally underneath a horizontal straight line. From this it is easy to deduce the expression for the velocity of the waves.

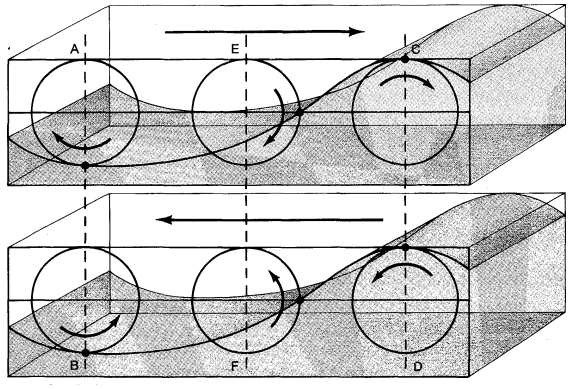


Figure 50: Water particle motion in waves traveling in either direction.

If an adverse current is imposed and the waves are brought to rest, the water particles will slide over the waves as if they were perfectly smooth. Let the radius of the wheel, rolling under the line XY (Figure 51), be K, and let P be the point on the spoke a distance A from the centre O. Let the wheel turn with angular velocity ω.

When P was at its highest point, P′, it lay at a distance (K − A) below the line XY, and the wheel was turning instantaneously about the point O′. The velocity of P at any other point is:

$$CP\omega = \omega\sqrt{K^2 + A^2 - 2KA\cos\theta}.$$

Now P is A $(1 - \cos\theta)$ below P′, and $K\omega^2 = g$, in which g is the acceleration due to gravity.

The wavelength will be the distance rolled by the wheel in one revolution, so that $\lambda = 2\pi K$, in which λ is again the wavelength and π is the constant 3.1416.

The velocity of the waves will be that of the adverse current, which, in turn, will be equal to the velocity of the wheel, namely:

$$c = K\omega = \sqrt{gK} = \sqrt{\frac{\lambda g}{2\pi}}.$$

As with shallow-water waves, the water particles move forward with the waves on the crests and in the opposite direction in the troughs.

Just as a large number of the properties of waves in shallow water were summarized by the equation $c = \sqrt{gH}$, so many of the properties of deep-water waves can be deduced from the equation $c = \sqrt{\lambda g/2\pi}$. The most important properties arise from the fact that the velocity varies as the square root of the wavelength.

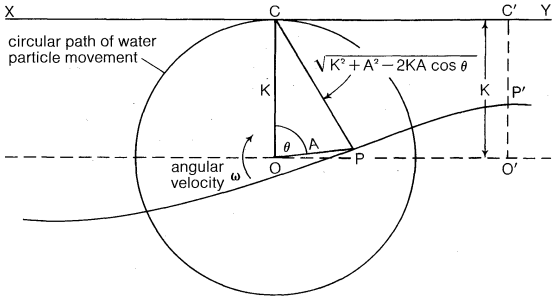


Figure 51: Trochoidal wave and particle motion relations (see text).

Group velocity. The phenomenon of group velocity as distinct from wave velocity is a case in point. If waves occur in a group, as they normally do, the individual waves travel with the velocity $c = \sqrt{\lambda g/2\pi}$, but those in the front of the group die out, whereas those at the back grow. The group as a whole thus travels more slowly than the waves of which it is composed. The waves, in fact, are continuously passing through the group, forming at the back and dying out in front.

The importance of the group velocity arises from the fact that the energy of the waves is related to the velocity of the group, and not the velocity of the waves. It has many important applications. For gravity waves, the group velocity is equal to one-half of the wave velocity. The reason for this is as follows. Consider two trains of waves traveling in the same direction but of slightly different wavelengths. At some point the crest of one wave will fall upon the trough of another, and the two waves will neutralize one another, as at A and C in Figure 52. At other points, such as B, they will reinforce each other. If the wavelengths of the two waves differ by $d\lambda$ and they travel with velocities differing by dc, the point of coincidence will fall back one wavelength when the faster, longer wave passes the slower, shorter one by a distance $d\lambda$—*i.e.*, in a time $d\lambda/dc$. Thus the group velocity is

$$c - \lambda\frac{dc}{d\lambda}.$$

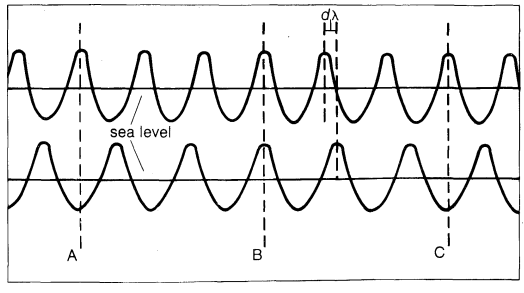


Figure 52: Waves of different wavelengths moving in the same direction. Neutralization occurs at A and C because the crests of waves in the upper wave train coincide with troughs in the lower train. Reinforcement occurs at B because two wave crests are aligned.

However, for gravity waves on deep water,

$$c = \sqrt{\frac{\lambda g}{2\pi}}$$

and

$$\lambda \frac{dc}{d\lambda} = \frac{\lambda}{2}\sqrt{\frac{g}{2\pi\lambda}} = \frac{1}{2}\sqrt{\frac{\lambda g}{2\pi}} = \frac{c}{2},$$

and the group velocity of these waves is equal to half the wave velocity.

Wave wakes

An interesting consequence of the difference between wave and group velocity is to be found in the wave wakes created by vessels moving through water. These take the form of a series of more-or-less parallel waves that follow the vessel astern and keep pace with it, and an arrowhead formation with the vessel at the apex of the arrowhead (Figure 53). The angle between the arms of the arrow is always 39° regardless of the speed of the vessel if it does not plane.

As a vessel plows its way through the water and generates waves, those that keep pace with it are continually being augmented, while others travel away and disappear. Those waves that travel astern possess the same velocity as the vessel, but they form a group, and their group velocity is only half that of the waves or the vessel. The group of stern waves in deep water, therefore, stretches back from the vessel for a distance equal to half that which the vessel has traversed; this group of waves gradually, of course, becomes dissipated.

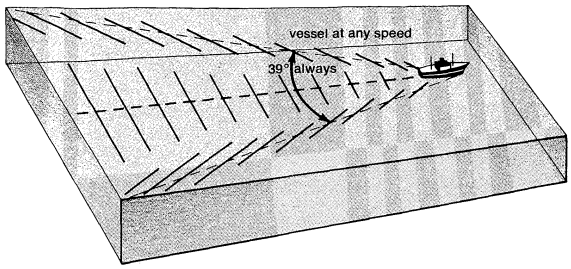


Figure 53: *Waves created by the movement of a vessel in deep water.*
One set of waves follows directly astern, while the other set of waves occurs inside an arrowhead formation, as shown.

In shallow water a stern surge follows the ship at a speed of $\sqrt{gH}$. The vessel is practically limited to this speed unless it can be lifted out of the water and made to plane. If its speed can once be increased beyond this value, however, then no stern surge can form because no wave can travel at a speed greater than this limiting value. Wave resistance, which ordinarily forms the major part of the drag on a vessel, is practically eliminated. (This scientific discovery was first made when a horse bolted while towing a barge on the Glasgow and Ardrossan Canal.) The principle underlies the performance of speedboats, certain sailing craft such as catamarans, and the surf rider.

(R.A.R.T./Ed.)

BALLISTICS

Primary aspects of the field

Ballistics is the subject concerned with the phenomena and laws of projectiles, their propulsion, flight, and impact. These three aspects of the field can be studied separately when applied to guns and are called individually interior (or internal), exterior (or external), and terminal ballistics, respectively, and a wide range of phenomena is subsumed under them. Rocket propulsion in free flight involves the special problem of interior and exterior ballistics at the same time because the thrust of the projectile and its aerodynamics are interdependent during flight. Ballistics beyond the atmosphere (that is, space flight) is a subcategory of exterior ballistics known as geoballistics. The ballistics of cartridge-actuated devices concerns the use of propellants to move heavy loads short distances, or to work against large resistance in short periods of time. Still another branch of study is hydroballistics, the study of projectiles moving through the water. The general effects treated in ballistics are reproducible muzzle velocity; the lowest gun pressure that will produce the desired velocity within a given length of the gun barrel; the stability of the projectile in flight; long and reproducible trajectories; and maximum interaction with a selected target.

HISTORY OF BALLISTICS THEORY

Ballistics theory depends on many scientific disciplines. Interior ballistics involves the application of theories of chemical combustion, compressible gas dynamics, and mechanics; exterior ballistics theory is built on a knowledge of mechanics, aerodynamics, and, sometimes, geodesy—the shape and physical character of the Earth; terminal ballistics draws on the theories of explosive technology, metallurgy, solid-state physics, and mechanics. Because all ballistic phenomena occur at extremely high rates of speed, and a great complexity of interactions between them develops, the fastest computers, microsecond photography, supersonic wind tunnels, and pressure gauges capable of measuring several thousand megabars (a megabar is equivalent to about one million atmospheres of pressure) in microseconds are required in ballistic study.

Ballistics problems have always challenged the most powerful and sophisticated scientific techniques available and have stimulated new ones. Among the prominent early contributors to ballistics were Leonardo da Vinci, Galileo, and Newton.

The first hurled object was a problem in ballistics. The advantages of being able to hurl farther or harder led to devices such as slings and throwing sticks for extending the arm. Further increases in effectiveness were achieved with the bow, which combined increased mechanical advantage with the accuracy derived from a more stable platform than provided by the human arm in the act of throwing. Arrows could reach the parapets of a defensive structure, but breaching the walls or causing internal structural damage necessitated the invention of the ballista, from which ballistics derives its name.

Invention of the ballista

The intimate relationship between the human body and ballistics in propulsion by throwing or using the ordinary bow makes it doubtful that hunters and warriors felt a need for an intellectual ballistics principle. With the advent of the ballista, however, the intimate sense of feel and the integration of arms and eyes are absent. The ballista propelled missiles in one of three ways: through torque, equipoise dropping of weights, or flexure. Its capacity was determined by the tension, lever, and weight values applied to the device. By the time of Leonardo da Vinci, the principles of levers were well known, so that methods of increasing initial velocity of various weights could be understood characteristically, if not analytically. Because Leonardo and his contemporaries had mistaken views of the magnitude of gravitational effects, the effective use of the ballista required experiments to test the range of each missile and each variation in the ballista's design and setting; a completely empirical set of data, therefore, had to be obtained for accurate use of the machine.

The work of Leonardo da Vinci

During the centuries preceding the introduction of guns, then, the principles of ballistics included the application of the principles of levers to ballista and of those of geometry to aiming.

It might also be noted that although Leonardo had an excellent knowledge of static mechanics, his version of dynamics was empirical and theoretically incorrect. He wrongly deduced that force is directly proportional to velocity, rather than a change of velocity with time; that is to say, he did not seem to comprehend that force is directly proportional to acceleration. Thus, he could have no correct theory of interior or exterior ballistics.

Until the first half of the 16th century it was assumed that a projectile traveled in a straight line to the zenith of

its trajectory, and then fell on the other leg of an isosceles triangle. It was observed that flight was curved, but the principle applied was simply to approximate the curvature of flight with a circular arc in the upper portion of flight. No concept of drag is included in this description, and it was 1740 before Benjamin Robins, an English mathematician and engineer, first measured with any accuracy the loss of projectile velocity in flight. His results were incredible to scientists of that time. It had been thought that the influence of drag was about the same as that of gravity, but experiments have shown it to be many times greater. The measuring device that Robins invented was a ballistic pendulum that operated on the principle that the swing of a pendulum of large mass, when struck by a projectile with less mass, can be used to calculate the impact velocity.

Contribution of Benjamin Robins

Muscle and gravity were eventually replaced by gunpowder for propelling projectiles, but there is dispute over its origin. It could have been imported from China, though Marco Polo never mentioned such a curiosity in his writings; or it could have been invented in Italy because chemicals of the proper quality were mined there; or it could have been discovered by Roger Bacon. Whatever the source, gunpowder provided a human experience with dynamical forces far, far beyond the effects of muscle, levers, or weights. There were no theoretical tools in chemistry, physics, or mathematics with which to begin a theory of gun ballistics, but the history of science since that time is replete with mathematical, physical, and chemical discoveries that were made in a search for solutions to problems in ballistics.

PRINCIPLES OF BALLISTICS

Interior ballistics. *General considerations.* It is useful to begin with the simplest ballistic case—a bullet fired from a gun. The bullet is driven along the barrel by the pressure of the gas generated when the gunpowder propellant burns at an extremely high rate in a chamber (called the breech) at the rear of the gun. The process is analogous to that causing the motion of a piston within a cylinder of an internal-combustion engine, the projectile being the piston and the gun barrel being the cylinder. The gross variables of this process are the weight and cross-sectional piston area of the projectile, the gas pressure applied to its base, and the length of travel in the gun along which the propellant gases can exert force. If a mathematical relationship can be found among all these factors, then the ballistic behaviour of any bullet fired from any gun using any propellant can be approximated.

In the broadest sense, the classical laws of motion are applicable to any moving object: the force ($\boldsymbol{F}$), causing the mass to accelerate, is equal to the product of the mass (m) and the acceleration ($\boldsymbol{a}$); that is, $\boldsymbol{F} = m\boldsymbol{a}$. Since the projectile's mass is usually constant throughout the time interval defined by interior ballistic period, its acceleration will result from the applied force. That force is the sum of: (1) the gas pressure (p) working on an area (A; effectively, the cross-sectional area of the gun bore, which is normally taken as a constant), and (2) the resistance to motion resulting from starting forces, friction, and spin (in the case of rifled gun barrels). The latter forces are small relative to that derived from the pressure, but it is essential to take them into account in deriving accurate useful formulations of the interior ballistic trajectory—especially the shot-start pressure (pressure at instant of firing) required to engrave the projectile's soft metal band, usually of copper, that then rotates the projectile along the grooves of the barrel. By using classical physics, the foregoing ballistic data are related in an equation such that the acceleration ($\boldsymbol{a}$) equals the pressure (p) times the area (A) plus the resistant force ($\boldsymbol{f}$), divided by the mass (m), the equation being written:

$$\boldsymbol{a} = \frac{pA + \boldsymbol{f}}{m}.$$

The principal variable in this process is the pressure on the projectile base; if it can be measured, then the acceleration can be calculated. The pressure that is effective on the base of the projectile depends on the propellant's properties, on the amount of it that is burned, and on the spatial distribution of gas down the bore. The outcome of these variables is a pressure that is ordinarily measured by instrumentation installed at a fixed location—the breech end of the gun.

Chemistry of propellants. Propellants are substances that ignite and burn extremely fast, producing voluminous quantities of gas in a very short time. In general, the burning process is a chemical reaction involving oxidation, an oxidizing agent being any substance that removes electrons. The shorter the period during which a given mass of propellant burns, the more explosive will be the production and expansion of the gas. In a gun the propulsive gases are generated by chemical changes that involve so-called monopropellants—*i.e.,* compounds in which one part of the structure serves as the fuel and another part serves as the oxidizer. Guns differ in this respect from internal-combustion engines, into which a mixture of two distinct substances—fuel vapour and atmospheric oxygen—is metered. Propellants must be chemically and physically stable in ordinary environments, and must react only to specific conditions that can be reproduced acceptably. They are ignited by a primer propellant that has been in turn ignited by heat either from an electrical-resistance wire or from an explosive percussion primer. Properly manufactured, prepared, placed, and ignited, the propellant material burns at a rate that depends on the pressure of the product gases and, secondarily, on the temperature of the unburnt material. The higher the temperature, the faster will be the reaction. This burning rate, usually expressed in terms of length burned per unit time, is determined by burning a strand of the material at various controlled pressures. When the strand is inhibited from burning along its sides and made to burn evenly across its face, the linear rate of burning is measured and an equation is derived in which an exponent of the pressure varies from 0.1 to 1, with conventional propellants; such a range of values makes determination of this variable a critical factor in calculating the ballistic behaviour of any particular gun or projectile.

Behaviour of gases. The chemistry of propellants involves the so-called equation of state of the material—*i.e.,* the relationship between pressure, volume, and temperature of a gas. If p is pressure, V the volume available to the gas, n the number of moles of the gas (weight in grams equal to the molecular weight), R the universal gas constant, and T its temperature (in degrees absolute), then the product of pressure and volume will equal the product of the number of moles, the gas constant, and the temperature, thus: $pV = nRT$.

Equation of state of the material

Since the gas itself must be accelerated down the gun bore behind the projectile, it is distributed at any time in such a way that local pressure decreases from the breech to the projectile's base as the latter travels. The average pressure along the gun behind the projectile can be used to calculate the instantaneous rate of burning. The energy consumed in doing the work of moving the propellant gas itself is obtained from the total pressure generated, and the projectile, therefore, receives that much less. Accordingly, acceleration can be estimated by dividing the force of the combustion (the mean pressure of the combustion times the effective area of the bullet) by the sum of the mass of the bullet plus one-third the weight of the propellant. The thermodynamic, or heat, efficiency of the system is also reduced by any gas leakage past the projectile, by potential work available in the gas remaining when the projectile leaves the muzzle, and by heat lost to the gun itself. When all these factors are considered, what is left for the kinetic energy imparted to the projectile is about one-third of the potential chemical energy available from the propellant charge.

There is clearly an intricate relationship between the rate of propellant consumption, the pressure at which the propellant burns, and the volume of gas behind the projectile; the relationship is expressed in several simultaneous differential equations, which can be conveniently programmed on an analogue computer to provide curves showing how the gas pressure and projectile travel depend upon the time. These vary for each assembly of devices, but when for any one system enough measurements have

been made, the formula can be used to predict desired factors (*e.g.*, the amount of a propellant needed to move a certain projectile at a desired acceleration).

Rotational acceleration

Projectiles are often given a spin in order to obtain a more stable and uniform flight pattern. This rotational acceleration of projectiles, as well as other secondary variables, such as frictional resistance to motion and leakage of gas, must be taken into account in a full development of interior ballistics. Although they dissipate collectively, only about 5 percent of the energy consumed in the whole process, they can affect projectile performance significantly and are therefore important to ballistic calculations. For example, the inferred empirical value for shot-start pressure is vital in establishing the conditions for burning during projectile travel.

The interior ballistic portion of the total ballistic trajectory ends with the velocity, spin, and direction imparted to the projectile at the gun muzzle as well as to the projectile's angle of attack to its direction of motion.

Exterior ballistics. Exterior ballistics deals with the flight of projectiles through the atmosphere, and of primary interest are the manner in which the velocity of projectiles is gradually decreased as a result of a factor called drag; how they maintain a consistently favourable attitude (a variable factor called stability); and how they are influenced by variations in interior ballistics and by their flight environment.

Principal concerns

General considerations. First, the interaction between the projectile and the medium through which it passes must be examined and measurements made in terms of coefficients that indicate the amount of changes in the projectile's behaviour as the conditions change. The aerodynamics of a nonspinning projectile are represented in terms of three dimensionless parameters that appear as coefficients of drag, of lift, and of moment (a moment measures the forces tending to turn an object) in equations describing how a projectile is slowed by the frictional resistance offered by the air, how it is displaced laterally (lifted or depressed) by the aerodynamic pressure distribution over the projectile in flight, and how the projectile tends to tumble because of an overturning aerodynamic moment. These coefficients are all functions of the speed of the projectile and properties of the air such as density and elasticity. They are expressible as functions of projectile velocity divided by the local speed of sound (*i.e.*, Mach number; Mach 1 being a speed equal to the velocity of sound in the same medium) and of the so-called Reynolds number (based on the velocity at which streamlined flow of a fluid becomes turbulent) of the air or fluid through which the projectile is passing. The theory of the Reynolds number is complex, and a good approximation to conventional ballistics can be obtained by ignoring it and regarding the three coefficients as functions of Mach number only. (This is not the case for high speed—hypersonic—or hydroballistics. For those the change in Reynolds number is sufficient to require its use in conversion of experimental data.) These coefficients are also different for different projectile shapes. Specifically, for typical shapes the drag coefficient is near 0.08 at low velocities, increases very sharply to about 0.2 at the speed of sound, and decreases slightly at higher velocity. The coefficients are normally determined experimentally either from direct force measurements in a wind tunnel on a fixed projectile shape with air blown over it, or by inference from fitting flight trajectory data to a mathematical model. If ρ is the air density, d the projectile diameter, v the magnitude of the projectile velocity, and δ the angle of yaw—or the angle between the projectile's long axis and its direction of motion—and K_D, K_L, K_M are coefficients of drag, lift, and moment, and these forces are combined into the following linearized equations, then drag equals density times diameter squared times velocity squared times the drag coefficient; lift equals the same factors but substituting the coefficient of lift for the coefficient of drag, times the sine of the angle of yaw; and the moment equals the product of the density, diameter cubed, velocity squared, the coefficient of moment, and the sine and cosine of the angle of yaw. Written out symbolically, the equations are

Mach and Reynolds numbers

$$\text{Drag } D = \rho d^2 v^2 K_D$$
$$\text{Lift } L = \rho d^2 v^2 K_L \times \sin \delta$$
$$\text{Moment } M = \rho d^3 v^2 K_M \times \sin \delta \times \cos \delta.$$

Drag and spin. The drag effect at zero yaw can be broken up into three influences: that due to pressure of the atmosphere on the bow of the projectile resulting in a backward force component; that due to skin friction of the atmosphere along the body of the projectile and varying with its surface roughness and minor convolutions; and that due to base drag resulting from the turbulence induced in the atmosphere at the rear of the projectile. At usual projectile velocities, say 300 to 900 metres per second, pressure drag is reduced by sharpening and shallowing the curve of projectile noses, frictional drag by smoothing the surface of the body, and base drag by drawing the base more toward a conical shape.

Factors in stable flight

In stable projectile flight the orientation of lowest drag is presented to the airstream. Unstable flight leads to high drag, short range, and reduced accuracy. Static stability is achieved when the centre of pressure (resulting from combination of forces including drag) acting on the body is behind its centre of gravity. This is sometimes achieved by the use of fins as on bombs, mortar shells, and nonspinning rockets; the fins acquire the centre of pressure, whereas the centre of gravity remains somewhere around the midsection of the projectile. Dynamic stability is achievable even when the centre of pressure is forward of the centre of gravity by spinning the projectile and thereby inducing a gyroscopic action that counterbalances the wobbling or tumbling that the pressure would induce. Thus, unless suitable fins are provided, spinning the projectile is necessary for the stability of its longitudinal axis in flight. Too much spin will make a projectile so stable at high angles of fire that it will resist following its velocity vector near the zenith of its trajectory and assume an increasingly sideways orientation, and eventually the projectile will tumble. The mathematical treatment of spinning projectiles introduces complex variable notation to the force system described above; a number of separately identifiable torques (or moments) and forces result that are expressible as functions of the same basic parameters that are identified without spin, with the addition of the spin parameter and new dimensionless coefficients. These coefficients are also empirical. To make the mathematical problems tractable, projectile yaw is assumed to be small, and a linearized theory is used. There has been some analytical development of nonlinear yawing motion.

Geoballistic factors. Over long ranges of flight, including an "infinite" range postulated by Newton (*i.e.*, that of orbital satellites), effects that can be neglected in conventional ballistic systems become prominent and, at extreme altitudes, the significance of aerodynamic effects are reduced. For very long ranges and times of flight, the spinning of the Earth imposes a small force, called the Coriolis force, which causes projectiles in the Northern Hemisphere to have a rightward curving flight, and in the Southern Hemisphere a leftward curving flight. The explanation of this is that since the surface of the Earth, spinning on its axis, moves at different speeds at different Earth latitudes (a point on the Equator moves faster than any point north or south because the circumference at the Equator is greatest, while a point at either pole is stationary), when trajectories cover a range of latitudes this must be accounted for. Moreover, the oblate nonspherical configuration of the Earth affects orbital trajectories as does, to lesser degree, local variation in the magnitude and direction of gravity forces. All these effects are treated in a modern branch of ballistics called geoballistics.

Effect of the Coriolis force

Nonspinning projectiles. Bombs are fin stablized and are dealt with as nonspinning projectiles. The problem of aiming a bomb introduces the problem of locating a release point from a moving platform to make the bomb's trajectory intersect a target. Since a bomb-carrying aircraft counteracts air drag with its engine thrust while the falling bomb does not, after its release, the bomb immediately trails the aircraft while it falls. To select the point at which to release the bomb requires mathematical manipulations of data, and the aim is to find the most efficient way

of mechanizing the selection of release conditions. While being carried and during its release, the bomb is subjected to complicated airflow conditions around the aircraft.

Meteorological effects. Corrections for variation from standard, of either gun muzzle velocity or of projectile weight, are supplied by exterior ballistics. A more difficult correction problem is posed by meteorological conditions. Wind speed, direction, and air density vary with altitude and horizontal distance from the gun. Measurements are not generally available for the specific flight path intended, but, rather, one must use data gathered at a more or less nearby location and at an earlier time. Internationally agreed-upon standard atmospheres are used as a first approximation in standard ballistics tables giving range and time of flight for various gun elevations and selections of ammunition. Meteorological differentials are then applied, either for a single "average" value of wind and density variation from standard, or even for variations with altitude.

Terminal ballistics. *General considerations.* Of the various events that occur when a projectile hits a target, the simplest to discuss is penetration of the target by the projectile. Attempts to state analytically the three-dimensional process of penetration accounting for the material properties of target and missile have been only partly successful. Even the simplest two-dimensional attempts have encountered much computational difficulty. The phenomenology of penetration has evolved out of experimental observations and known characteristics of material failure under shock loading. It is most easily approached through projectile impact on a single plate of some thickness. Complete penetration through the plate is called perforation. Partial penetration that pushes whatever is left of the surface at the rear face of the plate around the penetration path is termed scabbing. If the projectile's impact causes a pressure shock of sufficient amplitude in the plate and if the plate thickness and its material permit it, a piece of the plate, thinner than the plate thickness, may be separated and driven by shock from the rear face of the plate without the projectile penetrating an appreciable distance. This piece, called a spall, is split off the plate by the extreme tensile force within the plate, caused by the rarefaction pulse that results when the shock front is reflected from the rear surface of the plate that is not penetrated. A theoretical description of spallation can be given from the impact geometry and velocity and from the equation of state of the materials. The depth of penetration for a projectile that survives impact without fracture is roughly proportional to the impact velocity and the projectile length, and it increases with the density of the penetrator, which may not be identical with the density of the projectile before impact. As the angle of impact on the plate varies from the right angle to the plate, the possibility of ricochet increases.

Spallation

Theory of terminal ballistic impact. A theory of terminal ballistic impact is now evolving as a result of simulation using digital computers, based on a number of hydrodynamic representations (considering the materials under high pressure as fluids). By the use of computer graphics, a picture of the fluid can be presented as it is being compressed or rarefied and as its geometry changes. This is done by calculating and plotting a lattice of connecting points of the fluid as a function of time. Initially, the lattice may look like a section of graph paper, but then with time the lattice is seen to distort and change configurations just as would the fluid when disturbed. Indeed, it is a motion picture of its history. Inputs in such computations include impact velocity and physical variables of projectile and target, such as density, speed of sound in the materials, material strengths, and equation of state, all relating pressure and density in the material. Output variables include dimensional deformation and internal pressure distributions. Varying impact conditions in the formulations beyond the range of available experimental evidence can suggest new directions, but the validity of computed extrapolation beyond experience is questionable until verified by experiment. The tremendous speed and analytical capability of the electronic digital computer nevertheless allow some success in attempts to smooth projectiles for flight, to increase muzzle velocity, and to penetrate the target more deeply. Consequently, in all areas of ballistics, a process of computer-simulated experimentation supported by selected physical experiments has become useful.

Use of computer graphics

Theory of terminal ballistics. A theoretical structure for terminal ballistics has been evolving since 1960. Although all fields of ballistics are centuries old, in an experimental sense, terminal ballistics is in its theoretical youth and has considerable promise in explaining physical relationships in materials at very high pressures. (R.I.Ro./Ed.)

FLUID MECHANICS

Fluid mechanics is concerned with the response of fluids to forces exerted upon them. It is a branch of classical physics with applications of great importance in hydraulic and aeronautical engineering, chemical engineering, meteorology, and zoology.

Hydrostatics and hydrodynamics

The most familiar fluid is of course water, and an encyclopaedia of the 19th century probably would have dealt with the subject under the separate headings of hydrostatics, the science of water at rest, and hydrodynamics, the science of water in motion. Archimedes founded hydrostatics in about 250 BC when, according to legend, he leapt out of his bath and ran naked through the streets of Syracuse crying "Heureka!"; it has undergone rather little development since. The foundations of hydrodynamics, on the other hand, were not laid until the 18th century when mathematicians such as Leonhard Euler and Daniel Bernoulli began to explore the consequences, for a virtually continuous medium like water, of the dynamical principles that Newton had enunciated for systems composed of discrete particles. Their work was continued in the 19th century by several mathematicians and physicists of the first rank, notably G.G. Stokes and William Thomson. By the end of the century explanations had been found for a host of intriguing phenomena having to do with the flow of water through tubes and orifices, the waves that ships moving through water leave behind them, raindrops on windowpanes, and the like. There was still no proper understanding, however, of problems as fundamental as that of water flowing past a fixed obstacle and exerting a drag force upon it; the theory of potential flow, which worked so well in other contexts, yielded results that at relatively high flow rates were grossly at variance with experiment. This problem was not properly understood until 1904, when the German physicist Ludwig Prandtl introduced the concept of the boundary layer (see below *Boundary layers and separation*). Prandtl's career spanned the period in which the first manned aircraft were developed. Since that time, the flow of air has been of as much interest to physicists and engineers as the flow of water, and hydrodynamics has, as a consequence, become fluid dynamics. The term fluid mechanics, as used here, embraces both fluid dynamics and the subject still generally referred to as hydrostatics.

One other representative of the 20th century who deserves mention here besides Prandtl is Geoffrey Taylor of England. Taylor remained a classical physicist while most of his contemporaries were turning their attention to the problems of atomic structure and quantum mechanics, and he made several unexpected and important discoveries in the field of fluid mechanics. The richness of fluid mechanics is due in large part to a term in the basic equation of the motion of fluids which is nonlinear—*i.e.,* one that involves the fluid velocity twice over. It is characteristic of systems described by nonlinear equations that under certain conditions they become unstable and begin behaving in ways that seem at first sight to be totally chaotic. In the case of fluids, chaotic behaviour is very common and is called turbulence. Mathematicians

have now begun to recognize patterns in chaos that can be analyzed fruitfully, and this development suggests that fluid mechanics will remain a field of active research well into the 21st century. (For a discussion of the concept of chaos, see PHYSICAL SCIENCE, PRINCIPLES OF.)

Fluid mechanics is a subject with almost endless ramifications, and the account that follows is necessarily incomplete. Some knowledge of the basic properties of fluids will be needed; a brief survey of the most relevant properties is therefore given in the next section. For further details, see THERMODYNAMICS, PRINCIPLES OF; and MATTER: *Liquid state.*

Basic properties of fluids

Fluids are not strictly continuous media in the way that all the successors of Euler and Bernoulli have assumed, for they are composed of discrete molecules. The molecules, however, are so small and, except in gases at very low pressures, the number of molecules per millilitre is so enormous that they need not be viewed as individual entities. There are a few liquids, known as liquid crystals, in which the molecules are packed together in such a way as to make the properties of the medium locally anisotropic, but the vast majority of fluids (including air and water) are isotropic. In fluid mechanics, the state of an isotropic fluid may be completely described by defining its mean mass per unit volume, or density (ρ), its temperature (T), and its velocity ($\mathbf{v}$) at every point in space, and just what the connection is between these macroscopic properties and the positions and velocities of individual molecules is of no direct relevance.

A word perhaps is needed about the difference between gases and liquids, though the difference is easier to perceive than to describe. In gases the molecules are sufficiently far apart to move almost independently of one another, and gases tend to expand to fill any volume available to them. In liquids the molecules are more or less in contact, and the short-range attractive forces between them make them cohere; the molecules are moving too fast to settle down into the ordered arrays that are characteristic of solids, but not so fast that they can fly apart. Thus samples of liquid can exist as drops or as jets with free surfaces, or they can sit in beakers constrained only by gravity, in a way that samples of gas cannot. Such samples may evaporate in time, as molecules one by one pick up enough speed to escape across the free surface and are not replaced. The lifetime of liquid drops and jets, however, is normally long enough for evaporation to be ignored.

There are two sorts of stress that may exist in any solid or fluid medium, and the difference between them may be illustrated by reference to a brick held between two hands. If the holder moves his hands toward each other, he exerts pressure on the brick; if he moves one hand toward his body and the other away from it, then he exerts what is called a shear stress. A solid substance such as a brick can withstand stresses of both types, but fluids, by definition, yield to shear stresses no matter how small these stresses may be. They do so at a rate determined by the fluid's viscosity. This property, about which more will be said later, is a measure of the friction that arises when adjacent layers of fluid slip over one another. It follows that the shear stresses are everywhere zero in a fluid at rest and in equilibrium, and from this it follows that the pressure (that is, force per unit area) acting perpendicular to all planes in the fluid is the same irrespective of their orientation (Pascal's law). For an isotropic fluid in equilibrium there is only one value of the local pressure (p) consistent with the stated values for ρ and T. These three quantities are linked together by what is called the equation of state for the fluid.

Equation of state

For gases at low pressures the equation of state is simple and well known. It is

$$p=\left(\frac{RT}{M}\right)\rho, \tag{33}$$

where R is the universal gas constant (8.3 joules per degree Celsius per mole) and M is the molar mass, or an average molar mass if the gas is a mixture; for air, the appropriate average is about 29×10^{-3} kilogram per mole. For other fluids knowledge of the equation of state is often incomplete. Except under very extreme conditions, however, all one needs to know is how the density changes when the pressure is changed by a small amount, and this is described by the compressibility of the fluid—either the isothermal compressibility, β_T, or the adiabatic compressibility, β_S, according to circumstance. When an element of fluid is compressed, the work done on it tends to heat it up. If the heat has time to drain away to the surroundings and the temperature of the fluid remains essentially unchanged throughout, then β_T is the relevant quantity. If virtually none of the heat escapes, as is more commonly the case in flow problems because the thermal conductivity of most fluids is poor, then the flow is said to be adiabatic, and β_S is needed instead. (The S refers to entropy, which remains constant in an adiabatic process provided that it takes place slowly enough to be treated as "reversible" in the thermodynamic sense.) For gases that obey equation (33), it is evident that p and ρ are proportional to one another in an isothermal process, and

$$\beta_T=\rho^{-1}\left(\frac{\partial\rho}{\partial p}\right)_T=p^{-1}. \tag{34}$$

In reversible adiabatic processes for such gases, however, the temperature rises on compression at a rate such that

$$T\propto\rho^{(\gamma-1)},\quad p\propto\rho^{\gamma}, \tag{35}$$

and

$$\beta_S=\rho^{-1}\left(\frac{\partial\rho}{\partial p}\right)_S=(\gamma p)^{-1}=\frac{\beta_T}{\gamma}, \tag{36}$$

where γ is about 1.4 for air and takes similar values for other common gases. For liquids the ratio between the isothermal and adiabatic compressibilities is much closer to unity. For liquids, however, both compressibilities are normally much less than p^{-1}, and the simplifying assumption that they are zero is often justified.

The factor γ is not only the ratio between two compressibilities; it is also the ratio between two principal specific heats. The molar specific heat is the amount of heat required to raise the temperature of one mole through one degree. This is greater if the substance is allowed to expand as it is heated, and therefore to do work, than if its volume is fixed. The principal molar specific heats, C_P and C_V, refer to heating at constant pressure and constant volume, respectively, and

Molar specific heat

$$\gamma=\frac{C_P}{C_V}. \tag{37}$$

For air, C_P is about 3.5 R.

Solids can be stretched without breaking, and liquids, though not gases, can withstand stretching, too. Thus, if the pressure is steadily reduced in a specimen of very pure water, bubbles will ultimately appear, but they may not do so until the pressure is negative and well below -10^7 newton per square metre; this is 100 times greater in magnitude than the (positive) pressure exerted by the Earth's atmosphere. Water owes its high ideal strength to the fact that rupture involves breaking links of attraction between molecules on either side of the plane on which rupture occurs; work must be done to break these links. However, its strength is drastically reduced by anything that provides a nucleus at which the process known as cavitation (formation of vapour- or gas-filled cavities) can begin, and a liquid containing suspended dust particles or dissolved gases is liable to cavitate quite easily.

Work also must be done if a free liquid drop of spherical shape is to be drawn out into a long thin cylinder or deformed in any other way that increases its surface area. Here again work is needed to break intermolecular links. The surface of a liquid behaves, in fact, as if it were an elastic membrane under tension, except that the tension exerted by an elastic membrane increases when the membrane is stretched in a way that the tension exerted by a liquid surface does not. Surface tension is what causes liquids to rise up capillary tubes, what supports hanging liquid drops, what limits the formation of ripples on the surface of liquids, and so on.

Surface tension

Hydrostatics

It is common knowledge that the pressure of the atmosphere (about 10^5 newtons per square metre) is due to the weight of air above the Earth's surface, that this pressure falls as one climbs upward, and correspondingly that pressure increases as one dives deeper into a lake (or comparable body of water). Mathematically, the rate at which the pressure in a stationary fluid varies with height z in a vertical gravitational field of strength g is given by

$$\frac{dp}{dz} = -\rho g. \tag{38}$$

If ρ and g are both independent of z, as is more or less the case in lakes, then

$$p(z) = p(0) - \rho g z. \tag{39}$$

This means that, since ρ is about 10^3 kilograms per cubic metre for water and g is about 10 metres per second squared, the pressure is already twice the atmospheric value at a depth of 10 metres. Applied to the atmosphere, equation (39) would imply that the pressure falls to zero at a height of about 10 kilometres. In the atmosphere, however, the variation of ρ with z is far from negligible and (39) is unreliable as a consequence; a better approximation is given below in the section *Compressible flow in gases.*

Differential manometers

Instruments for comparing pressures are called differential manometers, and the simplest such instrument is a U-tube containing liquid, as shown in Figure 54A. The two pressures of interest, p_1 and p_2, are transmitted to the two ends of the liquid column through an inert gas whose density is negligible by comparison with the liquid density, ρ, and the difference of height, h, of the two menisci is measured. It is a consequence of (39) that

$$p_1 - p_2 = \rho g h. \tag{40}$$

A barometer for measuring the pressure of the atmosphere in absolute terms is simply a manometer in which p_2 is made zero, or as close to zero as is feasible. The barometer invented in the 17th century by the Italian physicist and mathematician Evangelista Torricelli, and still in use today, is a U-tube that is sealed at one end (see Figure 54B). It may be filled with liquid, with the sealed end downward, and then inverted. On inversion, a negative pressure may momentarily develop at the top of the liquid column if the column is long enough; however, cavitation normally occurs there and the column falls away from the sealed end of the tube, as shown in the figure. Between the two exists what Torricelli thought of as a vacuum, though it may be very far from that condition if the barometer has been filled without scrupulous precautions to ensure that all dissolved or adsorbed gases, which would otherwise collect in this space, have first been removed. Even if no contaminating gas is present, the Torricellian vacuum always contains the vapour of the liquid, and this exerts a pressure which may be small but is never quite zero. The liquid conventionally used in a Torricelli barometer is of course mercury, which has a low vapour pressure and a high density. The high density means that h is only about 760 millimetres; if water were used, it would have to be about 10 metres instead.

Figure 54C illustrates the principle of the siphon. The top container is open to the atmosphere and the pressure in it, p_2, is therefore atmospheric. To balance this and the weight of the liquid column in between, the pressure p_1 in the bottom container ought to be greater by $\rho g h$. If the bottom container is also open to the atmosphere, then equilibrium is clearly impossible; the weight of the liquid column prevails and causes the liquid to flow downward. The siphon operates only as long as the column is continuous; it fails if a bubble of gas collects in the tube or if cavitation occurs. Cavitation therefore limits the level differences over which siphons can be used, and it also limits (to about 10 metres) the depth of wells from which water can be pumped using suction alone.

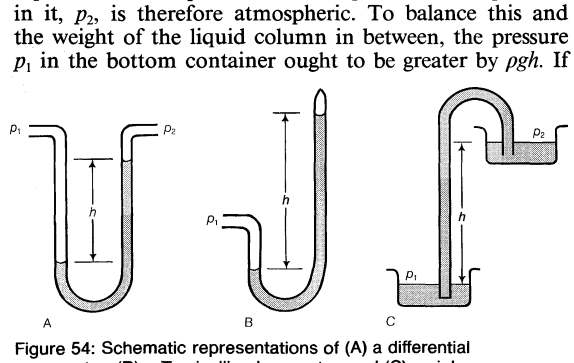


Figure 54: Schematic representations of (A) a differential manometer, (B) a Torricellian barometer, and (C) a siphon.

Consider now a cube of side d totally immersed in liquid with its top and bottom faces horizontal. The pressure on the bottom face will be higher than on the top by $\rho g d$, and, since pressure is force per unit area and the area of a cube face is d^2, the resultant upthrust on the cube is $\rho g d^3$. This is a simple example of the so-called Archimedes' principle, which states that the upthrust experienced by a submerged or floating body is always equal to the weight of the liquid that the body displaces. As Archimedes must have realized, there is no need to prove this by detailed examination of the pressure difference between top and bottom. It is obviously true, whatever the body's shape. It is obvious because if the solid body could somehow be removed and if the cavity thereby created could somehow be filled with more fluid instead, the whole system would still be in equilibrium. The extra fluid would, however, then be experiencing the upthrust previously experienced by the solid body, and it would not be in equilibrium unless this were just sufficient to balance its weight.

Archimedes' principle

Archimedes' problem was to discover, by what would nowadays be called a nondestructive test, whether the crown of King Hieron II was made of pure gold or of gold diluted with silver. He understood that the pure metal and the alloy would differ in density, and that he could determine the density of the crown by weighing it to find its mass and making a separate measurement of its volume. Perhaps the inspiration that struck him (in his bath) was that one can find the volume of any object by submerging it in liquid in something like a measuring cylinder (*i.e.*, in a container with vertical sides that have been suitably graduated) and measuring the displacement of the liquid surface. If so, he no doubt realized soon afterward that a more elegant and more accurate method for determining density can be based on the principle that bears his name. This method involves weighing the object twice, first, when it is suspended in a vacuum (suspension in air will normally suffice) and second, when it is totally submerged in a liquid of density ρ. If the density of the object is ρ', the ratio between the two weights must be

$$\frac{W_2}{W_1} = \frac{(\rho' - \rho)}{\rho'}. \tag{41}$$

If ρ' is less than ρ, then W_2, according to equation (41), is negative. What that means is that the object does not submerge of its own accord; it has to be pushed downward to make it do so. If an object with a mean density less than that of water is placed in a lake and not subjected to any downward force other than its own weight, it naturally floats on the surface, and Archimedes' principle shows that in equilibrium the volume of water which it displaces is a fraction ρ'/ρ of its own volume. A hydrometer is an object graduated in such a way that this fraction may be measured. By floating a hydrometer first in water of density ρ_0 and then in some other liquid of density ρ_1 and comparing the readings, one may determine the ratio ρ_1/ρ_0—*i.e.*, the specific gravity of the other liquid.

In what orientation an object floats is a matter of grave concern to those who design boats and those who travel in them. A simple example will suffice to illustrate the factors that determine orientation. Figure 55 shows three of the many possible orientations that a uniform square prism might adopt when floating, with half its volume submerged in a liquid for which $\rho = 2\rho'$; they are separated by rotations of 22.5°. In each of these diagrams, C is the centre of mass of the prism, and B, a point known as the centre of buoyancy, is the centre of mass of the displaced water. The distributed forces acting on the prism are equivalent to its weight acting downward through C and to the equal weight of the displaced water acting upward through B. In general, therefore, the prism experiences a

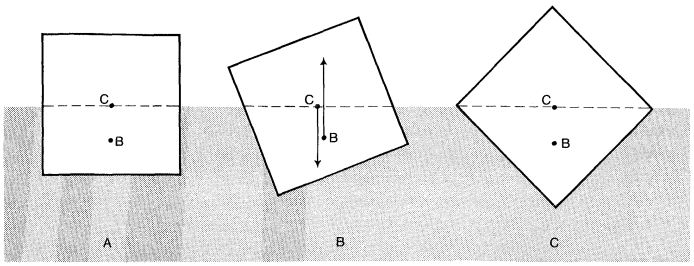


Figure 55: Three possible orientations of a uniform square prism floating in liquid of twice its density. The stable orientation is (C) (see text).

torque. In Figure 55B the torque is counterclockwise, and so it turns the prism away from 55A and toward 55C. In 55C the torque vanishes because B is now vertically below C, and this is the orientation that corresponds to stable equilibrium. The torque also vanishes in 55A, and the prism can in principle remain indefinitely in that orientation as well; the equilibrium in this case, however, is unstable, and the slightest disturbance will cause the prism to topple one way or the other. In fact, the potential energy of the system, which increases in a linear fashion with the difference of height between C and B, is at its smallest in orientation 55C and at its largest in orientation 55A. To improve the stability of a floating object one should, if possible, lower C relative to B. In the case of a boat, this may be done by redistributing the load inside.

Capillarity

Of the many hydrostatic phenomena in which the surface tension of liquids plays a role, the most significant is probably capillarity. Consider what happens when a tube of narrow bore, often called a capillary tube, is dipped into a liquid. If the liquid "wets" the tube (with zero contact angle), the liquid surface inside the tube forms a concave meniscus, which is a virtually spherical surface having the same radius, r, as the inside of the tube. The tube experiences a downward force of magnitude $2\pi r\sigma$, where σ is the surface tension of the liquid, and the liquid experiences a reaction of equal magnitude that lifts the meniscus through a height h such that

$$2\pi r\sigma = \pi r^2 h\rho g; \tag{42}$$

i.e., until the upward force for which surface tension is responsible is balanced by the weight of the column of liquid that has been lifted. If the liquid does not wet the tube, the meniscus is convex and depressed through the same distance h (see Figure 56). A simple method for determining surface tension involves the measurement of h in one or the other of these situations, and the use of equation (42) thereafter.

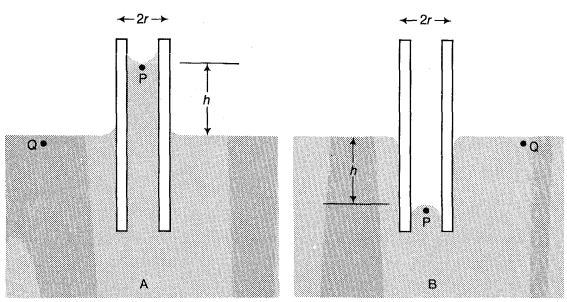


Figure 56: *Capillarity.*
(A) The liquid wets the tube and rises up in it. (B) The liquid does not wet the tube and is depressed.

It follows from equations (39) and (42) that the pressure at a point P just below the meniscus differs from the pressure at Q by an amount

$$\rho gh = \frac{2\sigma}{r}; \tag{43}$$

it is less than the pressure at Q in the case to which Figure 56A refers, and greater than the pressure at Q in the other case. Since the pressure at Q is just the atmospheric pressure, it is equal to the pressure at a point immediately above the meniscus. Hence, in both instances there is a pressure difference of $2\sigma/r$ between the two sides of the curved meniscus, and in both the higher pressure is on the inner side of the curve. Such a pressure difference is a requirement of equilibrium wherever a liquid surface is curved, and it plays a key role in many surface-tension problems. If the surface is curved but not spherical, the pressure difference is

$$\sigma(r_1^{-1} + r_2^{-1}), \tag{44}$$

where r_1 and r_2 are the two principal radii of curvature. If it is cylindrical, one of these radii is infinite; and, if it is curved in opposite directions, then for the purposes of (44) they should be treated as being of opposite sign.

The diagrams in Figure 56 were drawn to represent cross sections through cylindrical tubes, but they might equally well represent two vertical parallel plates, which are partly submerged in the liquid a small distance apart. Consideration of how the pressure varies with height shows that over the range of height h the plates experience a greater pressure on their outer surfaces than on their inner surfaces; this is true whether the liquid wets both plates or not. It is a matter of observation that small objects floating near one another on the surface of a liquid tend to move toward one another, and it is the pressure difference just referred to that makes them behave in this way.

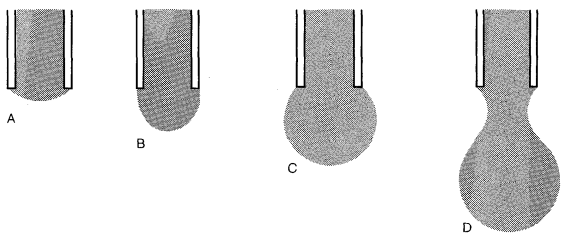


Figure 57: Stages in the formation of a liquid drop (see text).

One other problem having to do with surface tension will be considered here. The diagrams in Figure 57 show stages in the growth of a liquid drop on the end of a tube which the liquid is supposed to wet. In passing from stage A to stage B, by which time the drop is roughly hemispherical in shape, the radius of curvature of the drop diminishes; and it follows from (43) that, to bring about this growth, one must slowly increase the pressure of the liquid inside the tube. If the pressure could be held steady at the value corresponding to B, the drop would then become unstable because any further growth (*e.g.*, to the more or less spherical shape indicated in Figure 57C) would involve an increase in radius of curvature. The applied pressure would then exceed that required to hold the drop in equilibrium, and the drop would necessarily grow bigger still. In practice, however, it is easier to control the rate of flow of water through the tube, and hence the rate of growth of the drop, than it is to control the pressure. If the rate of flow is very small, drops will form the nonspherical shapes suggested by Figure 57D before they detach themselves and fall. It is not an easy matter to analyze the shape of a drop on the point of detachment, and there is no simple formula for the volume of the drop after it is detached.

Hydrodynamics

BERNOULLI'S LAW

Fluids in steady motion

Up to now the focus has been fluids at rest. This section deals with fluids that are in motion in a steady fashion such that the fluid velocity at each given point in space is not changing with time. Any flow pattern that is steady in this sense may be seen in terms of a set of streamlines, the trajectories of imaginary particles suspended in the fluid and carried along with it. In steady flow, the fluid is in motion but the streamlines are fixed. Where the streamlines crowd together, the fluid velocity is relatively high; where they open out, the fluid becomes relatively stagnant.

When Euler and Bernoulli were laying the foundations of hydrodynamics, they treated the fluid as an idealized inviscid substance in which, as in a fluid at rest in equilibrium, the shear stresses associated with viscosity are zero

and the pressure p is isotropic. They arrived at a simple law relating the variation of p along a streamline to the variation of v (the principle is credited to Bernoulli, but Euler seems to have arrived at it first), which serves to explain many of the phenomena that real fluids in steady motion display. To the inevitable question of when and why it is justifiable to neglect viscosity, there is no single answer. Some answers will be provided later in this article, but other matters will be taken up first.

Consider a small element of fluid of mass m, which—apart from the force on it due to gravity—is acted on only by a pressure p. The latter is isotropic and does not vary with time but may vary from point to point in space. It is a well-known consequence of Newton's laws of motion that, when a particle of mass m moves under the influence of its weight mg and an additional force $\boldsymbol{F}$ from a point P where its speed is v_P and its height is z_P to a point Q where its speed is v_Q and its height is z_Q, the work done by the additional force is equal to the increase in kinetic and potential energy of the particle—*i.e.*, that

$$\int_P^Q \boldsymbol{F}\cdot d\boldsymbol{s} = \left(\frac{1}{2}\right)m(v_Q^2 - v_P^2) + mg(z_Q - z_P). \qquad (45)$$

In the case of the fluid element under consideration, $\boldsymbol{F}$ may be related in a simple fashion to the gradient of the pressure, and one finds

$$\int_P^Q \boldsymbol{F}\cdot d\boldsymbol{s} = -m\int_P^Q \rho^{-1}dp. \qquad (46)$$

If the variations of fluid density along the streamline from P to Q are negligibly small, the factor ρ^{-1} may be taken outside the integral on the right-hand side of (46), which thereupon reduces to $\rho^{-1}(p_Q - p_P)$. Then (45) and (46) can be combined to obtain

$$\frac{p_P}{\rho} + \frac{v_P^2}{2} + gz_P = \frac{p_Q}{\rho} + \frac{v_Q^2}{2} + gz_Q. \qquad (47)$$

Since this applies for any two points that can be visited by a single element of fluid, one can immediately deduce Bernoulli's (or Euler's) important result that along each streamline in the steady flow of an inviscid fluid the quantity

$$\left(\frac{p}{\rho} + \frac{v^2}{2} + gz\right) \qquad (48)$$

is constant.

Under what circumstances are variations in the density negligibly small? When they are very small compared with the density itself—*i.e.*, when

$$\left(\frac{\Delta\rho}{\rho}\right) = \beta_S\Delta p = (\beta_S\rho)\Delta\left(\frac{v^2}{2} + gz\right) = \frac{\Delta\left(\frac{v^2}{2} + gz\right)}{V_S^2} \ll 1, \qquad (49)$$

where the symbol Δ is used to represent the extent of the change along a streamline of the quantity that follows it, and where V_s is the speed of sound (see below *Compressible flow in gases*). This condition is satisfied for all the flow problems having to do with water that are discussed below. If the fluid is air, it is adequately satisfied provided that the largest excursion in z is on the order of metres rather than kilometres and provided that the fluid velocity is everywhere less than about 100 metres per second.

Bernoulli's law indicates that, if an inviscid fluid is flowing along a pipe of varying cross section, then the pressure is relatively low at constrictions where the velocity is high, and relatively high where the pipe opens out and the fluid stagnates. Many people find this situation paradoxical when they first encounter it. Surely, they say, a constriction should increase the local pressure rather than diminish it? The paradox evaporates as one learns to think of the pressure changes along the pipe as cause and the velocity changes as effect, instead of the other way around; it is only because the pressure falls at a constriction that the pressure gradient upstream of the constriction has the right sign to make the fluid accelerate.

Paradoxical or not, predictions based on Bernoulli's law are well-verified by experiment. Try holding two sheets of paper so that they hang vertically two centimetres or so apart and blow downward so that there is a current of air between them. The sheets will be drawn together by the reduction in pressure associated with this current. Ships are drawn together for much the same reason if they are moving through the water in the same direction at the same speed with a small distance between them. In this case, the current results from the displacement of water by each ship's bow, which has to flow backward to fill the space created as the stern moves forward, and the current between the ships, to which they both contribute, is stronger than the current moving past their outer sides. As another simple experiment, listen to the hissing sound made by a tap that is almost, but not quite, turned off. What happens in this case is that the flow is so constricted and the velocity within the constriction so high that the pressure in the constriction is actually negative. Assisted by the dissolved gases that are normally present, the water cavitates as it passes through, and the noise that is heard is the sound of tiny bubbles collapsing as the water slows down and the pressure rises again on the other side.

Two practical devices that are used by hydraulic engineers to monitor the flow of liquids though pipes are based on Bernoulli's law. One is the venturi tube, a short length with a constriction in it of standard shape (see Figure 58A), which may be inserted into the pipe proper. If the velocity at point P, where the tube has a cross-sectional area A_P, is v_P and the velocity in the constriction, where the area is A_Q, is v_Q, the continuity condition—the condition that the mass flowing through the pipe per unit time has to be the same at all points along its length—suggests that $\rho_P A_P v_P = \rho_Q A_Q v_Q$, or that $A_P v_P = A_Q v_Q$ if the difference between ρ_P and ρ_Q is negligible. Then Bernoulli's law indicates

Venturi and pitot tubes

$$\rho gh = (p_P - p_Q) = \left(\frac{1}{2}\right)\rho v_P^2\left[\left(\frac{A_P}{A_Q}\right)^2 - 1\right]. \qquad (50)$$

Thus one should be able to find v_P, and hence the quantity Q ($= A_P v_P$) that engineers refer to as the rate of discharge, by measuring the difference of level h of the fluid in the two side tubes shown in the diagram. At low velocities the pressure difference $(p_P - p_Q)$ is greatly affected by viscosity (see below *Viscosity*), and equation (50) is unreliable in consequence. The venturi tube is normally used, however, when the velocity is large enough for the flow to be turbulent (see below *Turbulence*). In such a circumstance, equation (50) predicts values for Q that agree with values measured by more direct means to within a few parts percent, even though the flow pattern is not really steady at all.

The other device is the pitot tube, which is illustrated in Figure 58B. The fluid streamlines divide as they approach the blunt end of this tube, and at the point marked Q in the diagram there is complete stagnation, since the fluid at this point is moving neither up nor down nor to the right. It follows immediately from Bernoulli's law that

$$\rho gh = (p_Q - p_P) = \left(\frac{1}{2}\right)\rho v_P^2. \qquad (51)$$

As with the venturi tube, one should therefore be able to find v_P from the level difference h.

One other simple result deserves mention here. It concerns a jet of fluid emerging through a hole in the wall of a vessel filled with liquid under pressure. Observation of jets shows that after emerging they narrow slightly before settling down to a more or less uniform cross section known as the vena contracta. They do so because the streamlines are converging on the hole inside the vessel and are obliged to continue converging for a short while outside. It was Torricelli who first suggested that, if the pressure excess inside the vessel is generated by a head of liquid h, then the velocity v at the vena contracta is the velocity that a free particle would reach on falling through a height h—*i.e.*, that

$$v = \sqrt{(2gh)}. \qquad (52)$$

This result is an immediate consequence, for an inviscid fluid, of the principle of energy conservation that Bernoulli's law enshrines.

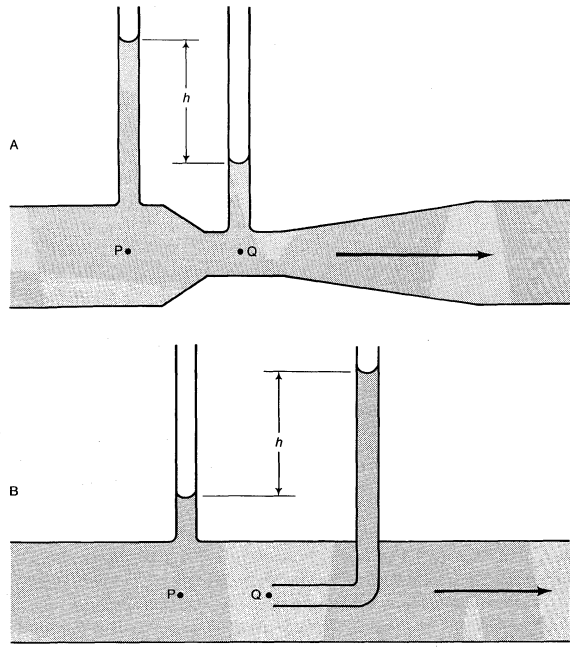


Figure 58: Schematic representation of (A) a venturi tube and of (B) a pitot tube.

In the following section, Bernoulli's law is used in an indirect way to establish a formula for the speed at which disturbances travel over the surface of shallow water. The explanation of several interesting phenomena having to do with water waves is buried in this formula. Analogous phenomena dealing with sound waves in gases are discussed below in *Compressible flow in gases,* where an alternative form of Bernoulli's law is introduced. This form of the law is restricted to gases in steady flow, but is not restricted to flow velocities that are much less than the speed of sound. The complication that viscosity represents is again ignored throughout these two sections.

WAVES ON SHALLOW WATER

Imagine a layer of water with a flat base that has a small step on its surface, dividing a region in which the depth of the water is uniformly equal to D from a region in which it is uniformly equal to $D(1+\varepsilon)$, with $\varepsilon << 1$. Let the water in the shallower region flow toward the step with some uniform speed V, as Figure 59A suggests, and let this speed be just sufficient to hold the step in the same position so that the flow pattern is a steady one. The continuity condition (*i.e.,* the condition that as much water flows out to the left per unit time as flows in from the right) indicates that in the deeper region the speed of the water is $V(1+\varepsilon)^{-1}$. Hence by applying Bernoulli's law to the points marked P and Q in the diagram, which lie on the same streamline and at both of which the pressure is atmospheric, one may deduce that

$$g\varepsilon D = \left(\frac{1}{2}\right)V^2[1-(1+\varepsilon)^{-2}] \approx \varepsilon V^2,$$

i.e., that

$$V = \sqrt{(gD)}. \tag{53}$$

This result shows that, if the water in the shallower region is in fact stationary (see Figure 59B), the step advances over it with the speed V that equation (53) describes, and it reveals incidentally that behind the step the deeper water follows up with speed $V[1-(1+\varepsilon)^{-1}] \approx \varepsilon V$. The argument may readily be extended to disturbances of the surface that are undulatory rather than steplike. Provided that the distance between successive crests—a distance known as the wavelength and denoted by λ—is much greater than the depth of the water, D, and provided that its amplitude is very much less than D, a wave travels over stationary water at a speed given by (53). Because their speed does not depend on wavelength, the waves are said to be nondispersive.

Non-dispersive waves

Evidently waves that are approaching a shelving beach should slow down as D diminishes. If they are approaching it at an angle, the slowing-down effect bends, or refracts, the wave crests so that they are nearly parallel to the shore by the time they ultimately break.

Suppose now that a small step of height εD ($\varepsilon << 1$) is traveling over stationary water of uniform depth D and that behind it is a second step of much the same height traveling in the same direction. Because the second step (suggested by a dotted line in Figure 59B) is traveling on a base that is moving at $\varepsilon\sqrt{(gD)}$ and because the thickness of that base is $(1+\varepsilon)D$ rather than D, the speed of the second step is approximately $(1+3\varepsilon/2)\sqrt{(gD)}$. Since this is greater than $\sqrt{(gD)}$, the second step is bound to catch up with the first. Hence, if there are a succession of infinitesimal steps that raise the depth continuously from D to some value D', which differs significantly from D, then the ramp on the surface is bound to become steeper as it advances. It may be shown that if D' exceeds about $1.3D$, the ramp ultimately becomes a vertical step of finite height and that the step then "breaks." A finite step that has broken dissipates energy as heat in the resultant foaming motion, and Bernoulli's equation is no longer applicable to it. A simple argument based on conservation of momentum rather than energy, however, suffices to show that its velocity of propagation is

$$\sqrt{\left(\frac{gD'(D'+D)}{2D}\right)}. \tag{54}$$

Tidal bores, which may be observed on some estuaries, are examples on the large scale of the sort of phenomena to which (54) applies. Examples on a smaller scale include the hydraulic jumps that are commonly seen below weirs and sluice gates where a smooth stream of water suddenly rises at a foaming front. In this case, (54) describes the speed of the water, since the front itself is more or less stationary.

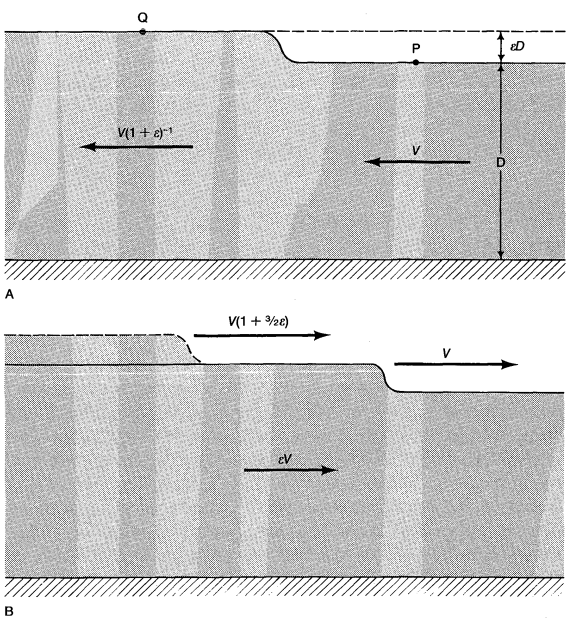


Figure 59: *Steps on the surface of shallow water.* In (A) the water is moving and the step is stationary. In (B) the water is stationary in front of the first step and the step is therefore moving; the second step (dotted line) is catching up to the first.

When water is shallow but not extremely shallow so that correction terms of the order of $(D/\lambda)^2$ are significant, waves of small amplitude become slightly dispersive (see below *Waves on deep water*). In this case, a localized disturbance on the surface of a river or canal, which is guided by the banks in such a way that it can propagate in one direction only, is liable to spread as it propagates. If its amplitude is not small, however, the tendency to spread due to dispersion may in special circumstances be subtly

balanced by the factors that cause waves of relatively large amplitude to form bores, and the result is a localized hump in the surface, of symmetrical shape, which does not spread at all. The phenomenon was first observed on a canal near Edinburgh in 1834 by a Scottish engineer named Scott Russell; he later wrote a graphic account of following on horseback, for well over a kilometre, a "large solitary elevation . . . which continued its course along the channel apparently without change of form." What Scott Russell saw is now called a soliton. Solitons on canals can have various widths, but the smaller the width the larger the height must be and the faster the soliton travels. Thus if a high, narrow soliton is formed behind a low, broad one, it will catch up with the low one. It turns out that when the high soliton does so, it passes through the low one and emerges with its shape unchanged (see Figure 60).

Solitons

It is now recognized that many of the nonlinear differential equations that appear in diverse branches of physics have solutions of large amplitude corresponding to solitons, and that the remarkable capacity of solitons for surviving encounters with other solitons is universal. This discovery has stimulated much interest among mathematicians and physicists, and understanding of solitons is expanding rapidly.

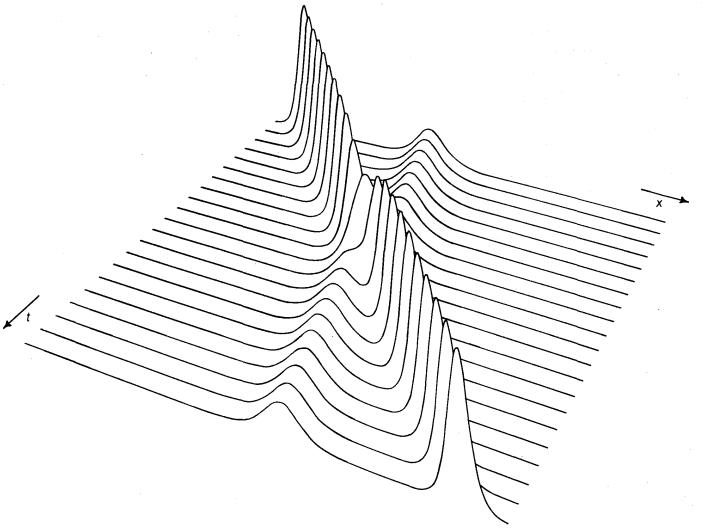


Figure 60: Interaction of two solitons (see text).

COMPRESSIBLE FLOW IN GASES

Compressible flow refers to flow at velocities that are comparable to, or exceed, the speed of sound. The compressibility is relevant because at such velocities the variations in density that occur as the fluid moves from place to place cannot be ignored.

Flow at velocities equal to or greater than the speed of sound

Suppose that the fluid is a gas at a low enough pressure for the ideal equation of state, equation (33), to apply and that its thermal conductivity is so poor that the compressions and rarefactions undergone by each element of the gas may be treated as adiabatic (see above). In this case, it follows from equation (35) that the change of density accompanying any small change in pressure, dp, is such that

$$\rho^{-1}dp = \left(\frac{\gamma}{(\gamma-1)}\right)d\left(\frac{p}{\rho}\right). \tag{55}$$

This makes it possible to integrate the right-hand side of equation (46), and one thereby arrives at a version of Bernoulli's law for a steady compressible flow of gases which states that

$$\frac{\gamma p}{(\gamma-1)\rho} + \frac{v^2}{2} + gz \tag{56}$$

is constant along a streamline. An equivalent statement is that

$$\frac{C_p T}{M} + \frac{v^2}{2} + gz \tag{57}$$

is constant along a streamline. It is worth noting that when a gas flows through a nozzle or through a shock front (see below), the flow, though adiabatic, may not be reversible in the thermodynamic sense. Thus the entropy of the gas is not necessarily constant in such flow, and as a consequence the application of equation (35) is open to question. Fortunately, the result expressed by (56) or (57) can be established by arguments that do not involve integration of (46). It is valid for steady adiabatic flow whether this is reversible or not.

Bernoulli's law in the form of (57) may be used to estimate the variation of temperature with height in the Earth's atmosphere. Even on the calmest day the atmosphere is normally in motion because convection currents (see below *Convection*) are set up by heat derived from sunlight that is released at the Earth's surface. The currents are indeed adiabatic to a good approximation, and their velocity is generally small enough for the term in v^2 in (57) to be negligible. One can therefore deduce without more ado that the temperature of the atmosphere should fall off in a linear fashion—*i.e.*, that

$$T(z) = T(0) - \beta z = T(0) - \left(\frac{Mg}{C_p}\right)z. \tag{58}$$

Here β is used to represent the temperature lapse rate, and the value suggested for this quantity, (Mg/C_p), is close to 10° C per kilometre for dry air.

Lapse rate

This prediction is not exactly fulfilled in practice. Within the troposphere (*i.e.*, to the heights of about 10 kilometres to which convection currents extend), the mean temperature does decrease with height in a linear fashion, but β is only about 6.5° C per kilometre. It is the water vapour in the atmosphere, which condenses as the air rises and cools, that lowers the lapse rate to this value by increasing the effective value of C_p. The fact that the lapse rate is smaller for moist air than for dry air means that a stream of moist air which passes over a mountain range and which deposits its moisture as rain or snow at the summit is warmer when it descends to sea level on the other side of the range than it was when it started. The foehn wind of the Alps owes its warmth to this effect.

The variation of the pressure of the atmosphere with height may be estimated in terms of β, using the equation

$$p(z) = p(0)\left[1 - \frac{\beta z}{T(0)}\right]^{\frac{Mg}{R\beta}}. \tag{59}$$

This is obtained by integration of (38), using (33) and (58).

In the form of equation (56), Bernoulli's law may be used to calculate the speed of sound in gases. The argument is directly analogous to the one applied in the previous section to waves on shallow water—and, indeed, the diagrams in Figure 59 can serve to illustrate the argument here too, if they are regarded as plots of gas density (or else of pressure or temperature, which go hand in hand with density in adiabatic flow) versus position. The results of the argument will be stated without proof. If there exists an infinitesimal step in the density of the gas, it will remain stationary provided that the gas flows uniformly through it toward the region of higher density, with a velocity

Determining the speed of sound in gases

$$V_s = \sqrt{\left(\frac{\gamma p}{\rho}\right)}. \tag{60}$$

If the gas is stationary, then (60) describes the velocity with which the step moves. It also describes the speed of propagation of the sort of undulatory variation of density that constitutes a sound wave of fixed frequency or pitch. Because the speed of sound is independent of pitch, sound waves, like waves on shallow water, are nondispersive. This is just as well. It is only because there is no dispersion that one can understand the words of a distant speaker or listen to a symphony orchestra with pleasure from the back of an auditorium as well as from the front.

It should be noted that the formula for the speed of sound in gases may be proved in other ways, and Newton came close to it a century before Bernoulli's time. However, because Newton failed to appreciate the distinction between adiabatic and isothermal flow, his answer lacked the factor γ occurring in (60). The first person to correct this error was Pierre-Simon Laplace.

The above statements apply to density steps or undula-

tions, the amplitude of which is infinitesimal, and they need some modification if the amplitude is large. In the first place it is found, as for waves on shallow water and for very much the same reasons, that where two small density steps are moving parallel to one another, the second is bound to catch up with the first. It follows that if there exists a propagating region in which the density rises in a continuous fashion from ρ to ρ', where $(\rho' - \rho)$ is not necessarily small, then the width of this region is bound to diminish as time passes. Ultimately a shock front develops over which the density—and hence the pressure and temperature—rises almost discontinuously. There are processes within the shock front, vaguely analogous on the molecular scale to the foaming of a breaking water wave, by which energy is dissipated as heat. The speed of propagation, V_{sh}, of a shock front in a gas that is stationary in front of it may be expressed in terms of V_s and V_s', the velocities of small-amplitude sound waves in front of the shock and behind it, respectively, by the equation

$$2V_{sh}^2 = \left[\frac{(\gamma+1)\rho'}{\gamma\rho}\right]V_s'^2 + V_s^2. \qquad (61)$$

Thus, if the shock is a strong one ($\rho' >> \rho$), V_{sh} may be significantly greater than both V_s and V_s'.

Even the gentlest sound wave, in which density and pressure initially oscillate in a smooth and sinusoidal fashion, develops into a succession of weak shock fronts in time. More noticeable shock fronts are a feature of the flow of gases at supersonic speeds through the nozzles of jet engines and accompany projectiles that are moving through stationary air at supersonic speeds. In certain circumstances when a supersonic aircraft is following a curved path, the accompanying shock wave may accidentally reinforce itself in places and thereby become offensively noticeable as a "sonic boom," which may break windowpanes and cause other damage. Strong shock fronts also occur immediately after explosions, of course, and when windowpanes are broken by an explosion, the broken glass tends to fall outward rather than inward. Such is the case because the glass is sucked out by the relatively low density and pressure that succeed the shock itself.

The diagrams in Figure 61 show a well-known construction attributed to the Austrian physicist Ernst Mach that explains the origin of the shock front accompanying a supersonic projectile. The circular arcs in this figure represent cross sections through spherical disturbances that are spreading with speed V_s from centres (S′, S″, etc.), which mark the position of the moving source S at the time when they were emitted. If the source is something like the tip of an arrow, which disturbs the air by parting it as it travels along but which is inaudible when stationary, then each "disturbance," due to some infinitesimal displacement of the tip, is a spherical shell of infinitesimal thickness within which a small radial velocity has been imparted to the air. There is an infinite number of such disturbances, overlapping one another, of which only a handful are represented in Figure 61. When the velocity of the source, U, is less than V_s (Figure 61A), the result of adding them together is the sort of steady backflow that is to be expected around a moving obstacle, and there is no sound emission in the normal sense; the source remains inaudible. When U exceeds V_s, however, the spherical disturbances reinforce one another, as Figure 61B shows, on a conical caustic surface, which makes an angle of $\sin^{-1}(U/V)$ to the line of travel of the source, and it is on this surface that a shock front is to be expected. The cone becomes sharper as the source speeds up.

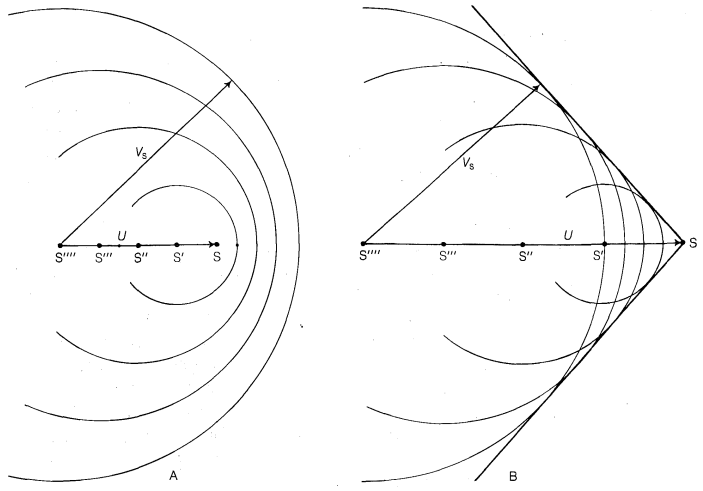


Figure 61: *Mach's construction.*
(A) Source speed U less than speed of sound V_s, (B) U greater than V_s (see text).

VISCOSITY

It has been shown in the three preceding sections that a number of phenomena of considerable physical interest can be discussed using little more than the law of conservation of energy, as expressed by Bernoulli's law. However, the argument has so far been restricted to cases of steady flow. To discuss cases in which the flow is not steady, an equation of motion for fluids is needed, and one cannot write down a realistic equation of motion without facing up to the problems presented by viscosity, which have so far been deliberately set aside.

The concept of viscosity was first formalized by Newton, who considered the shear stresses likely to arise when a fluid undergoes what is called laminar motion with the sort of velocity profile that is suggested in Figure 62A; the laminae here are planes normal to the x_2-axis, and they are moving in the direction of the x_1-axis with a velocity v_1, which increases in a linear fashion with x_2. Newton suggested that, as each lamina slips over the one below, it exerts a sort of frictional force upon the latter in the forward direction, in which case the upper lamina is bound to experience an equal reaction in the backward direction. The strength of these forces per unit area constitutes the component of shear stress normally written as σ_{12} (not to be confused with surface tension, for which the symbol σ has been used above). Figure 62B shows, in elevation, an enlarged view of an infinitesimal element of the fluid of cubic shape, and the directions of the forces experienced by this cube associated with σ_{12} are indicated by arrows. Other arrows show the directions of the forces associated with the so-called normal stresses σ_{11} and σ_{22}, which in the absence of motion of the fluid would both be equal, by Pascal's law, to $-p$. Now σ_{12} is clearly zero when the rate of variation of velocity, $\partial v_1/\partial x_2$, is zero, for then there is no slip, and presumably it increases monotonically as $\partial v_1/\partial x_2$ increases. Newton made the plausible assumption that the two are linearly related—*i.e.*, that

Stresses in laminar motion

$$\sigma_{12} = \eta\left(\frac{\partial v_1}{\partial x_2}\right). \qquad (62)$$

The full name for the coefficient η is shear viscosity to distinguish it from the bulk viscosity, b, which is defined below. The word shear, however, is frequently omitted in this context.

Now if the only shear stress acting on the cubic element of fluid sketched in Figure 62B were σ_{12}, the cube would

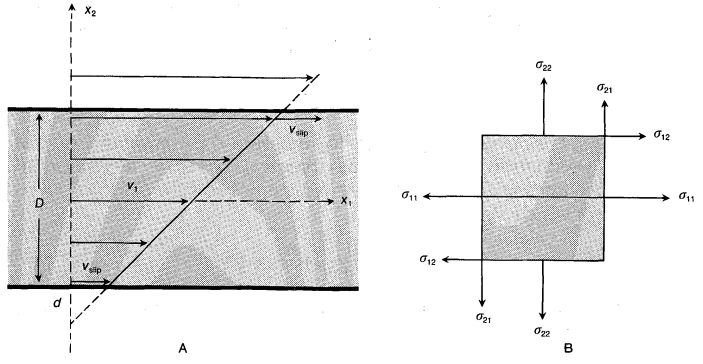


Figure 62: *Laminar motion and associated stresses.*
(A) Velocity profile for laminar flow between two plates, driven by motion of the upper plate, (B) an enlarged view of a cubic element of the fluid between the plates, showing the stresses that act upon it.

experience a torque tending to make it twist in a clockwise sense. Since the magnitude of the torque would vary like the third power of the linear dimensions of the cube, whereas the moment of inertia of the element would vary like the fifth power, the resultant angular acceleration for an infinitesimal cube would be infinite. One may infer that any tendency to twist in a clockwise sense gives rise instantaneously to an additional shear stress σ_{21}, the direction of which is indicated in the diagram, and that σ_{12} and σ_{21} are equal at all times. It follows that equation (62) cannot be a complete expression for these shear stresses, for it does not include the possibility that the fluid is moving in the x_2 direction, with a velocity v_2 that varies with x_1. The complete expression for what is called a Newtonian fluid is

$$\sigma_{12} = \sigma_{21} = \eta\left[\left(\frac{\partial v_1}{\partial x_2}\right) + \left(\frac{\partial v_2}{\partial x_1}\right)\right]. \tag{63}$$

Similar expressions may be written down for σ_{23} ($= \sigma_{32}$) and σ_{31} ($= \sigma_{13}$). Since Newton's day these hypothetical expressions have been fully substantiated for gases and simple liquids, not only by experiment but also by analysis of the molecular motions and molecular interactions in such fluids undergoing shear, and for such fluids one can even predict the magnitude of η with reasonable success. There do exist, however, more complicated fluids for which the Newtonian description of shear stress is inadequate, and some of these are very familiar in the home. In the whites of eggs, for example, and in most shampoos, there are long-chain molecules that become entangled with one another, and entanglement may hinder their efforts to respond to changes of environment associated with flow. As a result, the stresses acting in such fluids may reflect the deformations experienced by the fluid in the recent past as much as the instantaneous rate of deformation. Moreover, the relation between stress and rate of deformation may be far from linear. Non-Newtonian effects, interesting though they are, lie outside the scope of the present discussion, however.

Non-Newtonian effects

The sort of velocity profile that is suggested by Figure 62B may be established by containing the fluid between two parallel flat plates and moving one plate relative to the other. The possibility exists that in this situation the layers of fluid immediately in contact with each plate will slip over them with some finite velocity (indicated in the diagram by an arrow labeled v_{slip}). If so, the frictional stresses associated with this slip must be such as to balance the shear stress $\eta(\partial v_1/\partial x_2)$ exerted on each of these layers by the rest of the fluid. Little is known about fluid–solid frictional stresses, but intelligent guesswork suggests that they are proportional in magnitude to v_{slip} and that, in the circumstances to which Figure 62A refers, the distance d below the surface of the stationary bottom plate at which the straight line representing the variation of v_1 with x_2 extrapolates to zero should be of the same order of magnitude as the diameter of a molecule if the fluid is a liquid, or as the molecular "mean free path" if it is a gas. These distances are normally very small compared with the separation of the plates, D. Accordingly, fluid flow patterns may normally be treated as subject to the boundary condition that at a fluid–solid interface the relative velocity of the fluid is zero. No reliable evidence for failure of predictions based on this no-slip boundary condition has yet been found, except in the case of what is called Knudsen flow of gases (*i.e.*, flow at such low pressures that the mean free path is comparable in length with the dimensions of the apparatus).

If a fluid is flowing steadily between two parallel plates that are both stationary and if its velocity must be zero in contact with both of them, the velocity profile must necessarily have the form indicated in Figure 63. A force in the forward direction due to the shear stress $\eta(\partial v_1/\partial x_2)$ is transmitted to the plates, and an equal force in the backward direction acts on the fluid. The motion therefore cannot be maintained unless the pressure acting on the fluid is greater on the left of the diagram than it is on the right. A full analysis shows the velocity profile to be parabolic, and it indicates that the rate of discharge is related to the pressure gradient by the equation

$$Q = -\left(\frac{WD^3}{12\eta}\right)\left(\frac{dp}{dx_1}\right), \tag{64}$$

where W ($>> D$) is the width of the plates, measured perpendicular to the diagram in Figure 63. A similar analysis of the problem of steady flow through a (horizontal) cylindrical pipe of uniform diameter D, to which Figure 63 could equally well apply, shows the rate of discharge in this case to be given by

$$Q = -\left(\frac{\pi D^4}{128\eta}\right)\left(\frac{dp}{dx_1}\right); \tag{65}$$

this famous result is known as Poiseuille's equation, and the type of flow to which it refers is called Poiseuille flow.

Poiseuille's equation

Viscosity may affect the normal stress components, σ_{11}, σ_{22}, and σ_{33}, as well as the shear stress components. To see why this is so, one needs to examine the way in which stress components transform when one's reference axes are rotated. Here, the result will be stated without proof that the general expression for σ_{11} consistent with (63) is

$$\sigma_{11} = -p + \left(b + \frac{4\eta}{3}\right)\left(\frac{\partial v_1}{\partial x_1}\right) + \left(b - \frac{2\eta}{3}\right)\left(\frac{\partial v_2}{\partial x_2}\right) + \left(b - \frac{2\eta}{3}\right)\left(\frac{\partial v_3}{\partial x_3}\right). \tag{66}$$

On the right-hand side of this equation, p represents the equilibrium pressure defined in terms of local density and temperature by the equation of state, and b is another viscosity coefficient known as the bulk viscosity.

The bulk viscosity is relevant only where the density is changing. Thus it plays a role in attenuating sound waves in fluids and may be estimated from the magnitude of the attenuation. If the fluid is effectively incompressible, however, so that changes of density may be ignored, the flow is everywhere subject to the continuity condition that

$$\left(\frac{\partial v_1}{\partial x_1}\right) + \left(\frac{\partial v_2}{\partial x_2}\right) + \left(\frac{\partial v_3}{\partial x_3}\right) [\equiv \nabla \cdot \boldsymbol{v} \text{ or } \operatorname{div} \boldsymbol{v}] = 0. \tag{67}$$

The terms in (66) that involve b then cancel, and the expression simplifies to

$$\sigma_{11} = -p + 2\eta\left(\frac{\partial v_1}{\partial x_1}\right). \tag{68}$$

Similar equations may be written down for σ_{22} and σ_{33}. These simpler expressions provide the basis for the argument that follows, and the bulk viscosity can be left on one side.

A variety of methods are available for the measurement of shear viscosity. One standard method involves measurement of the pressure gradient along a pipe for various rates of flow and application of Poiseuille's equation. Other methods involve measurement either of the damping of the torsional oscillations of a solid disk supported between two parallel plates when fluid is admitted to the space between the plates, or of the effect of the fluid on the frequency of the oscillations.

The Couette viscometer deserves a fuller explanation. In this device, the fluid occupies the space between two coaxial cylinders of radii a and b ($> a$); the outer cylinder is rotated with uniform angular velocity ω_0, and the resultant torque transmitted to the inner stationary cylinder is measured. If both the terms on the right-hand side of equation (63) are taken into account, the shear stress in the circulating fluid is found to be proportional to

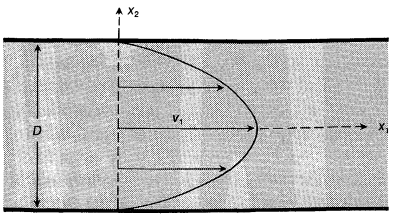


Figure 63: Velocity profile for laminar flow between two plates (or inside a cylindrical tube), driven by a pressure gradient (see text).

$r(d\omega/dr)$ rather than to (dv/dr)—not an unexpected result, since it is only if ω, the angular velocity of the fluid, varies with radius r that there is any slip between one cylindrical lamina of fluid and the next. The torque transmitted through the fluid is therefore proportional to $r^3(d\omega/dr)$. In the steady state, the opposing torques acting on the inner and outer surfaces of each cylindrical lamina of fluid must be of equal magnitude—otherwise the laminae accelerate—and this means that $r^3(d\omega/dr)$ must be independent of r. There are two basic modes of motion for a circulating fluid that satisfy this condition: in one, the liquid rotates as a solid body would, with an angular velocity that does not vary with r, and the torque is everywhere zero; in the other, ω varies like r^{-2}. The angular velocity of the fluid in a Couette viscometer can be viewed as a mixture of these two modes in proportions that satisfy the boundary conditions at $r = a$ and $r = b$. The torque transmitted per unit length of the cylinders turns out to be given by

$$4\pi\eta\omega_0\left[\frac{b^2a^2}{(b^2-a^2)}\right]. \tag{69}$$

It may be added that if the inner cylinder is absent, the steady flow pattern consists only of the first mode—*i.e.*, the fluid rotates like a solid body with uniform angular velocity ω_0. If the outer cylinder is absent, however, and the inner one rotates, it then consists only of the second mode. The angular velocity falls off like r^{-2}, and the velocity v falls off like r^{-1}.

In the equation of motion given in the following section, the shear viscosity occurs only in the combination (η/ρ). This combination occurs so frequently in arguments of fluid dynamics that it is given a special name—kinetic viscosity. The kinetic viscosity at normal temperatures and pressures is about 10^{-6} square metre per second for water and about 1.5×10^{-5} square metre per second for air.

Kinetic viscosity

NAVIER–STOKES EQUATION

One may have a situation where σ_{11} increases with x_1. The force that this component of stress exerts on the right-hand side of the cubic element of fluid sketched in Figure 62B will then be greater than the force in the opposite direction that it exerts on the left-hand side, and the difference between the two will cause the fluid to accelerate along x_1. Accelerations along x_1 will also result if σ_{12} and σ_{13} increase with x_2 and x_3, respectively. These accelerations, and corresponding accelerations in the other two directions, are described by the equation of motion of the fluid. For a fluid moving so slowly compared with the speed of sound that it may be treated as incompressible and in which the variations of temperature from place to place are insufficient to cause significant variations in the shear viscosity η, this equation takes the form

$$-\nabla\left(\frac{p}{\rho}+gz\right)-\left(\frac{\eta}{\rho}\right)[\nabla\times(\nabla\times\boldsymbol{v})]=\frac{D\boldsymbol{v}}{Dt}=\frac{\partial\boldsymbol{v}}{\partial t}+(\boldsymbol{v}\cdot\nabla)\boldsymbol{v}. \tag{70}$$

Euler derived all the terms in this equation except the one on the left-hand side proportional to (η/ρ), and without that term the equation is known as the Euler equation. The whole is called the Navier–Stokes equation.

The equation is written in a compact vector notation which many readers will find totally impenetrable, but a few words of explanation may help some others. The symbol ∇ represents the gradient operator, which, when preceding a scalar quantity X, generates a vector with components $(\partial X/\partial x_1, \partial X/\partial x_2, \partial X/\partial x_3)$. The vector product of this operator and the fluid velocity $\boldsymbol{v}$—*i.e.*, $(\nabla \times \boldsymbol{v})$—is sometimes designated as ***curl v*** [and $\nabla\times(\nabla\times\boldsymbol{v})$ is also ***curl curl v***]. Another name for $(\nabla \times \boldsymbol{v})$, which expresses particularly vividly the characteristics of the local flow pattern that it represents, is vorticity. In a sample of fluid that is rotating like a solid body with uniform angular velocity ω_0, the vorticity lies in the same direction as the axis of rotation and its magnitude is equal to $2\omega_0$. In other circumstances the vorticity is related in a similar fashion to the local angular velocity and may vary from place to place. As for the right-hand side of (70), $D\boldsymbol{v}/Dt$ represents the rate of change of velocity that one would see if the motion of a single element of the fluid could be followed—that is, it represents the acceleration of the element—while $\partial\boldsymbol{v}/\partial t$ represents the rate of change at a fixed point in space. If the flow is steady, then $\partial\boldsymbol{v}/\partial t$ is everywhere zero, but the fluid may be accelerating all the same, as individual fluid elements move from regions where the streamlines are widely spaced to regions where they are close together. It is the difference between $D\boldsymbol{v}/Dt$ and $\partial\boldsymbol{v}/\partial t$—*i.e.*, the final $(\boldsymbol{v}\cdot\nabla)\boldsymbol{v}$ term in (70)—that introduces into fluid dynamics the nonlinearity that makes the subject so rife with surprises.

Vorticity

POTENTIAL FLOW

This section is concerned with an important class of flow problems in which the vorticity is everywhere zero, and for such problems the Navier–Stokes equation may be greatly simplified. For one thing, the viscosity term drops out of it. For another, the nonlinear term, $(\boldsymbol{v}\cdot\nabla)\boldsymbol{v}$, may be transformed into $\nabla(v^2/2)$. Finally, it may be shown that when $(\nabla\times\boldsymbol{v})$ is zero, one may describe the velocity by means of a scalar potential φ, using the equation

$$\boldsymbol{v}=\nabla\varphi\ [\equiv \text{grad}\ \varphi]. \tag{71}$$

Thus (70) becomes

$$-\nabla\left(\frac{p}{\rho}+gz+\frac{v^2}{2}+\frac{\partial\varphi}{\partial t}\right)=0,$$

which may at once be integrated to show that

$$\left(\frac{p}{\rho}+gz+\frac{v^2}{2}+\frac{\partial\varphi}{\partial t}\right)=\text{constant}. \tag{72}$$

This result incorporates Bernoulli's law for an effectively incompressible fluid ([48]), as was to be expected from the disappearance of the viscosity term. It is more powerful than (48), however, because it can be applied to nonsteady flow in which $\partial\varphi/\partial t$ is not zero, and because it shows that in cases of potential flow the left-hand side of (72) is constant everywhere and not just constant along each streamline.

Vorticity-free, or potential, flow would be of rather limited interest were it not for the theorem, first proved by Thomson, that, in a body of fluid which is free of vorticity initially, the vorticity remains zero as the fluid moves. This theorem seems to open the door for relatively painless solutions to a great range of problems. Consider, for example, a stream of fluid in uniform motion approaching an obstacle of some sort. Well upstream of the obstacle the fluid is certainly vorticity-free, so it should, according to Thomson's theorem, be vorticity-free around the obstacle and downstream as well. In this case a flow potential should exist; and, if the fluid is effectively incompressible, it follows from equations (67) and (71) that it satisfies Laplace's equation,

Thomson's theorem

$$\left(\frac{\partial^2\varphi}{\partial x_1^2}\right)+\left(\frac{\partial^2\varphi}{\partial x_2^2}\right)+\left(\frac{\partial^2\varphi}{\partial x_3^2}\right)\ [\equiv\nabla^2\varphi]=0. \tag{73}$$

This is perhaps the most frequently occurring differential equation in physics, and methods for solving it, subject to appropriate boundary conditions, are very well established. Given a solution for φ, the fluid velocity $\boldsymbol{v}$ follows at once, and one may then discover how the pressure varies with position and time from equation (72).

The physicists and mathematicians who developed fluid dynamics during the 19th century relied heavily on this reasoning. They based splendid achievements upon it, a notable example being the theory of waves on deep water (see below). There was a touch of unreality, however, about some of their theorizing. If carried to extremes, the argument of the previous section implies that water initially stationary in a beaker can never be set into rotation by rotating the beaker or by stirring it with a spoon, and this is clearly nonsense. It suggests that vorticity-free water remains vorticity-free if it is squeezed into a narrow pipe, and this too is plainly nonsensical, for the well-established parabolic profile illustrated by Figure 63 is not vorticity-free. What is misleading about the argument in situations like these is that it pays inadequate attention to what happens at interfaces. Following the work of Prandtl, physicists now appreciate that vorticity is liable to be fed into the fluid at interfaces, whether these are interfaces between the fluid and some solid object or the free surfaces

of a liquid. Once the slightest trace of vorticity is present, it destroys the conditions on which the proof of Thomson's theorem depends. Moreover, vorticity admitted at interfaces spreads into the fluid in much the same way that a dye would spread, and whether or not the results of potential theory are useful depends on how much of the fluid is contaminated in the particular circumstances under discussion.

POTENTIAL FLOW WITH CIRCULATION: VORTEX LINES

The proof of Thomson's theorem depends on the concept of circulation, which Thomson introduced. This quantity is defined for a closed loop which is embedded in, and moves with, the fluid; denoted by K, it is the integral around the loop of $\boldsymbol{v} \cdot d\boldsymbol{l}$, where $d\boldsymbol{l}$ is an element of length along the loop. If the vorticity is everywhere zero, then so is the circulation around all possible loops, and vice versa. Thomson showed that K cannot change if the viscous term in (70) contributes nothing to the local acceleration, and it follows that both K and vorticity remain zero for all time.

Reference was made earlier to the sort of steady flow pattern that may be set up by rotating a cylindrical spindle in a fluid; the streamlines are circles around the spindle, and the velocity falls off like r^{-1}. This pattern of flow occurs naturally in whirlpools and typhoons, where the role of the spindle is played by a "core" in which the fluid rotates like a solid body; the axis around which the fluid circulates is then referred to as a vortex line. Each small element of fluid outside the core, if examined in isolation for a short interval of time, appears to be undergoing translation without rotation, and the local vorticity is zero. Were it not so, the viscous torques would not cancel and the flow pattern would not be a steady one. Nevertheless, the circulation is not zero if the loop for which it is defined is one that encloses the spindle or core. In such situations, a potential that obeys Laplace's equation outside the spindle or core can be found, but it is no longer, to use a technical term that may be familiar to some readers, single-valued.

Vortex lines

Readers who recognize this term are likely to have encountered it in the context of electromagnetism, and it is worth remarking that all the results of potential flow theory have electromagnetic analogues, in which streamlines become the lines of force of a magnetic field and vortex lines become lines of electric current. The analogy may be illustrated by reference to the Magnus effect.

Magnus effect

This effect (named for the German physicist and chemist H.G. Magnus, who first investigated it experimentally) arises when fluid flows steadily past a cylindrical spindle, with a velocity that at large distances from the spindle is perpendicular to the spindle's axis and uniformly equal to, say, v_0, while the spindle itself is steadily rotated. Rotation is communicated to the fluid, and in the steady state the circulation around any loop that encloses the spindle (and encloses a layer of fluid adjacent to the spindle within which the vorticity is nonzero and potential theory is inapplicable) has some nonzero value K. The streamlines that describe the steady flow pattern (outside that "boundary layer") have the form suggested by Figure 64, though the details naturally depend on the magnitude of v_0 and K. The flow pattern has stagnation points at P and P′ and, since the pressure is high at such points, the spindle may be expected to experience a downward force perpendicular both to its axis and to the direction of v_0. Detailed calculations confirm this expectation and show that the magnitude of the force, per unit length of the spindle, is

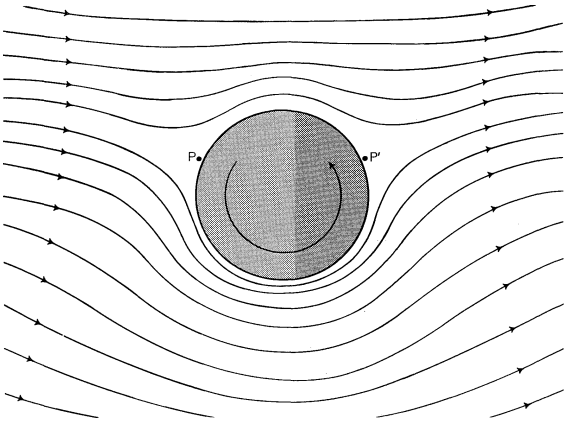


Figure 64: Streamlines for potential flow with circulation past a rotating cylinder. The cylinder experiences a downward Magnus force (see text).

$$\rho v_0 K. \tag{74}$$

This so-called Magnus force is directly analogous to the force that a transverse magnetic field B_0 exerts upon a wire carrying an electric current I, the magnitude of which, per unit length of the wire, is $B_0 I$.

The Magnus force on rotating cylinders has been utilized to propel experimental yachts, and it is closely related to the lift force on airfoils that enables airplanes to fly (see below *Lift*). The transverse forces that cause spinning balls to swerve in flight are, however, not Magnus forces, as is sometimes asserted. They are due to the asymmetrical nature of the eddies that develop at the rear of a spinning sphere (see below *Boundary layers and separation*). Cricket balls, unlike the balls used for baseball, tennis, and golf, have a raised equatorial seam that plays an important part in making the eddies asymmetric. A bowler in cricket who wants to make the ball swerve imparts spin to it, but he does so chiefly to ensure that the orientation of this seam remains steady as the ball moves toward the batsman.

It may be shown, by reference to the magnetic analogue or in other ways, that straight vortex lines of equal but opposite strength, $\pm K$, which are parallel and separated by a distance d, will drift sideways together through the fluid at a speed given by $K/2\pi d$. Similarly, a vortex line that has joined up on itself to form a closed vortex ring of radius a drifts along its axis with a speed given by

$$\left(\frac{K}{4\pi a}\right) \ln\left(\frac{a}{c}\right), \tag{75}$$

where c is the radius of the line's core, with ln standing for natural logarithm. This formula applies, for example, to smoke rings. The fact that such rings slow down as they propagate can be explained in terms of the increase of c with time, due to viscosity.

WAVES ON DEEP WATER

One particular solution of Laplace's equation that describes wave motion on the surface of a lake or of the ocean is

$$\varphi = \varphi_0 \cos\left\{2\pi\left[\left(\frac{x}{\lambda}\right) - ft\right]\right\} \sinh\left[2\pi\frac{(D+z)}{\lambda}\right]. \tag{76}$$

Here the x-axis is the direction of propagation and the z-axis is vertical; $z = 0$ describes the free surface of the water when it is undisturbed and $z = -D$ describes the bottom surface; φ_0 is an arbitrary constant that determines the amplitude of the motion; and f is the frequency of the waves and λ their wavelength. If λ is more than a few centimetres, surface tension is irrelevant and the pressure in the liquid just below its free surface is atmospheric for all x. It can be shown that in these circumstances the motion described by (76) is consistent with (72) only if the frequency and wavelength are related by the equation

$$f^2 = \left(\frac{2\pi g}{\lambda}\right) \tanh\left(\frac{2\pi D}{\lambda}\right), \tag{77}$$

and an expression for the speed of the waves may be deduced from this, since $V = f\lambda$. For shallow water ($D << \lambda$) one obtains the answer already quoted as equation (53), but for deep water ($D >> \lambda$) the answer is

$$V = \sqrt{\left(\frac{g\lambda}{2\pi}\right)}. \tag{78}$$

Waves on deep water are evidently dispersive, and surfers rely on this fact. A storm in the middle of the ocean disturbs the surface in a chaotic way that would be useless for surfing, but as the component waves travel toward the shore they separate; those with long wavelengths move ahead of those with short wavelengths because they travel faster. As a result, the waves seem nicely regular by the time that they arrive.

Dispersive character of waves on deep water

Anyone who has observed the waves behind a moving ship will know that they are confined to a V-shaped area of the water's surface, with the ship at its apex. The waves are particularly prominent on the arms of the V, but can also be discerned between these arms where the wave crests curve in the manner indicated in Figure 65. It seems to be widely believed that the angle of the V becomes more acute as the boat speeds up, much in the way that the conical shock wave accompanying a supersonic projectile becomes more acute (see Figure 61). That is not the case; the dispersive character of waves on deep water is such that the V has a fixed angle of $2\sin^{-1}(1/3) = 39°$. Thomson (Lord Kelvin) was the first to explain this, and so the V-shaped area is now known as the Kelvin wedge.

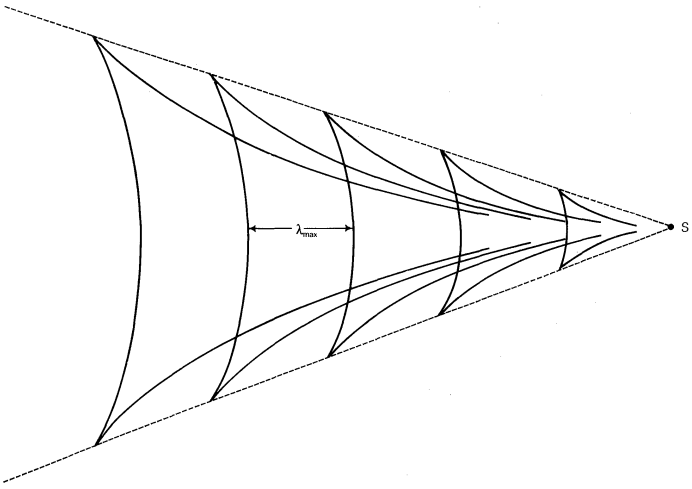


Figure 65: Wave crests in the Kelvin wedge behind a source S that is moving steadily from left to right. The maximum wavelength λ_{max} depends on the speed of the source, but the angle of the wedge does not (see text).

A version of Thomson's argument is illustrated by the diagram in Figure 66. Here S (the "source") represents the bow of the ship which is moving from left to right with uniform speed U, and the lines labeled C, C′, C″, etc., represent a set of parallel wave crests which are also moving from left to right. It can be shown that S will create this set of crests if, but only if, it rides continuously on the one labeled C. (It also can be shown that, though the crests in the set continue indefinitely to the left of C, there can be none to the right of this one.) The condition that S and C move together indicates that there is a relation between wavelength λ and inclination α expressed by the equation

$$\sin\alpha = \frac{V}{U} = \sqrt{\frac{g\lambda}{2\pi U^2}}. \tag{79}$$

This condition can evidently be satisfied by many other sets of crests besides the one represented by full lines in the figure—*e.g.*, by the set with slightly shorter wavelength λ' that is represented by broken lines. When one takes into consideration all the sets that satisfy (79) and have wavelengths intermediate between λ and λ', it becomes apparent that over most of the area behind the source they interfere destructively. They reinforce one another, however, near the intersections that are ringed in the figure. These intersections lie on a line through S of inclination β, where

$$\tan\beta = \frac{\tan\alpha}{(2+\tan^2\alpha)}. \tag{80}$$

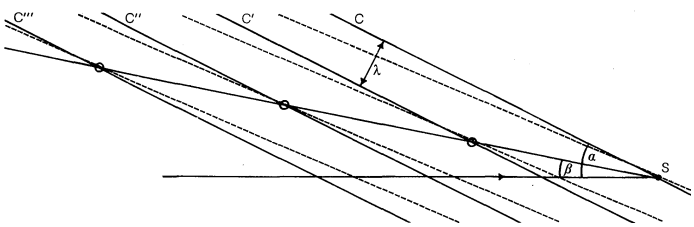


Figure 66: The curved wave crests of Figure 65 result from the superposition of many sets of straight wave crests like the two shown here. These two sets and others that are intermediate in wavelength reinforce one another near the line of inclination β and interfere destructively elsewhere.

It follows that, though the angle α can take any value between 90° (corresponding to $\lambda = \lambda_{max} = 2\pi U^2/g$) and zero, $\tan\beta$ can never exceed $\frac{1}{2}\sqrt{2}$, and $\sin\beta$ can never exceed $\frac{1}{3}$.

Ships lose energy to the waves in the Kelvin wedge, and they experience additional resistance on that account. The resistance is particularly high when the wave system created by the bow, where water is pushed aside, reinforces the wave system created by the "anti-source" at the stern, where the water closes in again. Such reinforcement is liable to occur when the effective length of the boat, L, is equal to $(2n+1)\lambda_{max}/2$ (with $n = 0, 1, 2, \ldots$) and therefore when the Froude number, $U/\sqrt{(Lg)}$, takes one of the values $[\sqrt{(2n+1)\pi}]^{-1}$. However, once a boat has been accelerated past $U = \sqrt{(Lg/\pi)}$, the bow and stern waves tend to cancel, and the resistance resulting from wave creation diminishes.

Waves on deep water whose wavelength is a few centimetres or less are generally referred to as ripples. In such waves, the pressure differences across the curved surface of the water associated with surface tension (see equation [44]) are not negligible, and the appropriate expression for their speed of propagation is

$$V = \sqrt{\left[\left(\frac{g\lambda}{2\pi}\right) + \left(\frac{2\pi\sigma}{\lambda\rho}\right)\right]}. \tag{81}$$

The wave velocity is therefore large for very short wavelengths as well as for very long ones. For water at normal temperatures, V has a minimum value of about 0.23 metre per second where the wavelength is about 17 millimetres, and it follows (note that equation [79] has no real root for α unless U exceeds V) that an object moving through water can create no ripples at all unless its speed exceeds 0.23 metre per second. A wind moving over the surface of water likewise creates no ripples unless its speed exceeds a certain critical value, but this is a more complicated phenomenon, and the critical speed in question is distinctly higher.

BOUNDARY LAYERS AND SEPARATION

It should be reiterated that vorticity is liable to enter a fluid that is initially undergoing potential flow where it makes contact with a solid and also at its free surface. The way in which, having entered, it spreads, may be illustrated by a simple example. Consider a large body of fluid, initially stationary, being set into motion by the motion in its own plane of a large solid plate immersed within the fluid. The motion is communicated from solid to fluid by the frictional forces that prevent slip between the two (see above *Viscosity*), and a velocity profile of the form suggested by Figure 67 is established. Its development with time turns out to be described by the partial differential equation

$$\rho\left(\frac{\partial v_1}{\partial t}\right) = \eta\left(\frac{\partial^2 v_1}{\partial x_2^2}\right). \tag{82}$$

In this situation the vorticity, which may be denoted by the symbol Ω, has one nonzero component, directed along the axis perpendicular to the diagram in Figure 67; it is $\Omega_3 = -(\partial v_1/\partial x_2)$. Differentiation of (82) with respect to x_2 shows at once that

$$\rho\left(\frac{\partial \Omega_3}{\partial t}\right) = \eta\left(\frac{\partial^2 \Omega_3}{\partial x_2^2}\right). \tag{83}$$

Diffusion equation

This is a diffusion equation. It indicates that if the plate oscillates to and fro with frequency f, then the so-called boundary layer within which Ω_3 is nonzero has a thickness δ given by

$$\delta \approx \sqrt{\left(\frac{\eta}{\pi\rho f}\right)}, \tag{84}$$

and in most instances of oscillatory motion this is small enough for the boundary layer to be neglected. For example, the boundary layer on the surface of the ocean has a

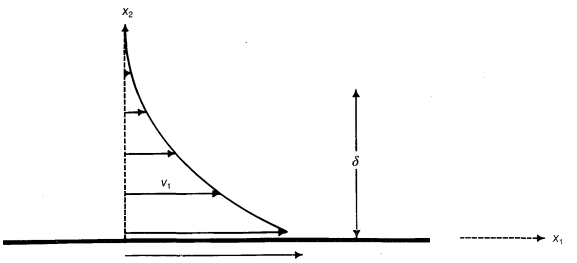


Figure 67: Velocity profile established by motion of a plate through stationary fluid (see text).

thickness of less than one millimetre when a wave with a frequency of about one hertz passes by; because the effects of viscosity are confined to this layer, they are too slight to affect the propagation of the wave to any significant degree. If the plate is kept moving at a uniform rate, however, the thickness of the boundary layer, as described by (83), will increase with the time t that has elapsed since the motion of the plate began, according to the equation

$$\delta \approx \sqrt{\left(\frac{2\eta t}{\rho}\right)}. \tag{85}$$

Prandtl suggested that when a stream of fluid flows steadily past an obstacle of finite extent, such as a sphere, the time that matters is the time for which fluid on a streamline just outside the boundary layer remains in contact with it. This time is of order D/v_0, where D is the diameter of the sphere and v_0 is the speed of the fluid well upstream. Hence, one would expect the thickness of the boundary layer at the rear of the sphere to be something like

$$\sqrt{\left(\frac{\eta D}{\rho v_0}\right)}. \tag{86}$$

If the velocity v_0 is so low that (85) is comparable with or greater than the diameter D, the flow pattern must be so contaminated by vorticity that the neglect of viscosity and reliance on Bernoulli's equation and on the other results of potential theory is clearly unjustified. If the velocity is high and (86) is much less than D, however, the boundary layer would seem to be of little importance. Surely then the results of potential theory are to be trusted?

Alas, that optimistic conclusion is not confirmed by experiment. What happens at high velocities is that the boundary layer comes unstuck from the surface of the sphere—it is said to separate. The reason why it does so is suggested by Figure 68A, which shows the streamlines to be expected when the boundary layer (shown in this figure by a shaded area still attached to the sphere) is relatively thin. Evidently the fluid velocity is higher near the equator of the sphere, at Q, than it is at either of the two poles, P and P′. Thus according to Bernoulli's equation, which can be relied on outside the boundary layer, the pressure near Q is less than it is near P and P′. The pressure gradient acts on the fluid in the boundary layer, accelerating it between P and Q but decelerating it between Q and P′. As the flow velocity increases, so does the pressure gradient, and at a certain stage the decelerating effect between Q and P′ becomes so large that the direction of flow within the boundary layer reverses in sign near the point labeled R in the diagram. The backflow of fluid near R causes an accumulation of fluid that obliges the oncoming boundary layer to separate, and the fluid behind the sphere circulates slowly within the boundary layer as a ring-shaped eddy (Figure 68B).

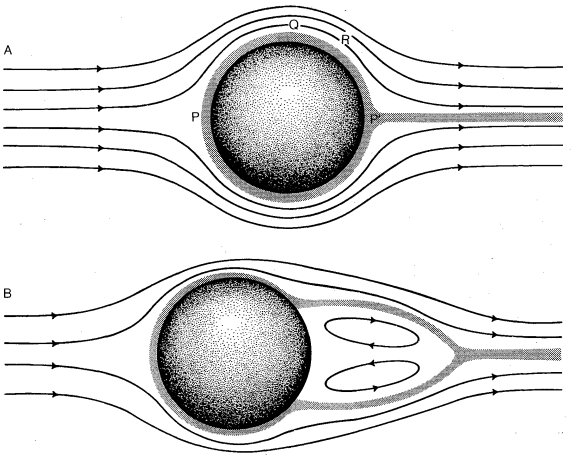


Figure 68: *Flow past a stationary solid sphere.* The fluid is free of vorticity outside a boundary layer, which is represented by shading. In (A) the boundary layer is still attached to the sphere, though it continues downstream of it. In (B) it has separated, and an eddy has formed behind the sphere.

The diagrams in Figure 68 might well refer to a cylinder rather than a sphere. If such were the case, however, the regions of circulating flow behind the obstacle that are shown in the second diagram would form parts of two separate straight eddies instead of a single ring-shaped one. At high velocities the eddies behind a cylinder become so large that they are blown off by the current and disappear downstream while new eddies form in their place; they are said to have been shed. The top and bottom eddies are shed alternately, and the cylinder experiences an oscillating force as a consequence. If the cylinder is something flexible like a telephone or power cable, it will move to and fro under this force; the singing noise produced by cables in high winds is due to a resonance between their natural frequency of transverse oscillation and the frequency of eddy shedding.

Eddy shedding

Similar processes are liable to occur behind obstacles of any shape, and the occurrence of eddies behind rocks or walls that interrupt the smooth flow of rivers is a familiar phenomenon.

DRAG

A fluid stream exerts a drag force F_D on any obstacle placed in its path, and the same force arises if the obstacle moves and the fluid is stationary. How large it is and how it may be reduced are questions of obvious importance to designers of moving vehicles of all sorts and equally to designers of cooling towers and other structures who want to be certain that the structures will not collapse in the face of winds.

An expression for the drag force on a sphere which is valid at such low velocities that the v^2 term in the Navier–Stokes equation is negligible, and thus at velocities such that the boundary layer thickness described by (86) is larger than the sphere diameter D, was first obtained by Stokes. Known as Stokes's law, it may be written as

Stokes's law

$$F_D = 3\pi\eta D v_0. \tag{87}$$

One-third of this force is transmitted to the sphere by shear stresses near the equator, and the remaining two-thirds are due to the pressure being higher at the front of the sphere than at the rear.

As the velocity increases and the boundary layer decreases in thickness, the effect of the shear stresses (or of what is sometimes called skin friction in this context) becomes less and less important compared with the effect of the pressure difference. It is impossible to calculate that difference precisely, except in the limit to which Stokes's law applies, but there are grounds for expecting that once eddies have formed it is about $\rho v_0^2/2$. Hence at high velocities one may expect

$$F_D = \left(\frac{\rho v_0^2}{2}\right) A', \tag{88}$$

where A' is some effective cross-sectional area, presumably comparable to its true cross-sectional area $A(= \pi D^2/4$ for a sphere) but not necessarily exactly equal to this. It is conventional to describe drag forces in terms of a dimensionless quantity called the drag coefficient; this is defined, irrespective of the shape of the body, as the ratio $[F_D/(\rho v_0^2/2)A]$ and is denoted by C_D. At high velocities, C_D is clearly the same thing as the ratio (A'/A) and should therefore be of order unity.

This is as far as theory can go with this problem. The principles of dimensional analysis can be invoked to show that, provided the compressibility of the fluid is irrelevant (*i.e.*, provided the flow velocity is well below the speed of sound), the drag coefficient must be some universal function of another dimensionless quantity known as the Reynolds number and defined as

$$R = \frac{\rho v_0 D}{\eta}. \qquad (89)$$

One must, however, resort to experiments to discover the form of this function. Fortunately, a limited number of experiments will suffice because the function is universal. They can be performed using whatever liquids and spheres are most convenient, provided that the whole range of R that is likely to be important is covered. Once the results have been plotted on a graph of C_D versus R, the graph can be used to predict the drag forces experienced by other spheres in other liquids at velocities that may be quite different from those so far employed. This point is worth emphasizing because it enshrines the principle of dynamical similarity, which is heavily relied on by engineers whenever they use results obtained with models to predict the behaviour of much larger structures.

The C_D versus R curve for spheres, plotted with logarithmic scales, is shown in Figure 69. Stokes's law, re-expressed in terms of C_D and R, becomes $C_D = 24/R$, and it is represented by the straight line on the left of the diagram. This law evidently fails when R exceeds about 1. There is a considerable range of R in the middle of the diagram over which C_D is about 0.5, but when R reaches about 3×10^{-5} it falls dramatically to about 0.1. The figure includes the corresponding curves for cylinders of diameter D whose axes are transverse to the direction of flow and for transverse disks of diameter D. The curve for cylinders is similar to that for spheres (though it has no straight-line part at low Reynolds number to correspond to Stokes's law), but the curve for disks is noticeably flatter. This flatness is linked to the fact that a disk has sharp edges around which the streamlines converge and diverge rapidly. The resulting large pressure gradients near the edge favour the formation and shedding of eddies. The drag force on a transverse flat plate of any shape can normally be estimated quite accurately, provided its edges are sharp, by assuming the drag coefficient to be unity.

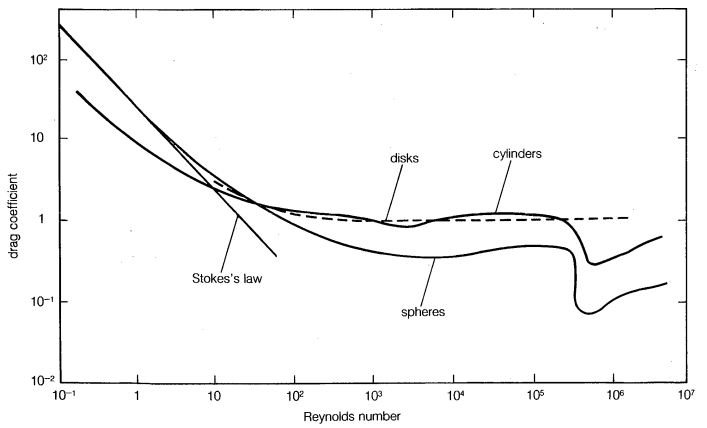


Figure 69: Variation of drag coefficient with Reynolds number for spheres, cylinders, and disks (see text).

Reducing drag

Since sharp edges favour the formation and shedding of eddies, and thereby increase the drag coefficient, one may hope to reduce the drag coefficient by streamlining the obstacle. It is at the rear of the obstacle that separation occurs, and it is therefore the rear that needs streamlining. By stretching this out in the manner suggested in Figure 70A, the pressure gradient acting on the boundary layer behind the obstacle can be much reduced. Other methods of reducing drag that have some practical applications are illustrated in Figures 70B and 70C. In 70B the obstacle is the wing of an aircraft with a slot through its leading edge; the current of air channeled through this slot imparts forward momentum to the fluid in the boundary layer on the upper surface of the wing to hinder this fluid from moving backward. The cowls that are often fitted to the leading edges of aircraft wings have a similar purpose. In Figure 70C, the obstacle is equipped with an internal device—a pump of some sort—which prevents the accumulation of boundary-layer fluid that would otherwise lead to separation by sucking it in through small holes in the surface of the obstacle, near Q; the fluid may be ejected again through holes near P′, where it will do no harm.

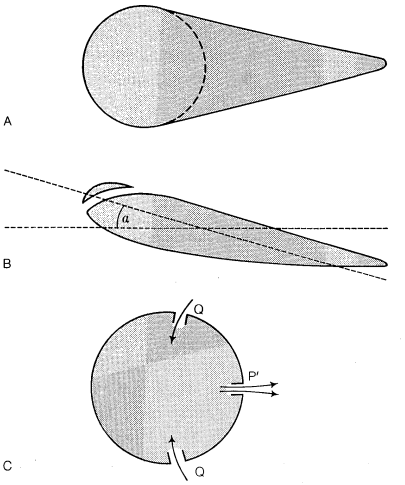


Figure 70: *Methods for reducing drag.* Each diagram represents a solid object that is stationary in the path of a fluid flowing from left to right, or that is moving from right to left through a fluid which is stationary: (A) a sphere that has been streamlined; (B) an aircraft wing, inclined at angle α, which is slotted along its leading edge; (C) a sphere equipped with an internal pump, which draws fluid in near Q and expels it at P′.

It should be stressed that the curves in Figure 69 are universal only so long as the velocity v_0 is much less than the speed of sound. When v_0 is comparable with the speed of sound, V_S, the compressibility of the fluid becomes relevant, which means that the drag coefficient has to be regarded as dependent on the dimensionless ratio $M = v_0/V_S$, known as the Mach number, as well as on the Reynolds number. The drag coefficient always rises as M approaches unity but may thereafter fall. To reduce drag in the supersonic region, it pays to streamline the front of obstacles or projectiles rather than the rear, as this reduces the intensity of the shock cone (see above *Compressible flow in gases*).

LIFT

If an aircraft wing, or airfoil, is to fulfill its function, it must experience an upward lift force, as well as a drag force, when the aircraft is in motion. The lift force arises because the speed at which the displaced air moves over the top of the airfoil (and over the top of the attached boundary layer) is greater than the speed at which it moves over the bottom and because the pressure acting on the airfoil from below is therefore greater than the pressure from above. It also can be seen, however, as an inevitable consequence of the finite circulation that exists around the airfoil. One way to establish circulation around an obstacle is to rotate it, as was seen earlier in the description of the Magnus effect. The circulation around an airfoil, however, is created by its forward motion; it arises as soon as the airfoil moves fast enough to shed its first eddy.

The lift force on an airfoil moving through stationary air at a steady speed v_0 is the same as the lift force on an identical airfoil that is stationary in air moving at v_0 the other way; the latter is easier to represent pictorially. Figure 71A shows a set of streamlines representing potential flow past a stationary inclined plate before any eddy has been shed. The pattern is a symmetrical one, and the pressure variations associated with it generate neither drag nor lift. At the rear of the plate, however, the streamlines diverge rapidly, so conditions exist for the

formation of an eddy there and the sense of its rotation will be counterclockwise. It grows more easily and is shed more quickly because the edges of the plate are sharp. Figure 71B shows some streamlines for the same plate a moment after shedding when the detached eddy, known as the starting vortex, is still in view. The circulation around the closed loop shown by a broken curve in this diagram was zero before the eddy formed and, according to Thomson's theorem (see above), it must still be zero. Passing through this loop, there thus must be a vortex line having clockwise circulation $-K$ to compensate for the circulation $+K$ of the starting vortex. This other line, known as the bound vortex, is not immediately apparent in the diagram because it is attached to the plate, and it remains thus attached as the starting vortex is swept away downstream. It does show up, however, in a modification of the flow pattern immediately behind the plate, where the streamlines no longer diverge as they do in Figure 71A. Because the divergence here has been eliminated, no further eddies are likely to be formed.

Starting vortex and bound vortex

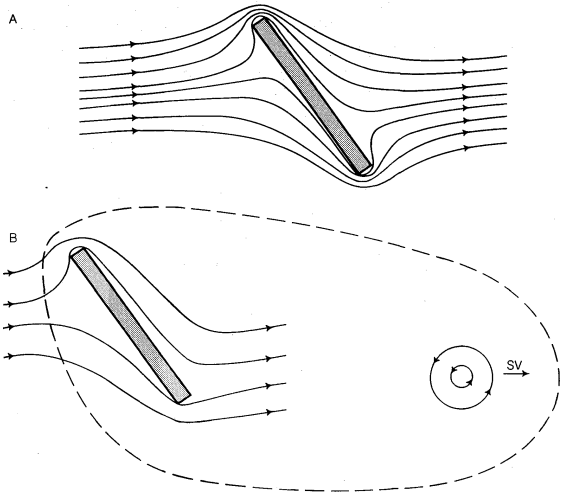


Figure 71: *Generation of lift force.*
(A) Streamlines for potential flow past a stationary inclined plate; conditions exist for formation of an eddy behind the plate, with counterclockwise rotation. (B) An eddy formed behind the plate in (A) has been shed as a starting vortex (SV), which is being carried downstream. The circulation remains zero around the large loop indicated by a broken line, and the streamlines no longer diverge behind the rear edge of the plate, as they do in (A) (see text).

In an earlier section, the formula $\rho v_0 K$ was quoted for the strength of the Magnus force per unit length of a rotating cylinder, and the same formula can be applied to the inclined plate in Figure 71B or to any airfoil that has shed a starting vortex and around which, consequently, there is circulation. The validity of the formula does not depend in any way on the precise shape of the airfoil, any more than the force exerted by a magnetic field on a wire carrying a current depends on the cross-sectional shape of the wire. The design of the airfoil, nevertheless, has a critical effect on the magnitude of the lift force because it determines the magnitude of K. The sort of cross section that is adopted for the wings of aircraft has been sketched already in Figure 70B. The rear edge is made as sharp as possible for reasons that have already been explained, and it may take the form of hinged flaps that are lowered at takeoff. Lowering the flaps increases K and therefore increases the lift, but the flaps need to be raised when the aircraft has reached its cruising altitude because they cause undesirable drag. The circulation and the lift can also be increased by increasing the angle α (see Figure 70B) at which the main part of the airfoil is inclined to the direction of motion. There is a limit to the lift that can be generated in this way, however, for if the inclination is too great the boundary layer separates behind the wing's leading edge, and the bound vortex, on which the lift depends, may be shed as a result. The aircraft is then said to stall. The leading edge is made as smooth and rounded as possible to discourage stalling.

Thomson's theorem can be used to prove that if the airfoil is of finite length then the starting vortex and the bound vortex must both be parts of a single, continuous vortex ring. They are joined by two trailing vortices, which run backward from the ends of the airfoil. As time passes, these trailing vortices grow steadily longer, and more and more energy is needed to feed the swirling motion of the fluid around them. It is clear, at any rate in the case where the airfoil is moving and the air is stationary, that this energy can come only from whatever agency propels the airfoil forward, and hence that the trailing vortices are a source of additional drag. The magnitude of the additional drag is proportional to K^2 but it does not increase, as the lift force does, if the airfoil is made longer while K is kept the same. For this reason, designers who wish to maximize the ratio of lift to drag will make the wings of their aircraft as long as they can—as long, that is, as is consistent with strength and rigidity requirements.

Trailing vortices

When a yacht is sailing into the wind, its sail acts as an airfoil of which the mast is the leading edge, and the considerations that favour long wings for aircraft favour tall masts as well.

TURBULENCE

The nonlinear nature of the $(\boldsymbol{v} \cdot \nabla)\boldsymbol{v}$ term in the Navier–Stokes equation—equation (70)—means that solutions of this equation cannot be superposed. The fact that $\boldsymbol{v}_1(\boldsymbol{R}, t)$ and $\boldsymbol{v}_2(\boldsymbol{R}, t)$ satisfy the equation does not ensure that $(\boldsymbol{v}_1 + \boldsymbol{v}_2)$ does so too. The nonlinear term provides a contact, in fact, through which two different modes of motion may exchange energy, so that one grows in amplitude at the expense of the other. A great deal of experimental and theoretical work has shown, in particular, that if a fluid is undergoing regular laminar motion (of the sort that was discussed in connection with Poiseuille's law, for example) at sufficiently high rates of shear, small periodic perturbations of this motion are liable to grow parasitically. Perturbations on a smaller scale still grow parasitically on those that are first established, until the flow pattern is so grossly disturbed that it is no longer useful to define a fluid velocity for each point in space; the description of the flow has to be a statistical one in terms of mean values and of correlated fluctuations about the mean. The flow is then said to be turbulent.

Turbulent flow

In the case (to which Poiseuille's law applies) of laminar flow through a uniform cylindrical pipe of diameter D, turbulence inevitably sets in when the Reynolds number R reaches a critical value that is about 10^5; in this context, the Reynolds number is defined (compare equation [89]) as

$$R = \frac{4\rho Q}{\pi D \eta} = \frac{\rho D \langle v \rangle}{\eta}, \qquad (90)$$

where Q is the rate of discharge and $\langle v \rangle$ is the mean fluid velocity. Turbulence sets in at much lower velocities, however, if the end of the pipe where the fluid enters is not carefully flared. The critical value of the Reynolds number for a pipe with a bluff entry may be as low as 2300, and this corresponds to a rate of discharge through a pipe for which D is, say, two centimetres, of only about three litres per minute. Thus pipe flow in engineering practice is more often turbulent than not. Once turbulence has set in, Q increases less rapidly with pressure gradient than Poiseuille's equation—equation (65)—predicts; it increases roughly as the square root of the pressure gradient or slightly more rapidly than this if the internal surface of the pipe is very smooth.

Turbulence arises not only in pipes but also within boundary layers around solid obstacles when the rate of shear within the boundary layer becomes large enough. Curiously enough, the onset of turbulence in the boundary layer can reduce the drag force on obstacles. In the case of a spherical obstacle, the point at which the boundary layer separates from the rear surface of the sphere shifts backward when the boundary layer becomes turbulent, away from the equator Q in Figure 68 and toward P′, and the eddies attached to the sphere therefore become smaller. It is turbulence in the boundary layer that is responsible for the dramatic drop in the drag coefficient for both spheres

Turbulence within boundary layers

and cylinders that occurs, as can be seen from Figure 69, when the Reynolds number is about 3×10^5. This drop enables golf balls to travel farther than they would do otherwise, and the dimples on the surface of golf balls are meant to encourage turbulence in the boundary layer. If swimsuits with rough surfaces help swimmers to move faster, as has been claimed, the same explanation may apply.

Where conditions for turbulence exist, flow rates of water through tubes may be increased and the drag forces exerted on obstacles by water diminished by dissolving small amounts of suitable polymers in the water. This is surprising because such additives increase viscosity, and in the preturbulent regime to which Poiseuille's law applies, their effect on the flow rate is quite the reverse. As has already been stated, the small perturbations that arise in a turbulent fluid tend to collapse into smaller perturbations and then into smaller perturbations still, until the motion is turbulent on a very fine scale—*i.e.*, on the scale of molecular dimensions—and until the energy stored in the perturbations is finally dissipated as heat. Polymer molecules seem to have the effect they do because, over the relatively large distances to which each such molecule extends, they impose a coherence on the fluid motion that would not otherwise be present.

Effect of polymer additives

CONVECTION

Apart from some remarks in the above section *Compressible flow in gases* about the circulation of the atmosphere, no attention has yet been paid to situations in which temperature differences are imposed upon a fluid by contact with hot and cold bodies. This subject will be briefly taken up here.

Consider first the case of two vertical plates with fluid between them, one at temperature T_1 and the other at T_2, in the presence of a vertical gravitational field. The hotter plate might be a domestic radiator and the colder plate the wall to which it is fixed. Thermal conduction ensures that the layer of air adjacent to the radiator is hotter than the rest of the air, and thermal expansion ensures that it is less dense. Consequently, the vertical pressure gradient which satisfies equation (38) in the rest of the air is too large to keep the layer adjacent to the radiator in equilibrium; that layer rises and, similarly, the cold layer adjacent to the wall falls. A circulating pattern of thermal convection is thereby established, and because this brings colder air into contact with the radiator, the rate at which heat is lost from the radiator is enhanced. The heat loss, once convection has been established, depends in a complicated manner on the separation between the plates (D) and on the thermal diffusivity (κ), specific heat, density, thermal expansion coefficient (α), and viscosity of the fluid. The heat loss also depends on $(T_1 - T_2)$, of course, and it is worthwhile noting that the manner in which it does so is not linear; the heat loss increases more rapidly than the temperature difference. Newton's law of cooling, which postulates a linear relationship, is obeyed only in circumstances where convection is prevented or in circumstances where it is forced (when a radiator is fan-assisted, for example).

Imagine a situation in which the same two plates are horizontal rather than vertical. In such a case, no convection can occur if the hot plate is above the cold one, and it is not obvious that it occurs in the reverse situation. Whether it does so or not depends on the magnitude of the temperature difference through a dimensionless combination of some of the relevant parameters, $\rho g \alpha D^3 (T_1 - T_2)/\eta\kappa$, which is known as the Rayleigh number. If the Rayleigh number is less than 1,708, the fluid is stable—or perhaps it would be more accurate to say that it is metastable—even though it is warmer at the bottom than at the top. However, when 1,708 is exceeded, a pattern of convective rolls known as Bénard cells is established between the plates. Evidence for the existence of such cells in the convecting atmosphere is sometimes seen in the regular columns of cloud that form over regions where the air is rising. Their periodicity can be astonishingly uniform.

Bénard cells

Macroscopic instabilities of a convective nature, of which the formation of Bénard cells provides just one example, are a feature of the oceans as well as of the atmosphere. These instabilities are frequently associated with gradients of salinity rather than with gradients of temperature. One cannot embark on a serious discussion of atmospheric and oceanic circulation on the Earth, however, without devoting more space to the dynamics of rotating fluids than is available here. (T.E.F.)

QUANTUM MECHANICS

Quantum mechanics deals with the behaviour of matter and light on the atomic and subatomic scale. It attempts to describe and account for the properties of molecules and atoms and their constituents—electrons, protons, neutrons, and other more esoteric particles such as quarks and gluons. These properties include the interactions of the particles with one another and with electromagnetic radiation (*i.e.*, light, X rays and gamma rays).

The behaviour of matter and radiation on the atomic scale often seems peculiar, and the consequences of quantum theory are accordingly difficult to understand and to believe. Its concepts frequently conflict with commonsense notions derived from observations of the everyday world. There is no reason, however, why the behaviour of the atomic world should conform to that of the familiar, large-scale world. It is important to realize that quantum mechanics is a branch of physics and that the business of physics is to describe and account for the way the world—on both the large and the small scale—actually is and not how one imagines it or would like it to be.

The study of quantum mechanics is rewarding for several reasons. First, it illustrates the essential methodology of physics. Second, it has been enormously successful in giving correct results in practically every situation to which it has been applied. There is, however, an intriguing paradox. In spite of the overwhelming practical success of quantum mechanics, the foundations of the subject contain unresolved problems—in particular, problems concerning the nature of measurement. An essential feature of quantum mechanics is that it is generally impossible, even in principle, to measure a system without disturbing it; the detailed nature of this disturbance and the exact point at which it occurs are obscure and controversial. For one or a combination of these reasons, quantum mechanics has attracted some of the ablest scientists of the 20th century, and they have erected what is perhaps the finest intellectual edifice of the period.

Historical basis of quantum theory

BASIC CONSIDERATIONS

At a fundamental level, both radiation and matter have characteristics of particles and waves. The gradual recognition by scientists that radiation has particle-like properties and that matter has wavelike properties provided the impetus for the development of quantum mechanics. Influenced by Newton, most physicists of the 18th century believed that light consisted of particles, which they called corpuscles. From about 1800, evidence began to accumulate for a wave theory of light. At about this time Thomas Young showed that if monochromatic light passes through a pair of slits, the two emerging beams interfere, so that a fringe pattern of alternately bright and dark bands appears on a screen. The bands are readily explained by a wave theory of light. According to the theory, a bright band is produced when the crests (or troughs) of the waves from the two slits arrive together at the screen; a dark band is produced when the crest of one wave arrives at the same time as the trough of the other, and the effects of the two light beams cancel. Beginning in 1815, a series of experiments by Augustin-Jean Fresnel of France and others showed that, when a parallel beam of light passes through

Support for a wave theory of light

a single slit, the emerging beam is no longer parallel but starts to diverge; this phenomenon is known as diffraction. Given the wavelength of the light and the geometry of the apparatus (*i.e.*, the separation and widths of the slits and the distance from the slits to the screen), one can use the wave theory to calculate the expected pattern in each case; the theory agrees precisely with the experimental data.

EARLY DEVELOPMENTS

Planck's radiation law. By the end of the 19th century, physicists almost universally accepted the wave theory of light. However, though the ideas of classical physics explain interference and diffraction phenomena relating to the propagation of light, they do not account for the absorption and emission of light. All bodies radiate electromagnetic energy as heat; in fact, a body emits radiation at all wavelengths. The energy radiated at different wavelengths is a maximum at a wavelength that depends on the temperature of the body; the hotter the body, the shorter the wavelength for maximum radiation. Attempts to calculate the energy distribution for the radiation from a blackbody using classical ideas were unsuccessful. (A blackbody is a hypothetical ideal body or surface that absorbs and reemits all radiant energy falling on it.) One formula, proposed by Wilhelm Wien of Germany, did not agree with observations at long wavelengths, and another, proposed by Lord Rayleigh (John William Strutt) of England, disagreed with those at short wavelengths.

In 1900 the German theoretical physicist Max Planck made a bold suggestion. He assumed that the radiation energy is emitted, not continuously, but rather in discrete packets called quanta. The energy E of the quantum is related to the frequency ν by

Concept of quantum

$$E = h\nu. \tag{91}$$

The quantity h, now known as the Planck constant, is a universal constant with the approximate value $h = 6.63 \times 10^{-34}$ joule-second. (The more precise, currently accepted value is 6.626075×10^{-34} joule-second.) Planck showed that the calculated energy spectrum then agreed with observation over the entire wavelength range.

Einstein and the photoelectric effect. In 1905 Einstein extended Planck's hypothesis to explain the photoelectric effect, which is the emission of electrons by a metal surface when it is irradiated by light or X rays. The kinetic energy of the emitted electrons depends on the frequency ν of the radiation, not on its intensity; for a given metal, there is a threshold frequency ν_0 below which no electrons are emitted. Furthermore, emission takes place as soon as the light shines on the surface; there is no detectable delay. Einstein showed that these results can be explained by two assumptions: (1) that light is composed of corpuscles or photons, the energy of which is given by Planck's relationship, and (2) that an atom in the metal can absorb either a whole photon or nothing. Part of the energy of the absorbed photon frees an electron, which requires a fixed energy W, known as the work function of the metal; the rest is converted into the kinetic energy $\frac{1}{2}m_e u^2$ of the emitted electron (m_e is the mass of the electron and u is its velocity). Thus, the energy relation is

$$h\nu = W + \frac{1}{2}m_e u^2. \tag{92}$$

If ν is less than ν_0, where $h\nu_0 = W$, no electrons are emitted. Not all the experimental results mentioned above were known in 1905, but all of Einstein's predictions have been verified since.

Bohr's theory of the atom. A major contribution to the subject was made by Niels Bohr of Denmark, who applied the quantum hypothesis to atomic spectra in 1913. The spectra of light emitted by gaseous atoms had been studied extensively since the mid-19th century. It was found that radiation from gaseous atoms at low pressure consists of a set of discrete wavelengths. This is quite unlike the radiation from a solid, which is distributed over a continuous range of wavelengths. The set of discrete wavelengths from gaseous atoms is known as a line spectrum, because the image of the linear slit in the spectrometer is a series of sharp lines. The wavelengths of the lines are characteristic of the element and may form extremely complex patterns. The simplest spectra are those of atomic hydrogen and the alkali atoms (*e.g.*, lithium, sodium, and potassium). For hydrogen, the wavelengths λ are given by the empirical formula

Line spectrum

$$\frac{1}{\lambda} = R\left(\frac{1}{m^2} - \frac{1}{n^2}\right), \tag{93}$$

where m and n are positive integers with $n > m$ and R, known as the Rydberg constant, has the value 1.10×10^7 per metre. For a given value of m, the lines for varying n form a series. The lines for $m = 1$, the Lyman series, lie in the ultraviolet part of the spectrum. Those for $m = 2$, the Balmer series, lie in the visible spectrum, and so on.

Bohr started with a model suggested by the New Zealand-born British physicist Ernest Rutherford. The model was based on the experiments of Hans Geiger and Ernest Marsden, who in 1909 bombarded gold atoms with massive, fast-moving alpha particles; when some of these particles were deflected backward, Rutherford concluded that the atom has a massive, charged nucleus. In Rutherford's model, the atom resembles a miniature solar system with the nucleus acting as the Sun and the electrons as the circulating planets. Bohr made three assumptions. First, he postulated that, in contrast to classical mechanics where an infinite number of orbits is possible, an electron can only be in one of a discrete set of orbits, which he termed stationary states. Second, he postulated that the only orbits allowed are those for which the angular momentum of the electron is a whole number n times $\hbar$ ($\hbar$ stands for $h/2\pi$). Third, Bohr assumed that Newton's laws of motion, so successful in calculating the paths of the planets around the Sun, also applied to electrons orbiting the nucleus. The force on the electron (the analogue of the gravitational force between the Sun and a planet) is the electrostatic attraction between the positively charged nucleus and the negatively charged electron. With these simple assumptions, he showed that the energy of the orbit has the form

$$E_n = -\frac{E_0}{n^2}, \tag{94}$$

where E_0 is a constant which may be expressed by a combination of the known constants e, m_e, and $\hbar$. While in a stationary state, the atom does not give off energy as light; however, when an electron makes a transition from a state with energy E_n to one with lower energy E_m, a quantum of energy is radiated with frequency ν, given by the equation

$$h\nu = E_n - E_m. \tag{95}$$

Inserting the expression for E_n into this equation and using the relation $\lambda\nu = c$, where c is the speed of light, Bohr derived the formula for the wavelengths of the lines in the hydrogen spectrum, with the correct value of the Rydberg constant.

Bohr's theory was a brilliant step forward. Its two most important features have survived in present-day quantum mechanics. They are (1) the existence of stationary, nonradiating states, and (2) the relationship of radiation frequency to the energy difference between the initial and final states in a transition. Prior to Bohr, physicists had thought that the radiation frequency would be the same as the electron's frequency of rotation in an orbit.

Most important features of Bohr's theory

Scattering of X rays. Soon scientists were faced with the fact that another form of radiation, X rays, also exhibit both wave and particle properties. Max von Laue of Germany had shown in 1912 that crystals can be used as three-dimensional diffraction gratings for X rays; his technique constituted the fundamental evidence for the wavelike nature of X rays. The atoms of a crystal, which are arranged in a regular lattice, scatter the X rays. For certain directions of scattering, all the crests of the X rays coincide. (The scattered X rays are said to be in phase and to give constructive interference.) For these directions, the scattered X-ray beam is very intense. Clearly, this phenomenon demonstrates wave behaviour. In fact, given the interatomic distances in the crystal and the directions of constructive interference, the wavelength of the waves can be calculated.

In 1922 the American physicist Arthur Holly Compton showed that X rays scatter from electrons as if they are particles. Compton performed a series of experiments on the scattering of monochromatic, high-energy X rays by graphite. He found that part of the scattered radiation had the same wavelength λ_0 as the incident X rays but that there was an additional component with a longer wavelength λ. To interpret his results, Compton regarded the X-ray photon as a particle that collides and bounces off an electron in the graphite target as though the photon and the electron were a pair of (dissimilar) billiard balls. Application of the laws of conservation of energy and momentum to the collision leads to a specific relation between the amount of energy transferred to the electron and the angle of scattering. For X rays scattered through an angle θ, the wavelengths λ and λ_0 are related by the equation

$$\lambda - \lambda_0 = \frac{h}{m_e c}(1 - \cos\theta). \tag{96}$$

The experimental correctness of Compton's formula is direct evidence for the corpuscular behaviour of radiation.

De Broglie's wave hypothesis. Faced with evidence that electromagnetic radiation has both particle and wave characteristics, Louis-Victor de Broglie of France suggested a great unifying hypothesis in 1924. De Broglie proposed that matter has wave, as well as particle, properties. He suggested that material particles can behave as waves and that their wavelength λ is related to the linear momentum p of the particle by

Wave-particle hypothesis

$$\lambda = \frac{h}{p}. \tag{97}$$

In 1927 Clinton Davisson and Lester Germer of the United States confirmed de Broglie's hypothesis for electrons. Using a crystal of nickel, they diffracted a beam of monoenergetic electrons and showed that the wavelength of the waves is related to the momentum of the electrons by the de Broglie equation. Since Davisson and Germer's investigation, similar experiments have been performed with atoms, molecules, neutrons, protons, and many other particles. All behave like waves with the same wavelength–momentum relationship.

Basic concepts and methods

Bohr's theory, which assumed circular orbits, was extended by the German physicist Arnold Sommerfeld and others to include elliptic orbits and other refinements. Attempts were made to apply the theory to more complicated systems than the hydrogen atom. However, the ad hoc mixture of classical and quantum ideas made the theory and calculations increasingly unsatisfactory. Then in the 12 months starting in July 1925, a period of creativity without parallel in the history of physics, there appeared a series of papers by German scientists that set the subject on a firm conceptual foundation. The papers took two approaches: (1) matrix mechanics, proposed by Werner Heisenberg, Max Born, and Pascual Jordan; and (2) wave mechanics, put forward by Erwin Schrödinger. The protagonists were not always polite to each other. Heisenberg found the physical ideas of Schrödinger's theory "disgusting," and Schrödinger was "discouraged and repelled" by the lack of visualization in Heisenberg's method. However, Schrödinger, not allowing his emotions to interfere with his scientific endeavours, showed that, in spite of apparent dissimilarities, the two theories are equivalent mathematically. The present discussion follows Schrödinger's wave mechanics because it is less abstract and easier to understand than Heisenberg's matrix mechanics.

SCHRÖDINGER'S WAVE MECHANICS

Schrödinger expressed de Broglie's hypothesis concerning the wave behaviour of matter in a mathematical form that is adaptable to a variety of physical problems without additional arbitrary assumptions. He was guided by a mathematical formulation of optics, in which the straight-line propagation of light rays can be derived from wave motion when the wavelength is small compared to the dimensions of the apparatus employed. In the same way, Schrödinger set out to find a wave equation for matter that would give particle-like propagation when the wavelength becomes comparatively small. According to classical mechanics, if a particle of mass m_e is subjected to a force such that its potential energy is $V(x,y,z)$ at position x,y,z, then the sum of $V(x,y,z)$ and the kinetic energy $p^2/2m_e$ is equal to a constant, the total energy E of the particle. Thus,

$$\frac{p^2}{2m_e} + V(x,y,z) = E. \tag{98}$$

It is assumed that the particle is bound—*i.e.*, confined by the potential to a certain region in space because its energy E is insufficient for it to escape. Since the potential varies with position, two other quantities do also: the momentum and, hence, by extension from the de Broglie relation, the wavelength of the wave. Postulating a wave function $\Psi(x,y,z)$ that varies with position, Schrödinger replaced p in the above energy equation with a differential operator that embodied the de Broglie relation. He then showed that Ψ satisfies the partial differential equation

$$-\frac{\hbar^2}{2m_e}\left(\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2}\right) + V(x,y,z)\Psi = E\Psi. \tag{99}$$

This is the (time-independent) Schrödinger wave equation of 1926, which established quantum mechanics in a widely applicable form. An important advantage of Schrödinger's theory is that no further arbitrary quantum conditions need be postulated. The required quantum results follow from certain reasonable restrictions placed on the wave function—for example, that it should not become infinitely large at large distances from the centre of the potential.

Time-independent Schrödinger wave equation

Schrödinger applied his equation to the hydrogen atom, for which the potential function, given by classical electrostatics, is proportional to $-e^2/r$, where $-e$ is the charge on the electron. The nucleus (a proton of charge e) is situated at the origin, and r is the distance from the origin to the position of the electron. Schrödinger solved the equation for this particular potential with straightforward, though not elementary, mathematics. Only certain discrete values of E lead to acceptable functions Ψ. These functions are characterized by a trio of integers n, l, m, termed quantum numbers. The values of E depend only on the integers n (1, 2, 3, etc.) and are identical with those given by the Bohr theory. The quantum numbers l and m are related to the angular momentum of the electron; $\sqrt{l(l+1)}\hbar$ is the magnitude of the angular momentum, and $m\hbar$ is its component along some physical direction.

The square of the wave function, Ψ^2, has a physical interpretation. Schrödinger originally supposed that the electron was spread out in space and that its density at point x,y,z was given by the value of Ψ^2 at that point. Almost immediately Born proposed what is now the accepted interpretation—namely, that Ψ^2 gives the probability of finding the electron at x,y,z. The distinction between the two interpretations is important. If Ψ^2 is small at a particular position, the original interpretation implies that a small fraction of an electron will always be detected there. In Born's interpretation, nothing will be detected there most of the time, but, when something is observed, it will be a whole electron. Thus, the concept of the electron as a point particle moving in a well-defined path around the nucleus is replaced in wave mechanics by clouds that describe the probable locations of electrons in different states.

ELECTRON SPIN AND ANTIPARTICLES

In 1928 the English physicist Paul A.M. Dirac produced a wave equation for the electron that combined relativity with quantum mechanics. Schrödinger's wave equation does not satisfy the requirements of the special theory of relativity because it is based on a nonrelativistic expression for the kinetic energy ($p^2/2m_e$). Dirac showed that an electron has an additional quantum number m_s. Unlike the first three quantum numbers, m_s is not a whole integer and can have only the values $+\frac{1}{2}$ and $-\frac{1}{2}$. It corresponds to an additional form of angular momentum ascribed to a spinning motion. (The angular momentum mentioned

Dirac's relativistic wave equation

above is due to the orbital motion of the electron, not its spin.) The concept of spin angular momentum was introduced in 1925 by Samuel A. Goudsmit and George E. Uhlenbeck, two graduate students at the University of Leiden, Neth., to explain the magnetic moment measurements made by Otto Stern and Walther Gerlach of Germany several years earlier. The magnetic moment of a particle is closely related to its angular momentum; if the angular momentum is zero, so is the magnetic moment. Yet Stern and Gerlach had observed a magnetic moment for electrons in silver atoms, which were known to have zero orbital angular momentum. Goudsmit and Uhlenbeck proposed that the observed magnetic moment was attributable to spin angular momentum.

The electron-spin hypothesis not only provided an explanation for the observed magnetic moment but also accounted for many other effects in atomic spectroscopy, including changes in spectral lines in the presence of a magnetic field (Zeeman effect), doublet lines in alkali spectra, and fine structure (close doublets and triplets) in the hydrogen spectrum.

The Dirac equation also predicted additional states of the electron that had not yet been observed. Experimental confirmation was provided in 1932 by the discovery of the positron by the American physicist Carl David Anderson. Every particle described by the Dirac equation has to have a corresponding antiparticle, which differs only in charge. The positron is just such an antiparticle of the negatively charged electron, having the same mass as the latter but a positive charge.

Discovery of the positron

IDENTICAL PARTICLES AND MULTIELECTRON ATOMS

Because electrons are identical to (*i.e.,* indistinguishable from) each other, the wave function of an atom with more than one electron must satisfy special conditions. The problem of identical particles does not arise in classical physics, where the objects are large-scale and can always be distinguished, at least in principle. There is no way, however, to differentiate two electrons in the same atom, and the form of the wave function must reflect this fact. The overall wave function Ψ of a system of identical particles depends on the coordinates of all the particles. If the coordinates of two of the particles are interchanged, the wave function must remain unaltered or, at most, undergo a change of sign; the change of sign is permitted because it is Ψ^2 that occurs in the physical interpretation of the wave function. If the sign of Ψ remains unchanged, the wave function is said to be symmetric with respect to interchange; if the sign changes, the function is antisymmetric.

The symmetry of the wave function for identical particles is closely related to the spin of the particles. In quantum field theory (see below *Quantum electrodynamics*), it can be shown that particles with half-integral spin ($1/2$, $3/2$, etc.) have antisymmetric wave functions. They are called fermions after the Italian-born physicist Enrico Fermi. Examples of fermions are electrons, protons, and neutrons, all of which have spin $1/2$. Particles with zero or integral spin (*e.g.,* mesons, photons) have symmetric wave functions and are called bosons after the Indian mathematician and physicist Satyendra Nath Bose, who first applied the ideas of symmetry to photons in 1924–25.

The requirement of antisymmetric wave functions for fermions leads to a fundamental result, known as the exclusion principle, first proposed in 1925 by the Austrian physicist Wolfgang Pauli. The exclusion principle states that two fermions in the same system cannot be in the same quantum state. If they were, interchanging the two sets of coordinates would not change the wave function at all, which contradicts the result that the wave function must change sign. Thus, two electrons in the same atom cannot have an identical set of values for the four quantum numbers n, l, m, m_s. The exclusion principle forms the basis of many properties of matter, including the periodic classification of the elements, the nature of chemical bonds, and the behaviour of electrons in solids; the last determines in turn whether a solid is a metal, insulator, or semiconductor (see ATOMS; MATTER).

Exclusion principle

The Schrödinger equation cannot be solved precisely for atoms with more than one electron. The principles of the calculation are well understood, but the problems are complicated by the number of particles and the variety of forces involved. The forces include the electrostatic forces between the nucleus and the electrons and between the electrons themselves, as well as weaker magnetic forces arising from the spin and orbital motions of the electrons. Despite these difficulties, approximation methods introduced by the English physicist Douglas R. Hartree and others in the 1920s have achieved considerable success. Such schemes start by assuming that each electron moves independently in an average electric field due to the nucleus and the other electrons—*i.e.,* correlations between the positions of the electrons are ignored. Each electron has its own wave function, called an orbital. The overall wave function for all the electrons in the atom satisfies the exclusion principle. Corrections to the calculated energies are then made, which depend on the strengths of the electron–electron correlations and the magnetic forces.

Hartree method

TIME-DEPENDENT SCHRÖDINGER EQUATION

At the same time that Schrödinger proposed his time-independent equation to describe the stationary states, he also proposed a time-dependent equation to describe how a system changes from one state to another. By replacing the energy E in equation (99) with a time-derivative operator, he generalized his wave equation to determine the time variation of the wave function as well as its spatial variation. The time-dependent Schrödinger equation reads

$$-\frac{\hbar^2}{2m_e}\left(\frac{\partial^2\Psi}{\partial x^2}+\frac{\partial^2\Psi}{\partial y^2}+\frac{\partial^2\Psi}{\partial z^2}\right)+V(x,y,z)\Psi=i\hbar\frac{\partial\Psi}{\partial t}. \quad (100)$$

The quantity i is the square root of -1. The function Ψ varies with time t as well as with position x,y,z. For a system with constant energy, E, Ψ has the form

$$\Psi(x,y,z,t)=\Psi(x,y,z)\exp\left(-\frac{iEt}{\hbar}\right), \quad (101)$$

where exp stands for the exponential function, and the time-dependent Schrödinger equation reduces to the time-independent form.

The probability of a transition between one atomic stationary state and some other state can be calculated with the aid of the time-dependent equation. For example, an atom may change spontaneously from one state to another state with less energy, emitting the difference in energy as a photon with a frequency given by the Bohr relation. If electromagnetic radiation is applied to a set of atoms and if the frequency of the radiation matches the energy difference between two stationary states, transitions can be stimulated. In a stimulated transition, the energy of the atom may increase; *i.e.,* the atom may absorb a photon from the electromagnetic radiation, or the energy of the atom may decrease, with the emission of a photon, which adds to the energy of the electromagnetic radiation. Such stimulated emission processes form the basic mechanism for the operation of lasers. The probability of a transition between one state and another depends on the values of the l, m, m_s quantum numbers of the initial and final states. For most values, the transition probability is effectively zero. However, for certain changes in the quantum numbers, summarized as selection rules, there is a finite probability. For example, according to one important selection rule, the l value changes by unity. The selection rules for radiation relate to the angular momentum properties of the stationary states. The absorbed or emitted photon has its own angular momentum, and the selection rules reflect the conservation of angular momentum between the atoms and the radiation.

Transitions between atomic states

TUNNELING

The phenomenon of tunneling, which has no counterpart in classical physics, is an important consequence of quantum mechanics. Consider a particle with energy E in the inner region of a one-dimensional potential well $V(x)$, as shown in Figure 72. (A potential well is a potential that has a lower value in a certain region of space than in the neighbouring regions.) In classical mechanics, if $E < V_0$ (the maximum height of the potential barrier), the particle remains in the well forever; if $E > V_0$, the particle escapes.

No counterpart in classical physics

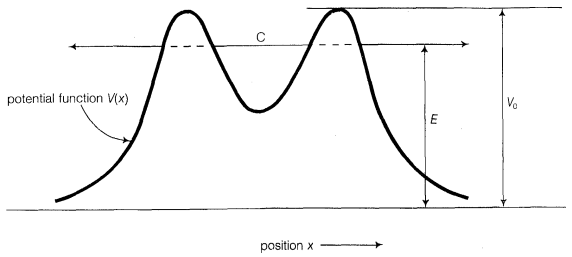


Figure 72: *The phenomenon of tunneling.*
Classically, a particle is bound in the central region C if its energy E is less than V_0, but in quantum theory the particle may tunnel through the potential barrier and escape.

In quantum mechanics, the situation is not so simple. The particle can escape even if its energy E is below the height of the barrier V_0, although the probability of escape is small unless E is close to V_0. In that case, the particle may tunnel through the potential barrier and emerge with the same energy E.

The phenomenon has many important applications. For example, it describes a type of radioactive decay in which a nucleus emits an alpha particle (a helium nucleus). According to the quantum explanation given independently by George Gamow and by Ronald W. Gurney and Edward Condon in 1928, the alpha particle is confined before the decay by a potential of the shape shown in Figure 72. For a given nuclear species, it is possible to measure the energy E of the emitted alpha particle and the average lifetime τ of the nucleus before decay. The lifetime of the nucleus is a measure of the probability of tunneling through the barrier—the shorter the lifetime, the higher the probability. With plausible assumptions about the general form of the potential function, it is possible to calculate a relationship between τ and E that is applicable to all alpha emitters. This theory, which is borne out by experiment, shows that the probability of tunneling, and hence the value of τ, is extremely sensitive to the value of E. For all known alpha-particle emitters, the value of E varies from about 2 to 8 megaelectron volts, or MeV (1 MeV = 10^6 electron volts). Thus, the value of E varies only by a factor of 4, whereas the range of τ is from about 10^{11} years down to about 10^{-6} second, a factor of 10^{24}. It would be difficult to account for this sensitivity of τ to the value of E by any theory other than quantum mechanical tunneling.

AXIOMATIC APPROACH

Axiomatic approach based on observables and states

Although the two Schrödinger equations form an important part of quantum mechanics, it is possible to present the subject in a more general way. Dirac gave an elegant exposition of an axiomatic approach based on observables and states in a classic textbook entitled *The Principles of Quantum Mechanics.* (The book, published in 1930, is still in print.) An observable is anything that can be measured—energy, position, a component of angular momentum, and so forth. Every observable has a set of states, each state being represented by an algebraic function. With each state is associated a number that gives the result of a measurement of the observable. Consider an observable with N states, denoted by $\varphi_1, \varphi_2, \ldots \varphi_N$, and corresponding measurement values $a_1, a_2, \ldots a_N$. A physical system—*e.g.,* an atom in a particular state—is represented by a wave function Ψ, which can be expressed as a linear combination, or mixture, of the states of the observable. Thus, the Ψ may be written as

$$\Psi = c_1\varphi_1 + c_2\varphi_2 + \ldots + c_N\varphi_N. \qquad (102)$$

For a given Ψ, the quantities c_1, c_2, etc., are a set of numbers that can be calculated. In general, the numbers are complex, but, in the present discussion, they are assumed to be real numbers.

The theory postulates, first, that the result of a measurement must be an a-value—*i.e.,* a_1, a_2, or a_3, etc. No other value is possible. Second, before the measurement is made, the probability of obtaining the value a_1 is c_1^2, and that of obtaining the value a_2 is c_2^2, and so on. If the value obtained is, say, a_5, the theory asserts that after the measurement the state of the system is no longer the original Ψ but has changed to φ_5, the state corresponding to a_5.

A number of consequences follow from these assertions. First, the result of a measurement cannot be predicted with certainty. Only the probability of a particular result can be predicted, even though the initial state (represented by the function Ψ) is known exactly. Second, identical measurements made on a large number of identical systems, all in the identical state Ψ, will produce different values for the measurements. This is, of course, quite contrary to classical physics and common sense, which say that the same measurement on the same object in the same state must produce the same result. Moreover, according to the theory, not only does the act of measurement change the state of the system, but it does so in an indeterminate way. Sometimes it changes the state to φ_1, sometimes to φ_2, and so forth.

There is an important exception to the above statements. Suppose that, before the measurement is made, the state Ψ happens to be one of φs, say $\Psi = \varphi_3$. Then $c_3 = 1$ and all the other c's are zero. This means that, before the measurement is made, the probability of obtaining the value a_3 is unity and the probability of obtaining any other value of a is zero. In other words, in this particular case, the result of the measurement can be predicted with certainty. Moreover, after the measurement is made, the state will be φ_3, the same as it was before. Thus, in this particular case, measurement does not disturb the system. Whatever the initial state of the system, two measurements made in rapid succession (so that the change in the wave function given by the time-dependent Schrödinger equation is negligible) produce the same result.

The value of one observable can be determined by a single measurement. The value of two observables for a given system may be known at the same time, provided that the two observables have the same set of state functions $\varphi_1, \varphi_2, \ldots \varphi_N$. In this case, measuring the first observable results in a state function that is one of the φs. Because this is also a state function of the second observable, the result of measuring the latter can be predicted with certainty. Thus the values of both observables are known. (Although the φs are the same for the two observables, the two sets of a values are, in general, different.) The two observables can be measured repeatedly in any sequence. After the first measurement, none of the measurements disturbs the system, and a unique pair of values for the two observables is obtained.

INCOMPATIBLE OBSERVABLES

Observables with different sets of state functions

The measurement of two observables with different sets of state functions is a quite different situation. Measurement of one observable gives a certain result. The state function after the measurement is, as always, one of the states of that observable; however, it is not a state function for the second observable. Measuring the second observable disturbs the system, and the state of the system is no longer one of the states of the first observable. In general, measuring the first observable again does not produce the same result as the first time. To sum up, both quantities cannot be known at the same time, and the two observables are said to be incompatible.

A specific example of this behaviour is the measurement of the component of angular momentum along two mutually perpendicular directions. The Stern–Gerlach experiment mentioned above involved measuring the angular momentum of a silver atom in the ground state. In reconstructing this experiment, a beam of silver atoms is passed between the poles of a magnet. The poles are shaped so that the magnetic field varies greatly in strength over a very small distance (Figure 73). The apparatus determines the m_s quantum number, which can be $+1/2$ or $-1/2$. No other values are obtained. Thus in this case the observable has only two states—*i.e.,* $N = 2$. The inhomogeneous magnetic field produces a force on the silver atoms in a direction that depends on the spin state of the atoms. The result is shown schematically in Figure 74. A beam of silver atoms is passed through magnet A. The atoms in the state with $m_s = +1/2$ are deflected upward and emerge as beam 1, while those with $m_s = -1/2$ are deflected downward and emerge as beam 2. If the direction

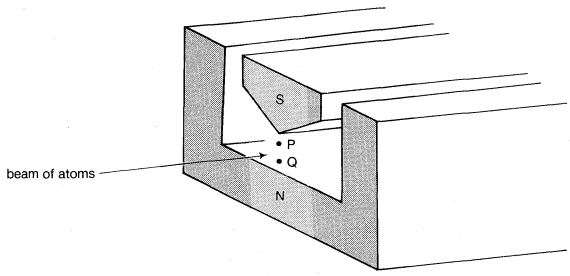


Figure 73: *Magnet in Stern–Gerlach experiment.*
N and S are the north and south poles of a magnet. The knife-edge of S results in a much stronger magnetic field at the point P than at Q.

of the magnetic field is the x-axis, the apparatus measures S_x, which is the x-component of spin angular momentum. The atoms in beam 1 have $S_x = +\frac{1}{2}\hbar$, while those in beam 2 have $S_x = -\frac{1}{2}\hbar$. In a classical picture, these two states represent atoms spinning about the direction of the x-axis with opposite senses of rotation.

The y-component of spin angular momentum S_y also can have only the values $+\frac{1}{2}\hbar$ and $-\frac{1}{2}\hbar$; however, the two states of S_y are not the same as for S_x. In fact, each of the states of S_x is an equal mixture of the states for S_y, and conversely. Again, the two S_y states may be pictured as representing atoms with opposite senses of rotation about the y-axis. These classical pictures of quantum states are helpful, but only up to a certain point. For example, quantum theory says that each of the states corresponding to spin about the x-axis is a superposition of the two states with spin about the y-axis. There is no way to visualize this; it has absolutely no classical counterpart. One simply has to accept the result as a consequence of the axioms of the theory. Suppose that, as in Figure 74, the atoms in beam 1 are passed into a second magnet B, which has a magnetic field along the y-axis perpendicular to x. The atoms emerge from B and go in equal numbers through its two output channels. Classical theory says that the two magnets together have measured both the x- and y-components of spin angular momentum and that the atoms in beam 3 have $S_x = +\frac{1}{2}\hbar$, $S_y = +\frac{1}{2}\hbar$, while those in beam 4 have $S_x = +\frac{1}{2}\hbar$, $S_y = -\frac{1}{2}\hbar$. However, classical theory is wrong, because if beam 3 is put through still another magnet C, with its magnetic field along x, the atoms divide equally into beams 5 and 6 instead of emerging as a single beam 5 (as they would if they had $S_x = +\frac{1}{2}\hbar$). Thus, the correct statement is that the beam entering B has $S_x = +\frac{1}{2}\hbar$ and is composed of an equal mixture of the states $S_y = +\frac{1}{2}\hbar$ and $S_y = -\frac{1}{2}\hbar$—*i.e.*, the x-component of angular momentum is known but the y-component is not. Correspondingly, beam 3 leaving B has $S_y = +\frac{1}{2}\hbar$ and is an equal mixture of the states $S_x = +\frac{1}{2}\hbar$ and $S_x = -\frac{1}{2}\hbar$; the y-component of angular momentum is known but the x-component is not. The information about S_x is lost because of the disturbance caused by magnet B in the measurement of S_y.

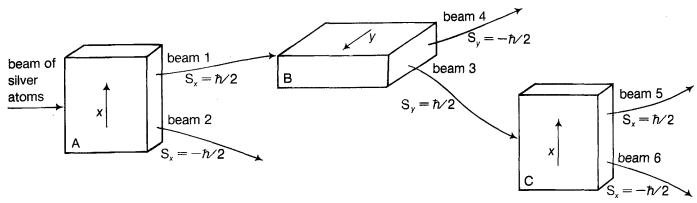


Figure 74: Measurements of the x and y components of angular momentum for silver atoms, S, in the ground state. A, B, and C are magnets with inhomogeneous magnetic fields. The arrows show the average direction of each magnetic field.

HEISENBERG UNCERTAINTY PRINCIPLE

The observables discussed so far have had discrete sets of experimental values. For example, the values of the energy of a bound system are always discrete, and angular momentum components have values that take the form $m\hbar$, where m is either an integer or a half-integer, positive or negative. On the other hand, the position of a particle or the linear momentum of a free particle can take continuous values in both quantum and classical theory. The mathematics of observables with a continuous spectrum of measured values is somewhat more complicated than for the discrete case, but presents no problems of principle. An observable with a continuous spectrum of measured values has an infinite number of state functions. The state function Ψ of the system is still regarded as a combination of the state functions of the observable, but the sum in equation (102) must be replaced by an integral.

Measurements can be made of position x of a particle and the x-component of its linear momentum, denoted by p_x. These two observables are incompatible because they have different state functions. The phenomenon of diffraction noted above illustrates the impossibility of measuring position and momentum simultaneously and precisely. If a parallel monochromatic light beam passes through a slit (Figure 75A), its intensity varies with direction, as shown in Figure 75B. The light has zero intensity in certain directions. Wave theory shows that the first zero occurs at an angle θ_0, given by $\sin\theta_0 = \lambda/b$, where λ is the wavelength of the light and b is the width of the slit. If the width of the slit is reduced, θ_0 increases—*i.e.*, the diffracted light is more spread out. Thus, θ_0 measures the spread of the beam.

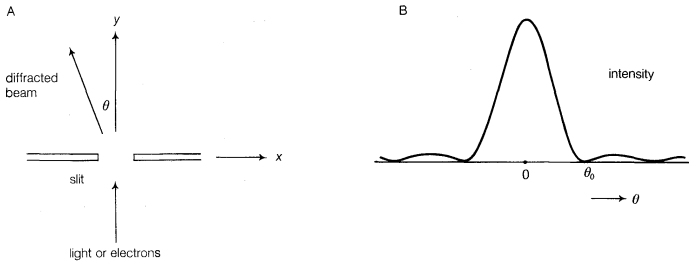


Figure 75: (A) Parallel monochromatic light incident normally on a slit, (B) variation in the intensity of the light with direction after it has passed through the slit. If the experiment is repeated with electrons instead of light, the same diagram would represent the variation in the intensity (*i.e.*, relative number) of the electrons.

The experiment can be repeated with a stream of electrons instead of a beam of light. According to de Broglie, electrons have wavelike properties; therefore, the beam of electrons emerging from the slit should widen and spread out like a beam of light waves. This has been observed in experiments. If the electrons have velocity u in the forward direction (*i.e.*, the y-direction in Figure 75A), their (linear) momentum is $p = m_e u$. Consider p_x, the component of momentum in the x-direction. After the electrons have passed through the aperture, the spread in their directions results in an uncertainty in p_x by an amount

Position–momentum uncertainty relation

$$\Delta p_x \approx p \sin\theta_0 = p\frac{\lambda}{b}, \tag{103}$$

where λ is the wavelength of the electrons and, according to the de Broglie formula, equals h/p. Thus, $\Delta p_x \approx h/b$. Exactly where an electron passed through the slit is unknown; it is only certain that an electron went through somewhere. Therefore, immediately after an electron goes through, the uncertainty in its x-position is $\Delta x \approx b/2$. Thus the product of the uncertainties is of the order of $\hbar$. More exact analysis shows that the product has a lower limit, given by

$$\Delta x\, \Delta p_x \geq \frac{\hbar}{2}. \tag{104}$$

This is the well-known Heisenberg uncertainty principle for position and momentum. It states that there is a limit to the precision with which the position and the momentum of an object can be measured at the same time. Depending on the experimental conditions, either quantity can be measured as precisely as desired (at least in principle), but the more precisely one of the quantities is measured, the less precisely the other is known.

The uncertainty principle is significant only on the atomic scale because of the small value of h in everyday units.

If the position of a macroscopic object with a mass of, say one gram, is measured with a precision of 10^{-6} metre, the uncertainty principle states that its velocity cannot be measured to better than about 10^{-25} metre per second. Such a limitation is hardly worrisome. However, if an electron is located in an atom about 10^{-10} metre across, the principle gives a minimum uncertainty in the velocity of about 10^{6} metre per second.

The above reasoning leading to the uncertainty principle is based on the wave–particle duality of the electron. When Heisenberg first propounded the principle in 1927 his reasoning was based, however, on the wave–particle duality of the photon. He considered the process of measuring the position of an electron by observing it in a microscope. Diffraction effects due to the wave nature of light result in a blurring of the image; the resulting uncertainty in the position of the electron is approximately equal to the wavelength of the light. To reduce this uncertainty, it is necessary to use light of shorter wavelength—*e.g.*, gamma rays. However, in producing an image of the electron, the gamma-ray photon bounces off the electron, giving the Compton effect (see above). As a result of the collision, the electron recoils in a statistically random way. The resulting uncertainty in the momentum of the electron is proportional to the momentum of the photon, which is inversely proportional to the wavelength of the photon. So it is again the case that increased precision in knowledge of the position of the electron is gained only at the expense of decreased precision in knowledge of its momentum. A detailed calculation of the process yields the same result as before (equation [104]). Heisenberg's reasoning brings out clearly the fact that the smaller the particle being observed, the more significant is the uncertainty principle. When a large body is observed, photons still bounce off it and change its momentum, but, considered as a fraction of the initial momentum of the body, the change is insignificant.

Energy–time uncertainty principle

The Schrödinger and Dirac theories give a precise value for the energy of each stationary state, but in reality the states do not have a precise energy. The only exception is in the ground state, which is the state with the lowest energy. Instead, the energies of the states are spread over a small range. The spread arises from the fact that, because the electron can make a transition to another state, the initial state has a finite lifetime. The transition is a random process, and so different atoms in the same state have different lifetimes. If the mean lifetime is denoted as τ, the theory shows that the energy of the initial state has a spread of energy ΔE, given by

$$\tau \Delta E \approx \hbar. \qquad (105)$$

This energy spread is manifested in a spread in the frequencies of emitted radiation. Therefore, the spectral lines are not infinitely sharp. (Some experimental factors can also broaden a line, but their effects can be reduced; however, the present effect, known as natural broadening, is fundamental and cannot be reduced.) Equation (105) is another type of Heisenberg uncertainty relation; generally, if a measurement with duration τ is made of the energy in a system, the measurement disturbs the system, causing the energy to be uncertain by an amount ΔE, the magnitude of which is given by the above equation.

QUANTUM ELECTRODYNAMICS

The application of quantum theory to the interaction between electrons and radiation requires a quantum treatment of Maxwell's field equations, which are the foundations of electromagnetism, and the relativistic theory of the electron formulated by Dirac (see above *Electron spin and antiparticles*). The resulting quantum field theory is known as quantum electrodynamics, or QED.

QED accounts for the behaviour and interactions of electrons, positrons, and photons. It deals with processes involving the creation of material particles from electromagnetic energy and with the converse processes in which a material particle and its antiparticle annihilate each other and produce energy. Initially the theory was beset with formidable mathematical difficulties because the calculated values of quantities such as the charge and mass of the electron proved to be infinite. However, an ingenious set of techniques developed (in the late 1940s) by Hans Bethe, Julian S. Schwinger, Tomonaga Shin'ichirō, Richard P. Feynman, and others dealt systematically with the infinities to obtain finite values of the physical quantities. Their method is known as renormalization. The theory has provided some remarkably accurate predictions.

Renormalization

According to the Dirac theory, two particular states in hydrogen with different quantum numbers have the same energy. QED, however, predicts a small difference in their energies; the difference may be determined by measuring the frequency of the electromagnetic radiation that produces transitions between the two states. This effect was first measured by Willis E. Lamb, Jr., and Robert Retherford in 1947. Its physical origin lies in the interaction of the electron with the random fluctuations in the surrounding electromagnetic field. These fluctuations, which exist even in the absence of an applied field, are a quantum phenomenon. The accuracy of experiment and theory in this area may be gauged by two recent values for the separation of the two states, expressed in terms of the frequency of the radiation that produces the transitions:

experiment (1982)	1,057,858 ± 2 kilohertz
theory (1975)	1,057,864 ± 14 kilohertz.

Magnetic dipole moment of the electron

An even more spectacular example of the success of QED is provided by the value for μ_e, the magnetic dipole moment of the free electron. Because the electron is spinning and has electric charge, it behaves like a tiny magnet, the strength of which is expressed by the value of μ_e. According to the Dirac theory, μ_e is exactly equal to $\mu_B = e\hbar/2m_e$, a quantity known as the Bohr magneton; however, QED predicts that $\mu_e = (1 + a)\mu_B$, where a is a small number, approximately $^1/_{860}$. Again, the physical origin of the QED correction is the interaction of the electron with random oscillations in the surrounding electromagnetic field. The best experimental determination of μ_e involves measuring, not the quantity itself, but the small correction term $\mu_e - \mu_B$. This greatly enhances the sensitivity of the experiment. The most recent results for the value of a are

experiment (1984)	$(115{,}965{,}219 \pm 1) \times 10^{-11}$
theory (1986)	$(115{,}965{,}227 \pm 10) \times 10^{-11}$.

Since a itself represents a small correction term, the magnetic dipole moment of the electron is measured with an accuracy of about one part in 10^{11}. One of the most precisely determined quantities in physics, the magnetic dipole moment of the electron can be calculated correctly from quantum theory to within about one part in 10^{10}.

The interpretation of quantum mechanics

Although quantum mechanics has been applied to problems in physics with great success, some of its ideas seem very strange. A few of their implications are considered here.

THE ELECTRON: WAVE OR PARTICLE?

Young's aforementioned experiment in which a parallel beam of monochromatic light is passed through a pair of narrow parallel slits (Figure 76A) has an electron counterpart. In Young's original experiment, the intensity of the light varies with direction after passing through the slits

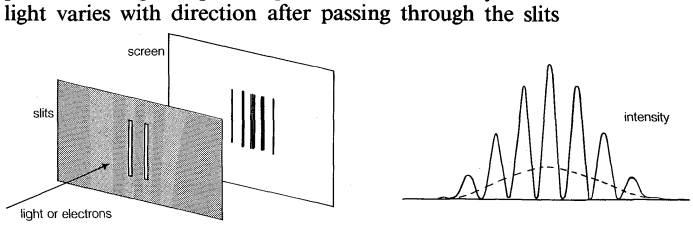


Figure 76: (A) Monochromatic light incident on a pair of slits gives interference fringes (alternate light and dark bands) on a screen, (B) variation in the intensity of the light at the screen when both slits are open. With a single slit, there is no interference pattern; the intensity variation is shown by the broken line. As with Figure 75B, the same diagram would give the variation in the intensity of electrons in the corresponding electron experiment.

Two-slit experiment

(Figure 76B). The intensity oscillates due to interference between the light waves emerging from the two slits, the rate of oscillation depending on the wavelength of the light and the separation of the slits. The oscillation creates a fringe pattern of alternating light and dark bands that is modulated by the diffraction pattern from each slit. If one of the slits is covered, the interference fringes disappear, and only the diffraction pattern (shown as a broken line in Figure 76B) is observed.

Young's experiment can be repeated with electrons all with the same momentum. The screen in the optical experiment is replaced by a closely spaced grid of electron detectors. There are many devices for detecting electrons; the most common are scintillators. When an electron passes through a scintillating material, such as sodium iodide, the material produces a light flash which gives a voltage pulse that can be amplified and recorded. The pattern of electrons recorded by each detector is the same as that predicted for waves with wavelengths given by the de Broglie formula. Thus, the experiment provides conclusive evidence for the wave behaviour of electrons.

If the experiment is repeated with a very weak source of electrons so that only one electron passes through the slits, a single detector registers the arrival of an electron. This is a well-localized event characteristic of a particle. Each time the experiment is repeated, one electron passes through the slits and is detected. A graph plotted with detector position along one axis and the number of electrons along the other looks exactly like the oscillating interference pattern in Figure 76B. Thus, the intensity function in the figure is proportional to the probability of the electron moving in a particular direction after it has passed through the slits. Apart from its units, the function is identical to Ψ^2, where Ψ is the solution of the time-independent Schrödinger equation for this particular experiment.

If one of the slits is covered, the fringe pattern disappears and is replaced by the diffraction pattern for a single slit. Thus, both slits are needed to produce the fringe pattern. However, if the electron is a particle, it seems reasonable to suppose that it passed through only one of the slits. The apparatus can be modified to ascertain which slit by placing a thin wire loop around each slit. When an electron passes through a loop, it generates a small electric signal, showing which slit it passed through. However, the interference fringe pattern then disappears, and the single-slit diffraction pattern returns. Since both slits are needed for the interference pattern to appear and since it is impossible to know which slit the electron passed through without destroying that pattern, one is forced to the conclusion that the electron goes through both slits at the same time.

Experimental proof of wave–particle duality

In summary, the experiment shows both the wave and particle properties of the electron. The wave property predicts the probability of direction of travel before the electron is detected; on the other hand, the fact that the electron is detected in a particular place shows that it has particle properties. Therefore, the answer to the question whether the electron is a wave or a particle is that it is neither. It is an object exhibiting either wave or particle properties, depending on the type of measurement that is made on it. In other words, one cannot talk about the intrinsic properties of an electron; instead, one must consider the properties of the electron and measuring apparatus together.

HIDDEN VARIABLES

The indeterminacy of quantum mechanics

A fundamental concept in quantum mechanics is that of randomness, or indeterminacy. In general, the theory predicts only the probability of a certain result. Consider the case of radioactivity. Imagine a box of atoms with identical nuclei that can undergo decay with the emission of an alpha particle. In a given time interval, a certain fraction will decay. The theory may tell precisely what that fraction will be, but it cannot predict which particular nuclei will decay. The theory asserts that, at the beginning of the time interval, all the nuclei are in an identical state and that the decay is a completely random process. Even in classical physics, many processes appear random. For example, one says that when a roulette wheel is spun, the ball will drop at random into one of the numbered compartments in the wheel. Based on this belief, the casino owner and the players give and accept identical odds against each number for each throw. However, the fact is that the winning number could be predicted if one noted the exact location of the wheel when the croupier released the ball, the initial speed of the wheel, and various other physical parameters. It is only ignorance of the initial conditions and the difficulty of doing the calculations that makes the outcome appear to be random. In quantum mechanics, on the other hand, the randomness is asserted to be absolutely fundamental. The theory says that, though one nucleus decayed and the other did not, they were previously in the identical state.

Many eminent physicists, including Einstein, have not accepted this indeterminacy. They have rejected the notion that the nuclei were initially in the identical state. Instead, they postulated that there must be some other property—presently unknown, but existing nonetheless—that is different for the two nuclei. This type of unknown property is termed a hidden variable; if it existed, it would restore determinacy to physics. If the initial values of the hidden variables were known, it would be possible to predict which nuclei would decay. Such a theory would, of course, also have to account for the wealth of experimental data which conventional quantum mechanics explains from a few simple assumptions. Attempts have been made by de Broglie, David Bohm, and others to construct theories based on hidden variables, but the theories are very complicated and contrived. For example, the electron would definitely have to go through only one slit in the two-slit experiment. To explain that interference occurs only when the other slit is open, it is necessary to postulate a special force on the electron which exists only when that slit is open. Such artificial additions make hidden variable theories unattractive, and there is little support for them among physicists.

The orthodox view of quantum mechanics—and the one adopted in the present article—is known as the Copenhagen interpretation because its main protagonist, Niels Bohr, worked in the city. The Copenhagen view of understanding the physical world stresses the importance of basing theory on what can be observed and measured experimentally. It therefore rejects the idea of hidden variables as quantities that cannot be measured. The Copenhagen view is that the indeterminacy observed in nature is fundamental and does not reflect an inadequacy in present scientific knowledge. One should therefore accept the indeterminacy without trying to "explain" it, and see what consequences come from it.

Rejection of the notion of hidden variables

Attempts have been made to link the existence of free will with the indeterminacy of quantum mechanics, but it is difficult to see how this feature of the theory makes free will more plausible. On the contrary, free will presumably implies rational thought and decision, whereas the essence of the indeterminism in quantum mechanics is that it is due to intrinsic randomness.

PARADOX OF EINSTEIN, PODOLSKY, AND ROSEN

In 1935 Einstein and two other physicists in the United States, Boris Podolsky and Nathan Rosen, analyzed a thought experiment to measure position and momentum in a pair of interacting systems. Employing conventional quantum mechanics, they obtained some startling results, which led them to conclude that the theory does not give a complete description of physical reality. Their results, which are so peculiar as to seem paradoxical, are based on impeccable reasoning, but, as will be seen, their conclusion that the theory is incomplete does not necessarily follow. Bohm simplified their experiment while retaining the central point of their reasoning; this discussion follows his account.

The proton, like the electron, has spin $1/2$; thus, no matter what direction is chosen for measuring the component of its spin angular momentum, the values are always $+\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$. (The present discussion relates only to spin angular momentum, and the word spin is omitted from now on.) It is possible to obtain a system consisting of a pair of protons in close proximity and with total angular momentum equal to zero. Thus, if the value of one of

the components of angular momentum for one of the protons is $+\frac{1}{2}\hbar$ along any selected direction, the value for the component in the same direction for the other particle must be $-\frac{1}{2}\hbar$. Suppose the two protons move in opposite directions until they are far apart. The total angular momentum of the system remains zero, and if the component of angular momentum along the same direction for each of the two particles is measured, the result is a pair of equal and opposite values. Therefore, after the quantity is measured for one of the protons, it can be predicted for the other proton; the second measurement is unnecessary. As previously noted, measuring a quantity changes the state of the system. Thus, if measuring S_x (the x-component of angular momentum) for proton 1 produces the value $+\frac{1}{2}\hbar$, the state of proton 1 after measurement corresponds to $S_x = +\frac{1}{2}\hbar$ and the state of proton 2 corresponds to $S_x = -\frac{1}{2}\hbar$. Any direction, however, can be chosen for measuring the component of angular momentum. Whichever direction is selected, the state of proton 1 after measurement corresponds to a definite component of angular momentum about that direction. Furthermore, since proton 2 must have the opposite value for the same component, it follows that the measurement on proton 1 results in a definite state for proton 2 relative to the chosen direction, notwithstanding the fact that the two particles may be millions of kilometres apart and are not interacting with each other at the time. Einstein and his two collaborators thought that this conclusion was so obviously false that the quantum mechanical theory on which it was based must be incomplete. They concluded that the correct theory would contain some hidden variable feature that would restore the determinism of classical physics.

Essential difference between quantum mechanics and classical physics

A comparison of how quantum theory and classical theory describe angular momentum for particle pairs illustrates the essential difference between the two outlooks. In both theories, if a system of two particles has a total angular momentum of zero, then the angular momenta of the two particles are equal and opposite. If the components of angular momentum are measured along the same direction, the two values are numerically equal, one positive and the other negative. Thus, if one component is measured, the other can be predicted. The crucial difference between the two theories is that, in classical physics, the system under investigation is assumed to have possessed the quantity being measured beforehand. The measurement does not disturb the system; it merely reveals the preexisting state. It may be noted that, if a particle were actually to possess components of angular momentum prior to measurement, such quantities would constitute hidden variables.

Does nature behave as quantum mechanics predicts? The answer comes from measuring the components of angular momenta for the two protons along different directions with an angle θ between them. A measurement on one proton can give only the result $+\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$. The experiment consists of measuring correlations between the plus and minus values for pairs of protons with a fixed value of θ, and then repeating the measurements for different values of θ, as in Figure 77. The interpretation of the results rests on an important theorem by the British physicist John Stewart Bell. Bell began by assuming the existence of some form of hidden variable with a value that would determine whether the measured angular momentum gives a plus or minus result. He further assumed locality—namely, that measurement on one proton (*i.e.*, the choice of the measurement direction) cannot affect the result of the measurement on the other proton. Both these assumptions agree with classical, commonsense ideas. He then showed quite generally that these two assumptions lead to a certain relationship, now known as Bell's inequality, for the correlation values mentioned above. Experiments have been conducted at several laboratories with photons instead of protons (the analysis is similar), and the results show fairly conclusively that Bell's inequality is violated. That is to say, the observed results agree with those of quantum mechanics and cannot be accounted for by a hidden variable (or deterministic) theory based on the concept of locality. One is forced to conclude that the two protons are a correlated pair and that a measurement on one affects the state of both, no matter how far apart they are. This may strike one as highly peculiar, but such is the way nature appears to be.

Bell's inequality

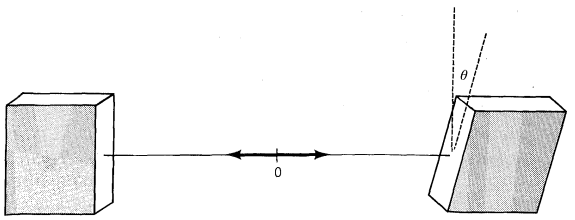


Figure 77: Experiment to determine the correlation in measured angular momentum values for a pair of protons with zero total angular momentum. The two protons are initially at the point 0 and move in opposite directions toward the two magnets.

It may be noted that the effect on the state of proton 2 following a measurement on proton 1 is believed to be instantaneous; the effect happens before a light signal initiated by the measuring event at proton 1 reaches proton 2. Alain Aspect and his coworkers in Paris demonstrated this result in 1982 with an ingenious experiment in which the correlation between the two angular momenta was measured, within a very short time interval, by a high-frequency switching device. The interval was less than the time taken for a light signal to travel from one particle to the other at the two measurement positions. Einstein's special theory of relativity states that no message can travel with a speed greater than that of light. Thus, there is no way that the information concerning the direction of the measurement on the first proton could reach the second proton before the measurement was made on it.

MEASUREMENT IN QUANTUM MECHANICS

The way quantum mechanics treats the process of measurement has caused considerable debate. Schrödinger's time-dependent wave equation (equation 100) is an exact recipe for determining the way the wave function varies with time for a given physical system in a given physical environment. According to the Schrödinger equation, the wave function varies in a strictly determinate way. On the other hand, in the axiomatic approach to quantum mechanics described above, a measurement changes the wave function abruptly and discontinuously. Before the measurement is made, the wave function Ψ is a mixture of the φs as indicated in equation (102). The measurement changes Ψ from a mixture of φs to a single φ. This change, brought about by the process of measurement, is termed the collapse or reduction of the wave function. The collapse is a discontinuous change in Ψ; it is also unpredictable, because starting with the same Ψ represented by the right-hand side of equation (102), the end result can be any one of the individual φs.

Collapse of the wave function Ψ

The Schrödinger equation, which gives a smooth and predictable variation of Ψ, applies between the measurements. The measurement process itself, however, cannot be described by the Schrödinger equation; it is somehow a thing apart. This appears unsatisfactory, inasmuch as a measurement is a physical process and ought to be the subject of the Schrödinger equation just like any other physical process.

The difficulty is related to the fact that quantum mechanics applies to microscopic systems containing one (or a few) electrons, protons, or photons. Measurements, however, are made with large-scale objects (*e.g.*, detectors, amplifiers, and meters) in the macroscopic world, which obeys the laws of classical physics. Thus, another way of formulating the question of what happens in a measurement is to ask how the microscopic quantum world relates and interacts with the macroscopic classical world. More narrowly, it can be asked how and at what point in the measurement process does the wave function collapse? So far, there are no satisfactory answers to these questions, although there are several schools of thought.

One approach stresses the role of a conscious observer in the measurement process and suggests that the wave function collapses when the observer reads the measuring instrument. Bringing the conscious mind into the mea-

surement problem seems to raise more questions than it answers, however.

Copenhagen interpretation

As discussed above, the Copenhagen interpretation of the measurement process is essentially pragmatic. It distinguishes between microscopic quantum systems and macroscopic measuring instruments. The initial object or event—*e.g.*, the passage of an electron, photon, or atom—triggers the classical measuring device into giving a reading; somewhere along the chain of events, the result of the measurement becomes fixed (*i.e.*, the wave function collapses). This does not answer the basic question, but says, in effect, not to worry about it. This is probably the view of most practicing physicists.

A third school of thought notes that an essential feature of the measuring process is irreversibility. This contrasts with the behaviour of the wave function when it varies according to the Schrödinger equation; in principle, any such variation in the wave function can be reversed by an appropriate experimental arrangement. However, once a classical instrument has given a reading, the process is not reversible. It may be that the key to the nature of the measurement process lies somewhere here. The Schrödinger equation is known to apply only to relatively simple systems. It is an enormous extrapolation to assume that the same equation applies to the large and complex system of a classical measuring device. It may be that the appropriate equation for such a system has features that produce irreversible effects (*e.g.*, wave-function collapse) which differ in kind from those for a simple system.

The many-worlds interpretation

One may also mention the so-called many-worlds interpretation, proposed by Hugh Everett III in 1957, which suggests that, when a measurement is made for a system in which the wave function is a mixture of states, the universe branches into a number of noninteracting universes. Each of the possible outcomes of the measurement occurs, but in a different universe. Thus, if $S_x = \hbar/2$ is the result of a Stern–Gerlach measurement on a silver atom (see above), there is another universe identical to ours in every way (including clones of people), except that the result of the measurement is $S_x = -\hbar/2$. Although this fanciful model solves some measurement problems, it has few adherents among physicists.

Because the various ways of looking at the measurement process lead to the same experimental consequences, trying to distinguish between them on scientific grounds may be fruitless. One or another may be preferred on the grounds of plausibility, elegance, or economy of hypotheses, but these are matters of individual taste. Whether one day a satisfactory quantum theory of measurement will emerge, distinguished from the others by its verifiable predictions, remains an open question.

Applications of quantum mechanics

As has been noted, quantum mechanics has been enormously successful in explaining microscopic phenomena in all branches of physics. The three phenomena described in this section are examples that demonstrate the quintessence of the theory.

DECAY OF THE K⁰ MESON

The K^0 meson, discovered in 1953, is produced in high-energy collisions between nuclei and other particles. It has zero electric charge, and its mass is about one-half the mass of the proton. It is unstable and, once formed, rapidly decays into either 2 or 3 pi-mesons. The average lifetime of the K^0 is about 10^{-10} second.

In spite of the fact that the K^0 meson is uncharged, quantum theory predicts the existence of an antiparticle with the same mass, decay products, and average lifetime; the antiparticle is denoted by $\bar{K}^0$. During the early 1950s, several physicists questioned the justification for postulating the existence of two particles with such similar properties. In 1955, however, Murray Gell-Mann and Abraham Pais made an interesting prediction about the decay of the K^0 meson. Their reasoning provides an excellent illustration of the quantum mechanical axiom that the wave function Ψ can be a superposition of states; in this case, there are two states, the K^0 and $\bar{K}^0$ mesons themselves.

A K^0 meson may be represented formally by writing the wave function as $\Psi = K^0$; similarly $\Psi = \bar{K}^0$ represents a $\bar{K}^0$. From the two states, K^0 and $\bar{K}^0$, the following two new states are constructed:

$$K_1 = \frac{(K^0 + \bar{K}^0)}{\sqrt{2}}, \tag{106a}$$

$$K_2 = \frac{(K^0 - \bar{K}^0)}{\sqrt{2}}. \tag{106b}$$

From these two equations it follows that

$$K^0 = \frac{(K_1 + K_2)}{\sqrt{2}}, \tag{106c}$$

$$\bar{K}^0 = \frac{(K_1 - K_2)}{\sqrt{2}}. \tag{106d}$$

The reason for defining the two states K_1 and K_2 is that, according to quantum theory, when the K^0 decays, it does not do so as an isolated particle; instead, it combines with its antiparticle to form the states K_1 and K_2. The state K_1 decays into two pi-mesons with a very short lifetime (about 10^{-10} second), while K_2 decays into three pi-mesons with a longer lifetime (about 10^{-7} second).

The physical consequences of these results may be demonstrated in the following experiment. K^0 particles are produced in a nuclear reaction at the point A (Figure 78). They move to the right in the figure and start to decay. At point A, the wave function is $\Psi = K^0$, which, from equation (106c), can be expressed as the sum of K_1 and K_2. As the particles move to the right, the K_1 state begins to decay rapidly. If the particles reach point B in about 10^{-8} second, nearly all the K_1 component has decayed, although hardly any of the K_2 component has done so. Thus, at point B, the beam has changed from one of pure K^0 to one of almost pure K_2, which equation (106b) shows is an equal mixture of K^0 and $\bar{K}^0$. In other words, $\bar{K}^0$ particles appear in the beam simply because K_1 and K_2 decay at different rates. At point B, the beam enters a block of absorbing material. Both the K^0 and $\bar{K}^0$ are absorbed by the nuclei in the block, but the $\bar{K}^0$ are absorbed more strongly. Thus, even though the beam is an equal mixture of K^0 and $\bar{K}^0$, when the beam enters the absorber, it is almost pure K^0 upon exit at point C. The beam begins and ends as K^0.

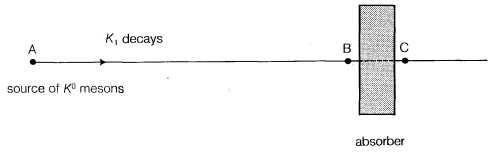


Figure 78: Decay of the K^0 meson.

Gell-Mann and Pais predicted all this, and experiments subsequently verified it. The experimental observations are that the decay products are primarily two pi-mesons with a short decay time near A, three pi-mesons with longer decay time near B, and two pi-mesons again near C. (This account exaggerates the changes in the K_1 and K_2 components between A and B and in the K^0 and $\bar{K}^0$ components between B and C; the argument, however, is unchanged.) The phenomenon of generating the $\bar{K}^0$ and regenerating the K_1 decay is purely quantum. It rests on the quantum axiom of the superposition of states and has no classical counterpart.

CESIUM CLOCK

Quantum oscillation frequency

The cesium clock is the most accurate type of clock yet developed. This device makes use of transitions between the spin states of the cesium nucleus, and produces a frequency which is so regular that it has been adopted for establishing the time standard.

Like electrons, many atomic nuclei have spin. The spin of these nuclei produces a set of small effects in the spectra, known as hyperfine structure. (The effects are small because, though the angular momentum of a spinning nucleus is of the same magnitude as that of an electron, its magnetic moment, which governs the energies of the atomic levels, is relatively small.) The nucleus of the

cesium atom has spin quantum number $^7/_2$. The total angular momentum of the lowest energy states of the cesium atom is obtained by combining the spin angular momentum of the nucleus with that of the single valence electron in the atom. (Only the valence electron contributes to the angular momentum because the angular momenta of all the other electrons total zero. Another simplifying feature is that the ground states have zero orbital momenta, so only spin angular momenta need to be considered.) When nuclear spin is taken into account, the total angular momentum of the atom is characterized by a quantum number, conventionally denoted by F, which for cesium is 4 or 3. These values come from the spin value $^7/_2$ for the nucleus and $^1/_2$ for the electron. If the nucleus and the electron are visualized as tiny spinning tops, the value $F = 4$ ($^7/_2 + {}^1/_2$) corresponds to the tops spinning in the same sense, and $F = 3$ ($^7/_2 - {}^1/_2$) corresponds to spins in opposite senses. The energy difference ΔE of the states with the two F values is a precise quantity. If electromagnetic radiation of frequency ν_0, where

$$h\nu_0 = \Delta E, \qquad (107)$$

is applied to a system of cesium atoms, transitions will occur between the two states. An apparatus that can detect the occurrence of transitions thus provides an extremely precise frequency standard. This is the principle of the cesium clock.

Principle of the cesium clock

The apparatus is shown schematically in Figure 79. A beam of cesium atoms emerges from an oven at a temperature of about 100° C. The atoms pass through an inhomogeneous magnet A, which deflects the atoms in state $F = 4$ downward and those in state $F = 3$ by an equal amount upward. The atoms pass through slit S and continue into a second inhomogeneous magnet B. Magnet B is arranged so that it deflects atoms with an unchanged state in the same direction that magnet A deflected them. The atoms follow the paths indicated by the broken lines in the figure and are lost to the beam. However, if an alternating electromagnetic field of frequency ν_0 is applied to the beam as it traverses the centre region C, transitions between states will occur. Some atoms in state $F = 4$ will change to $F = 3$, and vice versa. For such atoms, the deflections in magnet B are reversed. The atoms follow the whole lines in the diagram and strike a tungsten wire, which gives electric signals in proportion to the number of cesium atoms striking the wire. As the frequency ν of the alternating field is varied, the signal has a sharp maximum for $\nu = \nu_0$. The length of the apparatus from the oven to the tungsten detector is about one metre.

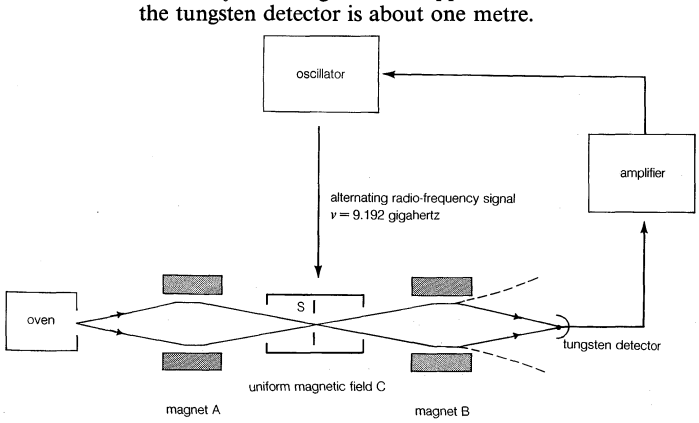


Figure 79: Cesium clock.

Each atomic state is characterized not only by the quantum number F but also by a second quantum number m_F. For $F = 4$, m_F can take integral values from 4 to −4. In the absence of a magnetic field, these states have the same energy. A magnetic field, however, causes a small change in energy proportional to the magnitude of the field and to the m_F value. Similarly, a magnetic field changes the energy for the $F = 3$ states according to the m_F value which, in this case, may vary from 3 to −3. The energy changes are indicated in Figure 80. In the cesium clock, a weak constant magnetic field is superposed on the alternating electromagnetic field in region C. The theory shows that the alternating field can bring about a transition only between pairs of states with m_F values that are the same or that differ by unity. However, as can be seen from the figure, the only transitions occurring at the frequency ν_0 are those between the two states with $m_F = 0$. The apparatus is so sensitive that it can discriminate easily between such transitions and all the others.

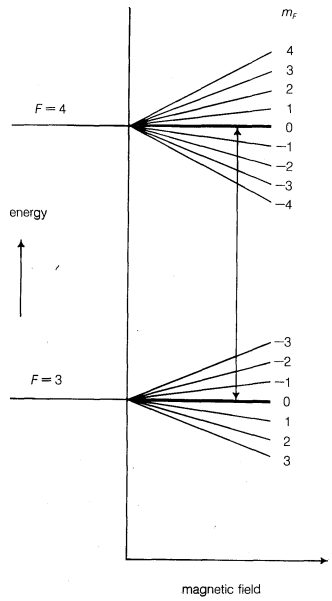


Figure 80: Variation of energy with magnetic-field strength for the $F = 4$ and $F = 3$ states in cesium-133.

If the frequency of the oscillator drifts slightly so that it does not quite equal ν_0, the detector output drops. The change in signal strength produces a signal to the oscillator to bring the frequency back to the correct value. This servo system keeps the oscillator frequency automatically locked to ν_0.

The cesium clock is exceedingly stable. The frequency of the oscillator remains constant to about one part in 10^{13}. For this reason, the device has been used to redefine the second. This base unit of time in the SI system is defined as equal to 9,192,631,770 cycles of the radiation corresponding to the transition between the levels $F = 4$, $m_F = 0$ and $F = 3$, $m_F = 0$ of the ground state of the cesium-133 atom. Prior to 1964, the second was defined in terms of the motion of the Earth. The latter, however, is not nearly as stable as the cesium clock. Specifically, the fractional variation of the Earth's rotation is a few hundred times larger than that of the frequency of the cesium clock.

Definition of the second

A QUANTUM VOLTAGE STANDARD

Quantum theory has been used to establish a voltage standard, and this standard has proven to be extraordinarily accurate and consistent from laboratory to laboratory.

If two layers of superconducting material are separated by a thin insulating barrier, a supercurrent (*i.e.*, a current of paired electrons) can pass from one superconductor to the other. This is another example of the tunneling process described earlier. Several effects based on this phenomenon were predicted in 1962 by the British physicist Brian D. Josephson. Demonstrated experimentally soon afterwards, they are now referred to as the Josephson effects.

If a DC (direct-current) voltage V is applied across the two superconductors, the energy of an electron pair changes by an amount of 2 eV as it crosses the junction. As a result, the supercurrent oscillates with frequency ν given by the Planck relationship—equation (91). Thus,

$$2\ \text{eV} = h\nu. \qquad (108)$$

This oscillatory behaviour of the supercurrent is known as the AC (alternating-current) Josephson effect. Measurement of V and ν permits a direct verification of the Planck relationship. Although the oscillating supercurrent

The AC Josephson effect

has been detected directly, it is extremely weak. A more sensitive method of investigating equation (108) is to study effects resulting from the interaction of microwave radiation with the supercurrent.

Several carefully conducted experiments have verified equation (108) to such a high degree of precision that it has been used to determine the value of $2e/h$. This value can in fact be determined more precisely by the AC Josephson effect than by any other method. The result is so reliable that laboratories now employ the AC Josephson effect to set a voltage standard. The numerical relationship between V and ν is

$$\frac{2e}{h} = \frac{\nu}{V} = 483{,}597.7 \times 10^9 \text{ hertz per volt.}$$

In this way, measuring a frequency, which can be done with great precision, gives the value of the voltage. Before the Josephson method was used, the voltage standard in metrological laboratories devoted to the maintenance of physical units was based on high-stability Weston cadmium cells. These cells, however, tend to drift, and so caused inconsistencies between standards in different laboratories. The Josephson method has provided a standard giving agreement to within a few parts in 10^8 for measurements made at different times and in different laboratories.

The experiments described in the preceding two sections are only two examples of high-precision measurements in physics. The values of the fundamental constants, such as c, h, e, and m_e, are determined from a wide variety of experiments based on quantum phenomena. The results are so consistent that the values of the constants are thought to be known in most cases to better than one part in 10^6. Physicists may not know what they are doing when they make a measurement, but they do it extremely well. (G.L.S.)

BIBLIOGRAPHY

Classical mechanics: ROBERT BRUCE LINDSAY and HENRY MARGENAU, *Foundations of Physics* (1936, reprinted 1981), contains a concise treatment of the foundations of classical mechanics. C. TRUESDELL, *Essays in the History of Mechanics* (1968), is a readable account of the development of some basic concepts, with copious references to original sources. PHYSICAL SCIENCE STUDY COMMITTEE, *Physics,* 6th ed. (1986); and SIDNEY ROSEN, ROBERT SIEGFRIED, and JOHN M. DENNISON, *Concepts in Physical Science* (1965), contain well-illustrated elementary treatments of the principles of mechanics suitable for a beginner and do not utilize calculus. J.L. MERIAM and L.G. KRAIGE, *Engineering Mechanics,* 2nd ed., 2 vol. (1986); and J.P. DEN HARTOG, *Mechanics* (1948, reprinted 1961), are both designed for undergraduate engineering students. CHARLES INGLIS, *Applied Mechanics for Engineers* (1951, reissued 1963), is noteworthy in both style and content but assumes some knowledge of the principles of mechanics. *Berkeley Physics Course,* vol. 1, *Mechanics,* by CHARLES KITTEL, WALTER D. KNIGHT, and MALVIN A. RUDEMAN, 2nd ed. (1973); IRVING H. SHAMES, *Engineering Mechanics,* 3rd ed., 2 vol. (1980); and A. DOUGLAS DAVIS, *Classical Mechanics* (1986), are also useful. J.H. JEANS, *An Elementary Treatise on the Theoretical Mechanics* (1907, reprinted 1967); and WILLIAM FOGG OSGOOD, *Mechanics* (1937, reprinted 1965), are more theoretical in their approach than the previous three books and cover more advanced topics.

JOHN L. SYNGE and BYRON A. GRIFFITH, *Principles of Mechanics,* 3rd ed. (1959); HERBERT GOLDSTEIN, *Classical Mechanics,* 2nd ed. (1980), a standard work; GEORGE W. HOUSNER and DONALD E. HUDSON, *Applied Mechanics,* 2nd ed., 2 vol. (1959–61); JAMES W. BROXON, *Mechanics* (1960); KEITH R. SYMON, *Mechanics,* 3rd ed. (1971); LAWRENCE E. GOODMAN and WILLIAM H. WARNER, *Dynamics* (1963); and JERRY B. MARION and STEPHEN T. THORNTON, *Classical Dynamics of Particles and Systems,* 3rd ed. (1988), are all of an intermediate level, utilizing algebraic rather than geometric methods in the treatment of vectors; in most cases an introductory chapter on vector analysis is included. FERDINAND P. BEER and E. RUSSELL JOHNSTON, JR., *Vector Mechanics for Engineers: Statics and Dynamics,* 5th ed. (1988), employs the full vector treatment. E.T. WHITTAKER, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies: With an Introduction to the Problem of Three Bodies,* 4th ed. (1937, reprinted 1970), is a well-known classic that presents an exhaustive and strictly analytic treatment of dynamics; it does not, however, use vector notation, and the scarcity of diagrams makes it difficult to follow. A.G. WEBSTER, *The Dynamics of Particles and of Rigid, Elastic, and Fluid Bodies,* 3rd ed. (1925), is a classic work. See also DAVID PARK, *Classical Dynamics and Its Quantum Analogues* (1979). Relativistic mechanics is discussed in R.D. SARD, *Relativistic Mechanics: Special Relativity and Classical Particle Dynamics* (1970). (A.Co./Ed.)

Celestial mechanics: Modern introductory treatments of celestial mechanics include: J.M.A. DANBY, *Fundamentals of Celestial Mechanics* (1962); I. BERNARD COHEN, *The Birth of a New Physics,* rev. and updated ed. (1985); S.W. MCCUSKEY, *Introduction to Celestial Mechanics* (1963); HARRY POLLARD, *Mathematical Introduction to Celestial Mechanics* (1966); and JEAN KOVALEVSKY, *Introduction to Celestial Mechanics* (1967; originally published in French, 1963). Applications to space dynamics and astrodynamics are offered by ROGER R. BATE, DONALD D. MUELLER, and JERRY E. WHITE, *Fundamentals of Astrodynamics* (1971); RICHARD H. BATTIN, *Astronautical Guidance* (1964); and ROBERT L.M. BAKER, JR. and MAUD W. MAKEMSON, *An Introduction to Astrodynamics,* 2nd ed. (1967).

Volumes of collected papers written by authorities in the field include VICTOR SZEBEHELY (ed.), *Celestial Mechanics and Astrodynamics* (1964), and *Dynamics of Planets and Satellites and Theories of Their Motion* (1978); and RAYNOR L. DUNCOMBE and VICTOR SZEBEHELY, *Methods in Astrodynamics and Celestial Mechanics* (1966). (V.G.S./Ed.)

Mechanics of elastic bodies: Apart from isolated chapters in many physics textbooks, most works on elasticity are for advanced study. Standard texts include STEPHEN P. TIMOSHENKO, *History of Strength of Materials: With a Brief Account of the History of Theory of Elasticity and Theory of Structures* (1953, reprinted 1983); A.H. COTTRELL, *The Mechanical Properties of Matter* (1964, reprinted 1981); W.D. BIGGS, *The Mechanical Behaviour of Engineering Materials* (1965); J.M. LESSELLS, *Strength and Resistance of Metals* (1954); ALFRED M. FREUDENTHAL, *Introduction to the Mechanics of Solids* (1966); I.S. SOKOLNIKOFF, *Mathematical Theory of Elasticity,* 2nd ed. (1956, reprinted 1983); and N.I. MUSKHELISHVILI, *Some Basic Problems of the Mathematical Theory of Elasticity: Fundamental Equations, Plane Theory of Elasticity, Torsion, and Bending,* 2nd ed. (1963, reissued 1977; originally published in Russian, 4th ed., 1954). See also J.H. WEINER, *Statistical Mechanics of Elasticity* (1983); and JERROLD E. MARSDEN and THOMAS J.R. HUGHS, *Mathematical Foundations of Elasticity* (1983). (D.S.D.)

Deformation and flow: J.C. JAEGER, *Elasticity, Fracture, and Flow, with Engineering and Geological Applications,* 3rd ed. (1969, reissued 1974), covers basic aspects. C. TRUESDELL, *The Elements of Continuum Mechanics* (1966), presents modern mathematical theory. JOHN D. FERRY, *Viscoelastic Properties of Polymers,* 3rd ed. (1980), discusses properties, experimental methods, and molecular theory. KARL F. HERZFELD and THEODORE A. LITOVITZ, *Absorption and Dispersion of Ultrasonic Waves* (1959), is an authoritative work, especially on gases and ordinary liquids. B.D. COLEMAN, H. MARKOVITZ, and W. NOLL, *Viscometric Flows of Non-Newtonian Fluids* (1966), covers theory and experiments on nonlinear steady flows. A. CEMAL ERINGEN, *Mechanics of Continua,* 2nd ed. (1980), outlines modern developments. BERNARD ROSEN (ed.), *Fracture Processes in Polymeric Solids: Phenomena and Theory* (1964), contains many authoritative articles. L.R.G. TRELOAR, *The Physics of Rubber Elasticity,* 3rd ed. (1975), discusses molecular theory and phenomenological finite elasticity theory and experiment. Other useful texts include A.S. LODGE, *Elastic Liquids: An Introductory Vector Treatment of Finite-Strain Polymer Rheology* (1964); and E.W. BILLINGTON and A. TATE, *The Physics of Deformation and Flow* (1981). (He.M.)

Tides and related phenomena: Discussions can be found in HORACE LAMB, *Hydrodynamics,* 6th ed. (1932, reissued 1945); GEORGE HOWARD DARWIN, *The Tides and Kindred Phenomena in the Solar System* (1898, reissued 1962); OTTO STRUVE, *Stellar Evolution: An Exploration from the Observatory* (1950); BARBARA M. MIDDLEHURST, "An Analysis of Lunar Events," *Reviews of Geophysics,* 5(2):173–189 (1967); and RUDOLF MEISSNER, GEORGE SUTTON, and FRED DUENNEBIER, "Mondbeben," *Umschau,* 71(4):111–115 (1971), on moonquakes, with an English summary. (W.Hn./B.M.M.)

Wave motion: Descriptive material intended for the reader with no technical background is available in WALTER R. FUCHS, *Physics for the Modern Mind* (1967; originally published in German, 1965), which covers many aspects of wave motion in a lucid manner; GEORGE GAMOW, "Vibrations and Waves," in *Matter, Earth, and Sky,* 2nd ed. (1965), an interesting introduction to these subjects by one of the most successful popularizers of science; ISAAC ASIMOV, "The Waves," ch. 8 in *Asimov's New Guide to Science* (1984), pp. 373–410, a very readable account of light, relativity, and quantum theory; S. TOLANSKY, *Curiosities of Light Rays and Light Waves* (1964), an excellent introduction; VERNON M. ALBERS, *The World of Sound* (1970), an exposition of the various aspects of sound; KONRAD B. KRAUSKOPF and ARTHUR BEISER, *Fundamentals of*

Physical Science, 6th ed. (1971), a standard presentation of waves; and JOHN R. PIERCE, *Waves and Messages* (1967), a treatment of electromagnetic waves and associated phenomena with very little mathematics. ALBERT EINSTEIN and LEOPOLD INFELD, *The Evolution of Physics: The Growth of Ideas from Early Concepts to Relativity and Quanta* (1938, reissued 1966), contains a discussion of waves, light, and ether. KEITH GORDON IRWIN, *The Romance of Physics* (1966), includes an elementary discussion of light.

Works on a more advanced level, written for the reader with some knowledge of physics and calculus, include JAMES LIGHTHILL, *Waves in Fluids* (1978); IAIN G. MAIN, *Vibrations and Waves in Physics,* 2nd ed. (1984); AKIRA HIROSE and KARL E. LONNGREN, *Introduction to Wave Phenomena* (1985); A.P. FRENCH, *Vibrations and Waves* (1971); and *Berkeley Physics Course,* vol. 3, *Waves,* by FRANK S. CRAWFORD, JR. (1968). RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, *The Feynman Lectures on Physics,* vol. 1 (1963, reprinted 1977), especially ch. 21, 23–38, and 47–51, deals with oscillations, resonance, optics, electromagnetic radiation, other aspects of radiation, and waves. For more on electromagnetic radiation, see CLAYTON R. PAUL and SYED A. NASAR, *Introduction to Electromagnetic Fields,* 2nd ed. (1987); VLADIMIR ROJANSKY, *Electromagnetic Fields and Waves* (1970, reprinted 1979); and F.H. READ, *Electromagnetic Radiation* (1980). (W.J.)

Ballistics: GREAT BRITAIN, MINISTRY OF SUPPLY, SCIENTIFIC ADVISORY COUNCIL, *Internal Ballistics* (1951), is a practical work accessible to the inexperienced ballistician. EDWARD J. MCSHANE, JOHN L. KELLEY and FRANKLIN V. RENO, *Exterior Ballistics* (1953), contains a historical appendix with a thorough account of the history of exterior ballistics with reference to other branches of ballistics. RAY KINSLOW (ed.), *High-Velocity Impact Phenomena* (1970), the first book on terminal ballistics, emphasizes experimental techniques and computer solutions. (R.I.Ro.)

Fluid mechanics: A classic text which enshrines all the results of 19th-century fluid dynamics is HORACE LAMB (*op. cit.*). This remains useful, but many later books, besides being more up-to-date, provide a better-balanced perspective of the subject and have better illustrations. N. CURLE and H.J. DAVIES, *Modern Fluid Dynamics,* vol. 1, *Incompressible Flow* (1968); and G.K. BATCHELOR, *Introduction to Fluid Dynamics* (1967, reissued 1973), can both be recommended to serious students who are not put off by mathematics. D.J. TRITTON, *Physical Fluid Dynamics,* 2nd ed. (1988), adopts a somewhat different approach and contains interesting material on turbulence and convective instabilities. Readers who are interested in the practical aspects of the subject and who want information concerning hydrostatics as well as fluid dynamics should consult one of the many good texts intended for engineers; among these, B.S. MASSEY, *Mechanics of Fluids,* 5th ed. (1983), is excellent. The development of the subject as a practical science is traced in HUNTER ROUSE and SIMON INCE, *History of Hydraulics* (1957, reprinted 1980). (T.E.F.)

Quantum mechanics: Scholarly but lively accounts of the early history of the subject are JAGDISH MEHRA and HELMUT RECHENBERG, *The Historical Development of Quantum Theory* (1982–), of which 5 vol. in 7 had appeared by 1989, covering the period from 1900 to 1926; and MAX JAMMER, *The Conceptual Development of Quantum Mechanics* (1966). The authors had the advantage, not usually accorded to historians, of being able to talk to their subjects.

There are a number of excellent texts on quantum mechanics at the undergraduate and graduate level. The following is a selection, beginning with the more elementary: A.P. FRENCH and EDWIN F. TAYLOR, *An Introduction to Quantum Physics* (1978); ALASTAIR I.M. RAE, *Quantum Mechanics,* 2nd ed. (1986); EUGEN MERZBACHER, *Quantum Mechanics,* 2nd ed. (1970); and ANTHONY SUDBERY, *Quantum Mechanics and the Particles of Nature: An Outline for Mathematicians* (1986), rather mathematical but including useful accounts and summaries of quantum metaphysics. RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, *The Feynman Lectures on Physics,* vol. 3, *Quantum Mechanics* (1965), is a personal and stimulating look at the subject. A good introduction to quantum electrodynamics is RICHARD P. FEYNMAN, *QED: The Strange Theory of Light and Matter* (1985).

J.C. POLKINGHORNE, *The Quantum World* (1984); and JOHN GRIBBIN, *In Search of Schrödinger's Cat: Quantum Physics and Reality* (1984), are both highly readable and instructive books written at a popular level. BERNARD D'ESPAGNAT, *Conceptual Foundations of Quantum Mechanics,* 2nd ed. (1976), is a technical account of the fundamental conceptual problems involved. The proceedings of a conference, *New Techniques and Ideas in Quantum Measurement Theory,* ed. by DANIEL M. GREENBERGER (1986), contain a wide-ranging set of papers that deal with both the experimental and theoretical aspects of the measurement problem.

Applications are presented by H. HAKEN and H.C. WOLF, *Atomic and Quantum Physics: An Introduction to the Fundamentals of Experiment and Theory,* 2nd enlarged ed. (1987; originally published in German, 2nd rev. and enlarged ed., 1983); EMILIO SEGRÈ, *Nuclei and Particles: An Introduction to Nuclear and Subnuclear Physics,* 2nd rev. and enlarged ed. (1977, reissued 1980); DONALD H. PERKINS, *Introduction to High Energy Physics,* 3rd ed. (1987); CHARLES KITTEL, *Introduction to Solid State Physics,* 6th ed. (1986); and RODNEY LOUDON, *The Quantum Theory of Light,* 2nd ed. (1983). B.W. PETLEY, *The Fundamental Physical Constants and the Frontier of Measurement* (1985), gives a good account of present knowledge of the fundamental constants. (G.L.S.)

Medicine

Until the revolutionary scientific discoveries of the 19th and 20th centuries, medical practice was generally restricted to folk medicine and proscriptive religious and cultural tenets and limited by uneven knowledge and religious beliefs. As the human body became better understood and the disease processes and agents that affect it became more rooted in fact, people began to turn their attention toward the various factors that were related to health.

The many important discoveries in human anatomy and physiology, infectious and other diseases, drugs, and therapeutic procedures that took place during the 19th and 20th centuries have had a direct bearing on the important developments that occurred in the field of public health. Philanthropists, social activists, and government officials began to take a critical look at the living conditions of the poor, the ill, and the working class and joined in a positive effort toward reform and education. The laws enacted during this period gradually brought about clean water, safer living conditions, lower morbidities and mortalities, preventive medicine, and a decrease in the incidence of infectious diseases.

Medical education for physicians and other health-care professionals has generally kept pace with advances in medicine, especially in the developed countries. Each country in the world maintains its own requirements for medical degrees and licenses, and medical boards and councils have been formed to set standards and oversee the quality of medical education. Boards of certification have created stringent requirements that physicians must meet before they can practice a specialty, and they put great stress on the need for practicing physicians to continue their education as well.

The disparity in medical services is especially obvious between developed and developing countries. This is often due to many problems, including finances, communication, education, and cultural and religious views. International organizations such as the World Health Organization and the United Nations International Children's Fund make substantial progress in this area by, among other things, considering the customs and culture in developing countries as an integral factor in the formation of the most efficient national health-care delivery system.

The concept of a general institution for the care of the sick can be traced to the Middle Ages, when monks set aside areas of their monasteries for this purpose. The modern general hospital fulfills a wide range of needs for the community both on an inpatient and an outpatient basis.

The complexity and dynamic nature of medical practice is reflected in the growing number of specialty areas of the average general hospital. Teaching also has become a concern of many general hospitals, and when designed as such, teaching hospitals provide services and knowledge often unavailable in other general hospitals.

Although religious and cultural laws as well as ethics have always been important in the history of medicine, it was not until the 1960s that the modern legal system became intimately involved in medical practice. Advances in therapeutics and diagnostics have created legal and moral issues in such complex areas as abortion, euthanasia, and patients' rights. Modern physicians especially have had to reexamine the ethical tenets by which their predecessors practiced medicine and to develop new standards by which they may be judged.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 424, 10/31, and 10/35. (Ed.)

This article is divided into the following sections:

The history of medicine and surgery 886
- Medicine and surgery before 1800 886
  - Primitive medicine and folklore
  - The ancient Middle East and Egypt
  - Traditional medicine and surgery in the Orient
  - The roots of Western medicine
  - Christian and Muslim reservoirs of learning
  - Medieval and Renaissance Europe
  - The Enlightenment
- The rise of scientific medicine in the 19th century 893
  - Physiology
  - Verification of the germ theory
  - Discoveries in clinical medicine and anesthesia
  - Advances at the end of the century
- Medicine in the 20th century 894
  - Infectious diseases and chemotherapy
  - Immunology
  - Endocrinology
  - Vitamins
  - Malignant disease
  - Tropical medicine
- Surgery in the 20th century 898
  - The opening phase
  - World War I
  - Between the world wars
  - World War II and after

The practice of modern medicine 902
- Health care and its delivery 902
  - Organization of health services
    - Levels of health care
    - Costs of health care
  - Administration of primary health care
  - Medical practice in developed countries
    - Britain
    - United States
    - Soviet Union
    - Japan
    - Other developed countries
  - Medical practice in developing countries
    - China
    - India
    - Other developing countries
  - Alternative or complementary medicine
  - Special practices and fields of medicine
    - Specialties in medicine
    - Teaching
    - Industrial medicine
    - Family health care
    - Geriatrics
    - Public health practice
    - Military practice
  - Clinical research
    - Historical notes
    - Clinical observation
    - Drug research
    - Surgery
  - Screening procedures
- Medical education 911
  - History of medical education
  - Modern patterns of medical education
    - Premedical education and admission to medical school
    - Undergraduate education
    - Postgraduate education
    - Continuing education
    - Medical school faculty
  - Requirements for practice
  - Economic aspects
  - Scientific and international aspects

Major medical institutions 914
- Hospitals 914
  - History of hospitals
  - The modern hospital
    - Comparison
    - Ownership, control, and financing
    - The general hospital
    - Specialized health- and medical-care facilities
    - Regional planning
- Public health services 917
  - History of public health
    - Beginnings in antiquity
    - The Middle Ages
    - The Renaissance
    - National developments in the 18th and 19th centuries
    - Developments from 1875
  - Modern organizational and administrative patterns
    - International organizations
    - Developed nations
    - Developing nations
  - Progress in public health
    - Developed nations
    - Developing nations
- Clinics 924
  - Hospital clinics
  - Public health clinics
  - Private clinics
  - Health centres
  - Polyclinics
  - Family planning clinics

Related fields 925
- Nursing 925
  - History of nursing
  - The practice of nursing
    - Kinds of nursing
    - Education
    - Licensing and registration
    - Organizations
  - Roles of international and multinational organizations
    - The International Red Cross
    - The World Health Organization
    - European Economic Community
- Dentistry 928
  - History of dentistry
  - The practice of dentistry
    - Licensure requirements
    - Types of practice
    - Dental specialties and subspecialties
    - Dental education
  - Ancillary dental fields
    - Dental hygienists
    - Dental nurses and dental auxiliaries
  - Organizations
- Pharmacy 932
  - History of pharmacy
  - The practice of pharmacy
    - Education
    - Licensing and regulation
    - Research
  - Organizations

Legal aspects of medicine 933
- Maintenance of professional standards 933
  - History
  - Relationship of law and ethics
- Law and medical practice 934
  - Government financing
  - License requirements for practice
    - The exclusive systems
    - The tolerant systems
    - The inclusive and integrated systems
  - Legal restrictions on practice
    - Determination of death
    - Termination of pregnancy
    - Sudden death
    - Public reporting
  - Legal redress
- Patient's rights 937

Bibliography 937

THE HISTORY OF MEDICINE AND SURGERY

Medicine and surgery before 1800

PRIMITIVE MEDICINE AND FOLKLORE

Unwritten history is not easy to interpret, and, although much may be learned from a study of the drawings, bony remains, and surgical tools of early man, it is difficult to reconstruct his mental attitude toward the problems of disease and death. It seems probable that humans, as soon as they had reached the stage of reasoning, discovered, by the process of trial and error, which plants might be used as foods, which of them were poisonous, and which of them had some medicinal value. Folk medicine or domestic medicine, consisting largely in the use of vegetable products, or herbs, originated in this fashion and still persists.

But that is not the whole story. Man did not at first regard death and disease as natural phenomena. Common maladies, such as colds or constipation, were accepted as part of existence and dealt with by means of such herbal remedies as were available. Serious and disabling diseases, however, were placed in a very different category. These were of supernatural origin. They might be the result of a spell cast upon the victim by some enemy, visitation by a malevolent demon, or the work of an offended god who had either projected some object—a dart, a stone, a worm—into the body of the victim or had abstracted something, usually the soul of the patient. The treatment then applied was to lure the errant soul back to its proper habitat within the body or to extract the evil intruder, be it dart or demon, by counterspells, incantations, potions, suction, or other means.

Belief in supernatural causes of illness

One curious method of providing the disease with means of escape from the body was by making a hole, 2.5 to five centimetres across, in the skull of the victim—the practice of trepanning, or trephining. Trepanned skulls of prehistoric date have been found in Britain, France, and other parts of Europe and in Peru. Many of them show evidence of healing and, presumably, of the patient's survival. The practice still exists among primitive people in parts of Algeria, in Melanesia, and perhaps elsewhere, though it is fast becoming extinct.

Magic and religion played a large part in the medicine of prehistoric or primitive man. Administration of the vegetable drug or remedy by mouth was accompanied by incantations, dancing, grimaces, and all the tricks of the magician. Thus, the first doctors, or "medicine men," were witch doctors or sorcerers. The use of charms and talismans, still prevalent in modern times, is of ancient origin.

Apart from the treatment of wounds and broken bones, the folklore of medicine is probably the most ancient aspect of the art of healing, for primitive physicians showed their wisdom by treating the whole person, soul as well as body. Treatments and medicines that produced no physical effects on the body could nevertheless make a patient feel better when both medicine man and patient believed in their efficacy. This so-called placebo effect is applicable even in modern clinical medicine.

THE ANCIENT MIDDLE EAST AND EGYPT

The establishment of the calendar and the invention of writing marked the dawn of recorded history. The clues to early knowledge are scanty, consisting of clay tablets bearing cuneiform signs and seals that were used by physicians of ancient Mesopotamia. In the Louvre there is preserved a stone pillar on which is inscribed the Code of Hammurabi, a Babylonian king of the 18th century BC. This code includes laws relating to medical practice, and the penalties for failure were severe. For example, "If the doctor, in opening an abscess, shall kill the patient, his hands shall be cut off"; if, however, the patient was a slave, the doctor was simply obliged to supply another slave.

The Greek historian Herodotus stated that every Babylonian was an amateur physician, since it was the custom to lay the sick in the street so that anyone passing by might offer advice. Divination, from the inspection of the liver of a sacrificed animal, was widely practiced to foretell the course of a disease. Little else is known regarding Babylonian medicine, and the name of not a single physician has survived.

When the medicine of ancient Egypt is examined, the picture becomes clearer. The first physician to emerge is Imhotep, chief minister to King Djoser in the 3rd millennium BC, who designed one of the earliest pyramids, the Step Pyramid at Ṣaqqārah, and who was later regarded as the Egyptian god of medicine and identified with the Greek god Asclepius. Surer knowledge comes from the study of Egyptian papyri, especially the Ebers and Edwin Smith papyri discovered in the 19th century. The former is a list of remedies, with appropriate spells or incantations, while the latter is a surgical treatise on the treatment of wounds and other injuries.

The physician Imhotep

Contrary to what might be expected, the widespread practice of embalming the dead body did not stimulate study of human anatomy. The preservation of mummies has, however, revealed some of the diseases suffered at that time, including arthritis, tuberculosis of the bone, gout, tooth decay, bladder stones, and gallstones; there is evidence too of the parasitic disease schistosomiasis, which remains a scourge still. There seems to have been no syphilis or rickets.

The search for information on ancient medicine leads naturally from the papyri of Egypt to Hebrew literature. Though the Bible contains little on the medical practices of Old Testament times, it is a mine of information on social and personal hygiene. The Jews were indeed pioneers in matters of public health. (D.J.G./P.Rh.)

TRADITIONAL MEDICINE AND SURGERY IN THE ORIENT

India. Indian medicine has a long history. Its earliest concepts are set out in the sacred writings called the Vedas, especially in the metrical passages of the Atharvaveda, which may possibly date as far back as the 2nd millennium BC. According to a later writer, the system of medicine called Āyurveda was received by a certain Dhanvantari from Brahma, and Dhanvantari was deified as the god of medicine. In later times his status was gradually reduced, until he was credited with having been an earthly king who died of snakebite.

The period of Vedic medicine lasted until about 800 BC. The Vedas are rich in magical practices for the treatment of diseases and in charms for the expulsion of the demons traditionally supposed to cause diseases. The chief conditions mentioned are fever (*takman*), cough, consumption, diarrhea, dropsy, abscesses, seizures, tumours, and skin diseases (including leprosy). The herbs recommended for treatment are numerous.

The golden age of Indian medicine, from 800 BC until about AD 1000, was marked especially by the production of the medical treatises known as the *Caraka-saṃhitā* and *Suśruta-saṃhitā*, attributed, respectively, to Caraka, a physician, and Suśruta, a surgeon. Estimates place the *Caraka-saṃhitā* in its present form as dating from the 1st century AD, although there were earlier versions. The *Suśruta-saṃhitā* probably originated in the last centuries BC and had become fixed in its present form by the 7th century AD. Of somewhat lesser importance are the treatises attributed to Vagbhata. All later writings on Indian medicine were based on these works.

Because Hindus were prohibited by their religion from cutting the dead body, their knowledge of anatomy was limited. The *Suśruta-saṃhitā* recommends that a body be placed in a basket and sunk in a river for seven days. On its removal the parts could be easily separated without cutting. As a result of these crude methods, the emphasis in Hindu anatomy was given first to the bones and then to the muscles, ligaments, and joints. The nerves, blood vessels, and internal organs were very imperfectly known.

The Hindus believed that the body contains three elementary substances, microcosmic representatives of the three divine universal forces, which they called spirit

The Hindu view of the body

(air), phlegm, and bile (comparable to the humours of the Greeks). Health depends on the normal balance of these three elementary substances. The seven primary constituents of the body—blood, flesh, fat, bone, marrow, chyle, and semen—are produced by the action of the elementary substances. Semen was thought to be produced from all parts of the body and not from any individual part or organ.

Both Caraka and Suśruta state the existence of a large number of diseases (Suśruta says 1,120). Rough classifications of diseases are given. In all texts "fever," of which numerous types are described, is regarded as important. Phthisis (wasting disease, especially pulmonary tuberculosis) was apparently prevalent, and the Hindu physicians knew the symptoms of cases likely to terminate fatally. Smallpox was common, and it is probable that smallpox inoculation was practiced.

Hindu physicians employed all five senses in diagnosis. Hearing was used to distinguish the nature of the breathing, alteration in voice, and the grinding sound produced by the rubbing together of broken ends of bones. They appear to have had a good clinical sense, and their discourses on prognosis contain acute references to symptoms that have grave import. Magical beliefs still persisted, however, until late in the classical period; thus, the prognosis could be affected by such fortuitous factors as the cleanliness of the messenger sent to fetch the physician, the nature of his conveyance, or the types of persons the physician met on his journey to the patient.

Dietetic treatment was important and preceded any medicinal treatment. Fats were much used, internally and externally. The most important methods of active treatment were referred to as the "five procedures": the administration of emetics, purgatives, water enemas, oil enemas, and sneezing powders. Inhalations were frequently administered, as were leeching, cupping, and bleeding.

The Indian materia medica was extensive and consisted mainly of vegetable drugs, all of which were from indigenous plants. Caraka knew 500 medicinal plants, and Suśruta knew 760. But animal remedies (such as the milk of various animals, bones, gallstones) and minerals (sulfur, arsenic, lead, copper sulfate, gold) were also employed. The physicians collected and prepared their own vegetable drugs. Among those that eventually appeared in Western pharmacopoeias were cardamom and cinnamon.

As a result of the strict religious beliefs of the Hindus, hygienic measures were important in treatment. Two meals a day were decreed, with indications of the nature of the diet, the amount of water to be drunk before and after the meal, and the use of condiments. Bathing and care of the skin were carefully prescribed, as were cleansing of the teeth with twigs from named trees, anointing of the body with oil, and the use of eyewashes.

Hindu surgery

In surgery, ancient Hindu medicine reached its zenith. Operations performed by Hindu surgeons included excision of tumours, incision and draining of abscesses, punctures to release fluid in the abdomen, extraction of foreign bodies, repair of anal fistulas, splinting of fractures, amputations, cesarean sections, and stitching of wounds.

A broad array of surgical instruments were used. According to Suśruta the surgeon should be equipped with 20 sharp and 101 blunt instruments of various descriptions. The instruments were largely of steel. Alcohol seems to have been used as a narcotic during operations, and bleeding was stopped by hot oils and tar.

In two types of operations especially, the Hindus were outstanding. Stone in the bladder (vesical calculus) was common in ancient India, and the surgeons frequently removed the stones by lateral lithotomy. They also introduced plastic surgery. Amputation of the nose was one of the prescribed punishments for adultery, and repair was carried out by cutting from the patient's cheek or forehead a piece of tissue of the required size and shape and applying it to the stump of the nose. The results appear to have been tolerably satisfactory, and the modern operation is certainly derived indirectly from this ancient source. Hindu surgeons also operated on cataracts by couching, or displacing the lens to improve vision.

China. The Chinese system of medicine is of great antiquity and is independent of any recorded external influences. According to legend, Emperor Huang Ti (the Yellow Emperor) wrote the canon of internal medicine called the *Nei ching* in the 3rd millennium BC; but there is some evidence that in its present form it dates from no earlier than the 3rd century BC. Most of the Chinese medical literature is founded on the *Nei ching,* and it is still regarded as a great authority. Other famous works are the *Mo ching* (known in the West as the "Pulse Classic"), composed about AD 300; and the *Golden Mirror,* a compilation, made about AD 1700, of medical writings of the Han dynasty (202 BC–AD 220). European medicine began to obtain a footing in China early in the 19th century, but the native system is still widely practiced.

The yin–yang principle

Basic to traditional Chinese medicine is the dualistic cosmic theory of the yin and the yang. The yang, the male principle, is active and light and is represented by the heavens; the yin, the female principle, is passive and dark and is represented by the earth. The human body, like matter in general, is made up of five elements: wood, fire, earth, metal, and water. With these are associated other groups of five, such as the five planets, the five conditions of the atmosphere, the five colours, and the five tones. Health, character, and the success of all political and private ventures are determined by the preponderance, at the time, of the yin or the yang; and the great aim of ancient Chinese medicine is to control their proportions in the body.

The teachings of the religious sects forbade the mutilation of the dead human body; hence traditional anatomy rests on no sure scientific foundation. One of the most important writers on anatomy, Wang Ch'ing-jen, gained his knowledge from the inspection of dog-torn children who had died in a plague epidemic in AD 1798. Traditional Chinese anatomy is based on the cosmic system, which postulates the presence of such hypothetical structures as the 12 channels and the three so-called burning spaces. The body contains five organs (heart, lungs, liver, spleen, and kidneys), which store up but do not eliminate; and five viscera (such as the stomach, intestines, gallbladder, and bladder), which eliminate but do not store up. Each organ is associated with one of the planets, colours, tones, smells, and tastes. There are 365 bones and 365 joints in the body.

According to the physiology of traditional Chinese medicine, the blood vessels contain blood and air, in proportions varying with those of the yin and the yang. These two cosmic principles circulate in the 12 channels and control the blood vessels and hence the pulse. The *Nei ching* says that "the blood current flows continuously in a circle and never stops. It may be compared to a circle without beginning or end." On this insubstantial evidence it has been claimed that the Chinese anticipated Harvey's discovery of the circulation of the blood. Traditional Chinese pathology is also dependent on the theory of the yin and the yang; this led to an elaborate classification of diseases in which most of the types listed are without scientific foundation.

In diagnosis, detailed questions are asked about the history of the illness and about such things as the patient's taste, smell, and dreams. Conclusions are drawn from the quality of the voice, and note is made of the colour of the face and of the tongue. The most important part of the investigation, however, is the examination of the pulse. Wang Shu-ho, who wrote the "Pulse Classic," lived in the 3rd century BC, and innumerable commentaries were written on his work. The pulse is examined in several places, at different times, and with varying degrees of pressure. The operation may take as long as three hours. It is often the only examination made, and it is used both for diagnosis and for prognosis. Not only are the diseased organs ascertained but the time of death or recovery may be foretold.

The pulse doctrine

The Chinese materia medica has always been extensive and consists of vegetable, animal (including human), and mineral remedies. There were famous herbals from ancient times; but all these, to the number of about 1,000, were embodied by Li Shih-chen in the compilation of *Pen-ts'ao kang-mu* (the "Great Pharmacopoeia") in the 16th century AD. This work, in 52 volumes, has been frequently revised

and reprinted and is still authoritative. The use of drugs is mainly to restore the harmony of the yin and the yang and is also related to such matters as the five organs, the five planets, and the five colours. The art of prescribing is therefore complex.

Among the drugs taken over by Western medicine from the Chinese are rhubarb, iron (for anemia), castor oil, kaolin, aconite, camphor, and *Cannabis sativa* (Indian hemp). Chaulmoogra oil was used by the Chinese for leprosy from at least the 14th century, and about a century ago it began to be used for this purpose by Western physicians. The herb mahuang (*Ephedra vulgaris*) has been used in China for at least 4,000 years, and the isolation of the alkaloid ephedrine from it has greatly improved the Western treatment of asthma and similar conditions.

The most famous and expensive of Chinese remedies is ginseng. Western analysis has shown that it has diuretic and other properties but is of doubtful value. In recent years reserpine, the active principle of the Chinese plant *Rauwolfia,* has been isolated; it is now effectively used in the treatment of high blood pressure and some emotional and mental conditions.

Hydrotherapy is probably of Chinese origin, since cold baths were used for fevers as early as 180 BC. The inoculation of smallpox matter, in order to produce a mild but immunizing attack of the disease, was practiced in China from ancient times and came to Europe about 1720. Another treatment is moxibustion, which consists in making a small, moistened cone (moxa) of powdered leaves of mugwort, or wormwood (*Artemisia* species), applying it to the skin, igniting it, and then crushing it into the blister so formed. Other substances are also used for the moxa. Dozens of these are sometimes applied at one sitting. The practice is often associated with acupuncture.

Acupuncture

Acupuncture consists of the insertion into the skin and underlying tissues of a metal needle, either hot or cold. The theory is that the needle affects the distribution of the yin and the yang in the hypothetical channels and burning spaces of the body. The site of the insertion is chosen to affect a particular organ or organs. The practice of acupuncture dates from before 2500 BC and is peculiarly Chinese. Little of practical importance has been added since that date, although there have been many well-known treatises on the subject.

A bronze model, *c.* AD 860, shows the hundreds of specified points for the insertion of the needle; this was the forerunner of countless later models and diagrams. The needles used are three to 24 centimetres (about one to nine inches) in length. They are often inserted with considerable force and after insertion may be agitated or screwed to the left or right. Acupuncture, often combined with moxibustion, is still widely used for many diseases, including fractures. Recently people in the Western world have turned to acupuncturists for relief from pain and other symptoms. There is some speculation that the treatment may trigger the brain to release morphinelike substances called endorphins, which presumably reduce the feeling of pain and its concomitant emotions.

Japan. The most interesting features of Japanese medicine are the extent to which it was derivative and the rapidity with which, after a slow start, it became westernized and scientific. In the early pre-Christian Era disease was regarded as sent by the gods or produced by the influence of evil spirits. Treatment and prevention were based largely on religious practices, such as prayers, incantations, and exorcism; at a later date drugs and bloodletting were also employed.

Beginning in AD 608, when young Japanese physicians were sent to China for a long period of study, Chinese influence on Japanese medicine was paramount. In 982, Tamba Yasuyori completed the 30-volume *Ishinhō,* the oldest Japanese medical work still extant. This work discusses diseases and their treatment, classified mainly according to the affected organs or parts. It is based entirely on older Chinese medical works, with the yin and yang concept underlying the theory of disease causation.

In 1570 a 15-volume medical work was published by Menase Dōsan, who also wrote at least five other works. In the most significant of these, the *Keitekishū* (a manual of the practice of medicine, 1574), diseases—or sometimes merely symptoms—are classified and described in 51 groups; the work is unusual in that it includes a section on the diseases of old age. Another distinguished physician and teacher of the period, Nagata Tokuhun, whose important books were the *I-no-ben* (1585) and the *Baika mujinzo* (1611), held that the chief aim of the medical art was to support the natural force, and consequently that it was useless to persist with stereotyped methods of treatment unless the physician had the cooperation of the patient.

European influences

European medicine was introduced into Japan in the 16th century by Jesuit missionaries and again in the 17th century by Dutch physicians. Translations of European books on anatomy and internal medicine were made in the 18th century, and in 1836 an influential Japanese work on physiology appeared. In 1857 a group of Dutch-trained Japanese physicians founded a medical school in Edo (later Tokyo) that is regarded as the beginning of the medical faculty of the Imperial University of Tokyo.

During the last third of the 18th century it became government policy to westernize Japanese medicine, and great progress was made in the foundation of medical schools and the encouragement of research. Important medical breakthroughs by the Japanese followed, among them the discovery of the plague bacillus in 1894, the discovery of a dysentery bacillus in 1897, the isolation of adrenaline (epinephrine) in crystalline form in 1901, and the first experimental production of a tar-induced cancer in 1918.

(E.A.U./P.Rh.)

THE ROOTS OF WESTERN MEDICINE

Early Greece. The transition from magic to science was a gradual process that lasted for centuries, and there is little doubt that ancient Greece inherited much from Babylonia and Egypt, and even India and China. Twentieth-century readers of the Homeric tales the *Iliad* and the *Odyssey* may well be bewildered by the narrow distinction between gods and men among the characters and between historical fact and poetic fancy in the story. Two characters, the military surgeons Podaleirius and Machaon, are said to have been sons of Asclepius, the god of medicine. The divine Asclepius may have originated in a human Asclepius who lived about 1200 BC and is said to have performed many miracles of healing.

Asclepius was worshiped in hundreds of temples throughout Greece, the remains of which may still be seen at Epidaurus, Cos, Athens, and elsewhere. To these resorts, or hospitals, sick persons went for the healing ritual known as incubation, or temple sleep. They lay down to sleep in the dormitory, or *abaton,* and were visited in their dreams by Asclepius or by one of his priests, who gave advice. In the morning the patient often is said to have departed cured. There are at Epidaurus many inscriptions recording cures, though there is no mention of failures or deaths.

Cure by incubation

Diet, baths, and exercises played their part in the treatment, and it would appear that these temples were the prototype of modern health resorts. Situated in a peaceful spot, with gardens and fountains, each had its theatre for amusements and its stadium for athletic contests. The cult of incubation continued far into the Christian Era. In Greece, some of the Aegean islands, Sardinia, and Sicily, sick persons are still taken to spend a night in certain churches in the hope of a cure.

It was, however, the work of the early philosophers, rather than that of the priests of Asclepius, that impelled Greeks to refuse to be guided solely by supernatural influence and moved them to seek out for themselves the causes and reasons for the strange ways of nature. The 6th-century philosopher Pythagoras, whose chief discovery was the importance of numbers, also investigated the physics of sound, and his views influenced the medical thought of his time. In the 5th century BC Empedocles set forth the view that the universe is composed of four elements—fire, air, earth, and water; this conception led to the doctrine of the four bodily humours: blood; phlegm; choler, or yellow bile; and melancholy, or black bile. The maintenance of health was held to depend upon the harmony of the four humours.

Hippocrates. Medical thought had reached this stage and had partially discarded the conceptions based upon magic and religion by 460 BC, the year that Hippocrates is said to have been born. Although he has been called the father of medicine, little is known of his life, and there may, in fact, have been several men of this name; or Hippocrates may have been the author of only some, or none, of the books that make up the Hippocratic Collection (*Corpus Hippocraticum*). Ancient writers held that Hippocrates taught and practiced medicine in Cos, the island of his birth, and in other parts of Greece, including Athens, and that he died at an advanced age.

Whether Hippocrates was one man or several, the works attributed to him mark the stage in Western medicine where disease was coming to be regarded as a natural rather than a supernatural phenomenon and doctors were encouraged to look for physical causes of illness. Some of the works, notably the *Aphorismi* (*Aphorisms*), were used as textbooks until the 19th century. The first and best-known aphorism is, "Life is Short, Art long, Occasion sudden and dangerous, Experience deceitful, and Judgment difficult" (often shortened to the Latin tag, "Ars longa, vita brevis"). This is followed by brief comments on diseases and symptoms, many of which remain valid.

The thermometer and the stethoscope were not then known; nor, indeed, did Hippocrates employ any aid to diagnosis beyond his own powers of observation and logical reasoning. He had an extraordinary ability to foretell the course of a malady, and he laid more stress upon the expected outcome, or prognosis, of a disease than upon its identification, or diagnosis. He had no patience with the idea that disease was a punishment sent by the gods. Writing of epilepsy, then called "the sacred disease," he said, "It is not any more sacred than other diseases, but has a natural cause, and its supposed divine origin is due to man's inexperience. Every disease," he continued, "has its own nature, and arises from external causes."

Hippocrates noted the effect of food, of occupation, and especially of climate in causing disease, and one of his most interesting books, entitled *De aëre, aquis et locis* (*Air, Waters and Places*), would today be classed as a treatise on human ecology. Pursuing this line of thought, Hippocrates stated that "our natures are the physicians of our diseases" and advocated that this tendency to natural cure should be fostered. He laid much stress on diet and the use of few drugs. He knew well how to describe illness clearly and concisely and recorded failures as well as successes; he viewed disease with the eye of the naturalist and studied the entire patient in his environment.

The Hippocratic oath

Perhaps the greatest legacy of Hippocrates is the charter of medical conduct embodied in the so-called Hippocratic oath, which has been adopted as a pattern by physicians throughout the ages:

> I swear by Apollo the physician, and Asclepius, and Health, and All-heal, and all the gods and goddesses . . . to reckon him who taught me this Art equally dear to me as my parents, to share my substance with him, and relieve his necessities if required; to look upon his offspring in the same footing as my own brothers, and to teach them this art, if they shall wish to learn it, without fee or stipulation; and that by precept, lecture, and every other mode of instruction, I will impart a knowledge of the Art to my own sons, and those of my teachers, and to disciples bound by a stipulation and oath according to the law of medicine, but to none others. I will follow that system of regimen which, according to my ability and judgment, I consider for the benefit of my patients, and abstain from whatever is deleterious and mischievous. I will give no deadly medicine to any one if asked, nor suggest any such counsel; and in like manner I will not give to a woman a pessary to produce abortion. . . . Into whatever houses I enter, I will go into them for the benefit of the sick, and will abstain from every voluntary act of mischief and corruption; and, further from the seduction of females or males, of freemen and slaves. Whatever, in connection with my professional practice or not, in connection with it, I see or hear, in the life of men, which ought not to be spoken of abroad, I will not divulge, as reckoning that all such should be kept secret.

Not strictly an oath, it was, rather, an ethical code or ideal, an appeal for right conduct. In one or other of its many versions, it has guided the practice of medicine throughout the world for more than 2,000 years.

Hellenistic and Roman medicine. In the following century the work of Aristotle, regarded as the first great biologist, was of inestimable value to medicine. A pupil of Plato at Athens and tutor to Alexander the Great, Aristotle studied the entire world of living things. He laid what can be identified as the foundations of comparative anatomy and embryology, and his views influenced scientific thinking for the next 2,000 years.

After the time of Aristotle, the centre of Greek culture shifted to Alexandria, where a famous medical school was established in about 300 BC. There, the two best medical teachers were Herophilus, whose treatise on anatomy may have been the first of its kind, and Erasistratus, regarded by some as the founder of physiology. Erasistratus noted the difference between sensory and motor nerves but thought that the nerves were hollow tubes containing fluid and that air entered the lungs and heart and was carried through the body in the arteries. Alexandria continued as a centre of medical teaching even after the Roman Empire had attained supremacy over the Greek world, and medical knowledge remained predominantly Greek.

Asclepiades of Bithynia (born 124 BC) differed from Hippocrates in that he denied the healing power of nature and insisted that disease should be treated safely, speedily, and agreeably. An opponent of the humoral theory, he drew upon the atomic theory of the 5th-century Greek philosopher Democritus in advocating a doctrine of *strictum et laxum*—the attribution of disease to the contracted or relaxed condition of the solid particles that he believed make up the body. To restore harmony among the particles and thus effect cures, Asclepiades used typically Greek remedies: massage, poultices, occasional tonics, fresh air, and corrective diet. He gave particular attention to mental disease, clearly distinguishing hallucinations from delusions. He released the insane from confinement in dark cellars and prescribed a regimen of occupational therapy, soothing music, soporifics (especially wine), and exercises to improve the attention and memory.

Asclepiades did much to win acceptance for Greek medicine in Rome. Aulus Cornelius Celsus, the Roman nobleman who wrote *De medicina* about AD 30, gave a classic account of Greek medicine of the time, including descriptions of elaborate surgical operations. His book, overlooked in his day, enjoyed a wide reputation during the Renaissance.

During the early centuries of the Christian Era, Greek doctors thronged to Rome. The most illustrious of them was Galen, who began practicing there in AD 161. He acknowledged his debt to Hippocrates and followed the Hippocratic method, accepting the doctrine of the humours. He laid stress on the value of anatomy, and he virtually founded experimental physiology. Galen recognized that the arteries contain blood and not merely air. He showed how the heart sets the blood in motion in an ebb and flow fashion, but he had no idea that the blood circulates. Dissection of the human body was at that time illegal, so that he was forced to base his knowledge upon the examination of animals, particularly apes. A voluminous writer who stated his views forcibly and with confidence, he remained for centuries the undisputed authority from whom no one dared to differ.

Galen

Another influential physician of the 2nd century AD was Soranus of Ephesus, who wrote authoritatively on childbirth, infant care, and women's diseases. An opponent of abortion, he advocated numerous means of contraception. He also described how to assist a difficult delivery by turning the fetus in the uterus (podalic version), a life-saving technique that was subsequently lost sight of until it was revived in the 16th century.

Although the contribution of Rome to the practice of medicine was negligible compared with that of Greece, in matters of public health the Romans set the world a great example. The city of Rome had an unrivaled water supply. Gymnasiums and public baths were provided, and there was even domestic sanitation and adequate disposal of sewage. The army had its medical officers, public physicians were appointed to attend the poor, and hospitals were built; a Roman hospital excavated near Düsseldorf, W.Ger., was found to be strikingly modern in design.

CHRISTIAN AND MUSLIM RESERVOIRS OF LEARNING

After the fall of Rome, learning was no longer held in high esteem, experiment was discouraged, and originality became a dangerous asset. During the early Middle Ages medicine passed into the widely diverse hands of the Christian Church and Arab scholars.

Translators and saints. It is sometimes stated that the early Christian Church had an adverse effect upon medical progress. Disease was regarded as a punishment for sin, and such chastening demanded only prayer and repentance. Moreover, the human body was held sacred and dissection was forbidden. But the infinite care and nursing bestowed upon the sick under Christian auspices must outweigh any intolerance shown toward medicine in the early days.

Preservation of Greek medical manuscripts

Perhaps the greatest service rendered to medicine by the church was the preservation and transcription of the classical Greek medical manuscripts. These were translated into Latin in many medieval monasteries, and the Nestorian Christians (an Eastern church) established a school of translators to render the Greek texts into Arabic. This famous school, and also a great hospital, were located at Jundi Shāhpūr in southwest Persia, where the chief physician was Jurjīs ibn Bukhtīshūʿ, the first of a dynasty of translators and physicians that lasted for six generations. A later translator of great renown was Ḥunayn ibn Isḥāq, or Johannitus (born AD 809), whose translations were said to be worth their weight in gold.

About this time there appeared a number of saints whose names were associated with miraculous cures. Among the earliest of these were twin brothers, Cosmas and Damian, who suffered martyrdom (*c.* AD 303) and who became the patron saints of medicine. Other saints were invoked as powerful healers of certain diseases, such as St. Vitus for chorea (or St. Vitus' dance) and St. Anthony for erysipelas (or St. Anthony's fire). The cult of these saints was widespread in medieval times, and a later cult, that of St. Roch for plague, was widespread during the plague-ridden years of the 14th century.

Arabian medicine. A second reservoir of medical learning during those times was the great Muslim empire, which extended from Persia to Spain. Although it is customary to speak of Arabian medicine in describing this period, not all of the physicians were Arabs or natives of Arabia. Nor, indeed, were they all Muslims: some were Jews, some Christians, and they were drawn from all parts of the empire. One of the earliest figures was Rhazes, a Persian born in the last half of the 9th century near modern Tehrān, who wrote a voluminous treatise on medicine, *Kitāb al-hāwī* ("Comprehensive Book"), but whose most famous work, *De variolis et morbillis* (*A Treatise on the Smallpox and Measles*), distinguishes between these two diseases and gives a clear description of both.

Avicenna

Of later date was Avicenna (980–1037), also a Persian, who has been called the prince of physicians and whose tomb at Hamadan has become a place of pilgrimage. He could repeat the Qurʾān before he was 10 years old and at the age of 18 became court physician. His principal medical work, *al-Qānūn fī aṭ-ṭibb* (*The Canon of Medicine*), became a classic and was used at many medical schools—at Montpellier, Fr., as late as 1650—and reputedly is still used in the East.

The greatest contribution of Arabian medicine was in chemistry and in the knowledge and preparation of medicines. The chemists of that time were alchemists, and their pursuit was mainly a search for the philosopher's stone, which supposedly would turn common metals into gold. In the course of their experiments, however, numerous substances were named and characterized, and some were found to have medicinal value. Many drugs now in use are of Arab origin, as are such processes as distillation and sublimation.

At that period, and indeed throughout most historical times, surgery was considered inferior to medicine, and surgeons were held in low regard. The renowned Spanish surgeon Abū al-Qāsim (Albucasis), however, did much to raise the status of surgery in Córdoba, an important centre of commerce and culture with a hospital and medical school equal to those of Cairo and Baghdad. A careful and conservative practitioner, he wrote the first illustrated surgical text, which held wide influence in Europe for centuries.

Another great doctor of Córdoba, born in the 12th century, just as the sun of Arabian culture was setting, was the Jewish philosopher Maimonides. Banished from the city because he would not become a Muslim, he eventually went to Cairo, where the law was more lenient and where he acquired a reputation so high that he became physician to Saladin, the Saracen leader. (He was the original of El Hakim in Sir Walter Scott's *Talisman.*) A few of his works, written in Arabic, were eventually translated into Latin and printed; perhaps the best known is *Yad-Hachazakah,* or *The Code of Maimonides,* an ethical guide still esteemed in Jewish medical circles.

MEDIEVAL AND RENAISSANCE EUROPE

Salerno and the medical schools. At about the same time that Arabian medicine flourished, the first organized medical school in Europe was established at Salerno, in southern Italy. Although the school of Salerno produced no brilliant genius and no startling discovery, it was the outstanding medical institution of its time and the parent of the great medieval schools soon to be founded at Montpellier and Paris, in France, and at Bologna and Padua, in Italy. Salerno drew scholars from near and far. Remarkably liberal in some of its views, Salerno admitted women as medical students. The school owed much to the enlightened Holy Roman emperor Frederick II, who decreed in 1221 that no one should practice medicine until he had been publicly approved by the masters of Salerno.

The Salernitan school also produced a literature of its own; the best-known work, of uncertain date and of composite authorship, was the *Regimen Sanitatis Salernitanum* ("Salernitan Guide to Health"). Written in verse, it has appeared in numerous editions and has been translated into many languages. Among its oft-quoted couplets is the following:

> Use three physicians still, first Doctor Quiet,
> Next Doctor Merryman, and Doctor Diet.

Montpellier

Salerno yielded its place as the premier medical school of Europe to Montpellier in about 1200. John of Gaddesden, the model for the "doctour of physick" in Chaucer's *Canterbury Tales,* was one of the English students there. That he relied upon astrology and upon the doctrine of the humours is evident from Chaucer's description:

> Well could he guess the ascending of the star
> Wherein his patient's fortunes settled were.
> He knew the course of every malady,
> Were it of cold or heat or moist or dry. . . .

Medieval physicians analyzed symptoms, examined excreta, and made their diagnoses. Then they might prescribe diet, rest, sleep, exercise, or baths; or they could administer emetics and purgatives or bleed the patient. Surgeons could treat fractures and dislocations, repair hernias, and perform amputations and a few other operations. Some of them prescribed opium, mandragora, or alcohol to deaden pain. Childbirth was left to midwives, who relied on folklore and tradition.

Great hospitals were established during the Middle Ages by religious foundations, and infirmaries were attached to abbeys, monasteries, priories, and convents. Doctors and nurses in these institutions were members of religious orders and combined spiritual with physical healing.

The spread of new learning. Among the teachers of medicine in the medieval universities there were many who clung to the past, but there were not a few who determined to explore new lines of thought. The new learning of the Renaissance, born in Italy, grew and expanded slowly. Two great 13th-century scholars who influenced medicine were Roger Bacon, an active observer and tireless experimenter, and Albertus Magnus, a distinguished philosopher and scientific writer.

About this time Mondino dei Liucci taught at Bologna. Prohibitions against human dissection were slowly lifting, and Mondino performed his own dissections rather than following the customary procedure of entrusting the task to a menial. Although he perpetuated the errors of Galen,

his *Anothomia,* published in 1316, was the first practical manual of anatomy. Foremost among the surgeons of the day was Guy de Chauliac, a physician to three popes at Avignon. His *Chirurgia magna* ("Great Surgery"), based on observation and experience, had a profound influence upon the progress of surgery.

The Renaissance in the 14th, 15th, and 16th centuries was much more than just a reviving of interest in Greek and Roman culture; it was rather a change of outlook, an eagerness for discovery, a desire to escape from the limitations of tradition and to explore new fields of thought and action. In medicine, it was perhaps natural that anatomy and physiology, the knowledge of the human body and its workings, should be the first aspects of medical learning to receive attention from those who realized the need for reform.

Vesalius

It was in 1543 that Andreas Vesalius, a young Belgian professor of anatomy at the University of Padua, published *De humani corporis fabrica* ("On the Structure of the Human Body"). Based on his own dissections, this seminal work corrected many of Galen's errors. By his scientific observations and methods, Vesalius showed that Galen could no longer be regarded as the final authority. His work at Padua was continued by Gabriel Fallopius and, later, by Hieronymus Fabricius ab Aquapendente; it was his work on the valves in the veins, *De venarum ostiolis* (1603), that suggested to his pupil William Harvey his revolutionary theory of the circulation of the blood, one of the great medical discoveries.

Surgery profited from the new outlook in anatomy, and the great reformer Ambroise Paré dominated the field in the 16th century. Paré was surgeon to four kings of France, and he has deservedly been called the father of modern surgery. In his autobiography, written after he had retired from 30 years of service as an army surgeon, Paré described how he had abolished the painful practice of cautery to stop bleeding and used ligatures and dressings instead. His favourite expression, "I dressed him; God healed him," is characteristic of this humane and careful doctor.

In Britain during this period surgery, which was performed by barber-surgeons, was becoming regulated and organized under royal charters. Companies were thus formed that eventually became the royal colleges of surgeons in Scotland and England. Physicians and surgeons united in a joint organization in Glasgow, and a college of physicians was founded in London.

The 16th-century medical scene was enlivened by the enigmatic physician and alchemist who called himself Paracelsus. Born in Switzerland, he traveled extensively throughout Europe, gaining medical skills and practicing and teaching as he went. In the tradition of Hippocrates, Paracelsus stressed the power of nature to heal; but unlike Hippocrates he believed also in the power of supernatural forces, and he violently attacked the medical treatments of his day. Eager for reform, he allowed his intolerance to outweigh his discretion, as when he prefaced his lectures at Basel by publicly burning the works of Avicenna and Galen. The authorities and medical men were understandably outraged. Widely famous in his time, Paracelsus remains a controversial figure to this day. Despite his turbulent career, however, he did attempt to bring a more rational approach to diagnosis and treatment, and he introduced the use of chemical drugs in place of herbal remedies.

A contemporary of Paracelsus, Girolamo Fracastoro of Italy was a scholar cast from a very different mold. His account of the disease syphilis, entitled *Syphilis sive morbus Gallicus* (1530; "Syphilis or the French Disease"), was written in verse. Although Fracastoro called syphilis the French disease, others called it the Neapolitan disease, for it was said to have been brought to Naples from America by the sailors of Christopher Columbus. Its origin is still questioned, however. Fracastoro was interested in epidemic infection, and he offered the first scientific explanation of disease transmission. In his great work, *De contagione et contagiosis morbis* (1546), he theorized that the seeds of certain diseases are imperceptible particles transmitted by air or by contact.

Fracastoro on syphilis

THE ENLIGHTENMENT

In the 17th century the natural sciences moved forward on a broad front. There were attempts to grapple with the nature of science, as expressed in the works of thinkers like Francis Bacon, Descartes, and Newton. New knowledge of chemistry superseded the theory that all things are made up of earth, air, fire, and water, and the old Aristotelian ideas began to be discarded. The supreme 17th-century achievement in medicine was Harvey's explanation of the circulation of blood.

Harvey and the experimental method. Born in Folkestone, Eng., William Harvey studied at Cambridge University and then spent several years at Padua, where he came under the influence of Fabricius. He established a successful medical practice in London and by precise observation and scrupulous reasoning developed his theory of circulation. In 1628 he published his classic book *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus* (*Concerning the Motion of the Heart and Blood*), often called *De Motu Cordis.*

That the book aroused controversy is not surprising. There were still many who adhered to the teaching of Galen that the blood follows an ebb and flow movement in the blood vessels. Harvey's work was the result of many careful experiments, but few of his critics took the trouble to repeat the experiments, simply arguing in favour of the older view. His second great book, *Exercitationes de generatione animalium* ("Experiments Concerning Animal Generation"), published in 1651, laid the foundation of modern embryology.

Harvey's discovery of the circulation of the blood was a landmark of medical progress; the new experimental method by which the results were secured was as noteworthy as the work itself. Following the method described by the philosopher Francis Bacon, he drew the truth from experience and not from authority.

There was one gap in Harvey's argument: he was obliged to assume the existence of the capillary vessels that conveyed the blood from the arteries to the veins. This link in the chain of evidence was supplied by Marcello Malpighi of Bologna (who was born in 1628, the year of publication of *De Motu Cordis*). With a primitive microscope Malpighi saw a network of tiny blood vessels in the lung of a frog. Harvey also failed to show why the blood circulated. After Robert Boyle had shown that air is essential to animal life, it was Richard Lower who traced the interaction between air and the blood. Eventually the importance of oxygen, which was confused for a time by some as phlogiston, was revealed, although it was not until the late 18th century that the great chemist Antoine-Laurent Lavoisier discovered the essential nature of oxygen and clarified its relation to respiration.

The microscope

Although the compound microscope had been invented slightly earlier, probably in Holland, its development, like that of the telescope, was the work of Galileo. He was the first to insist upon the value of measurement in science and in medicine, thus replacing theory and guesswork with accuracy. The great Dutch microscopist Antonie van Leeuwenhoek devoted his long life to microscopical studies and was probably the first to see and describe bacteria, reporting his results to the Royal Society of London. In England, Robert Hooke, who was Boyle's assistant and curator to the Royal Society, published his *Micrographia* in 1665, which discussed and illustrated the microscopic structure of a variety of materials.

The futile search for an easy system. Several attempts were made in the 17th century to discover an easy system that would guide the practice of medicine. A substratum of superstition still remained. Richard Wiseman, surgeon to Charles II, affirmed his belief in the "royal touch" as a cure for king's evil, or scrofula, while even the learned English physician Thomas Browne stated that witches really existed. There was, however, a general desire to discard the past and adopt new ideas.

The view of the French philosopher René Descartes that the human body is a machine and that it functions mechanically had its repercussions in medical thought. One group adopting this explanation called themselves the iatrophysicists; another school, preferring to view life as

a series of chemical processes, were called iatrochemists. Santorio Santorio, working at Padua, was an early exponent of the iatrophysical view and a pioneer investigator of metabolism. He was especially concerned with the measurement of what he called "insensible perspiration," described in his book *De statica medicina* (1614; "On Medical Measurement"). Another Italian, who developed the idea still further, was Giovanni Alfonso Borelli, a professor of mathematics at Pisa University, who gave his attention to the mechanics and statics of the body and to the physical laws that govern its movements.

The iatrochemical school was founded at Brussels by Jan Baptist van Helmont, whose writings are tinged with the mysticism of the alchemist. A more logical and intelligible view of iatrochemistry was advanced by Franciscus Sylvius, at Leiden; and in England a leading exponent of the same school was Thomas Willis, who is better known for his description of the brain in his *Cerebri anatome nervorumque descriptio et usus* ("Anatomy of the Brain and Descriptions and Functions of the Nerves"), published in 1664 and illustrated by Christopher Wren.

It soon became apparent that no easy road to medical knowledge and practice was to be found along these channels and that the best method was the age-old system of straightforward clinical observation initiated by Hippocrates. The need for a return to these views was strongly urged by Thomas Sydenham, well named "the English Hippocrates." Sydenham was not a voluminous writer and, indeed, had little patience with book learning in medicine; nevertheless he gave excellent descriptions of the phenomena of disease. His greatest service, much needed at the time, was to divert physicians' minds from speculation and lead them back to the bedside, where the true art of medicine could be studied.

Medicine in the 18th century. Even in the 18th century the search for a simple way of healing the sick continued. In Edinburgh the writer and lecturer John Brown expounded his view that there were only two diseases, sthenic (strong) and asthenic (weak), and two treatments, stimulant and sedative; his chief remedies were alcohol and opium. Lively and heated debates took place between his followers, the Brunonians, and the more orthodox Cullenians (followers of William Cullen, a professor of medicine at Glasgow), and the controversy spread to the medical centres of Europe.

At the opposite end of the scale, at least in regard to dosage, was Samuel Hahnemann, of Leipzig, the originator of homeopathy, a system of treatment involving the administration of minute doses of drugs whose effects resemble the effects of the disease being treated. His ideas had a salutary effect upon medical thought at a time when prescriptions were lengthy and doses were large, and his system has had many followers.

By the 18th century the medical school at Leiden had grown to rival that of Padua, and many students were attracted there from abroad. Among them was John Monro, an army surgeon, who resolved that his native city of Edinburgh should have a similar medical school. He specially educated his son Alexander with a view to having him appointed professor of anatomy, and the bold plan was successful. Alexander Monro studied at Leiden under Hermann Boerhaave, the central figure of European medicine and the greatest clinical teacher of his time. Subsequently, three generations of Alexander Monros taught anatomy at Edinburgh University over a continuous period of 126 years. Medical education was increasingly incorporated into the universities of Europe, and Edinburgh became the leading academic centre for medicine in Britain.

Leiden and Edinburgh

In 18th-century London, Scottish doctors were the leaders in surgery and obstetrics. The noted teacher John Hunter conducted extensive researches in comparative anatomy and physiology, founded surgical pathology, and raised surgery to the level of a respectable branch of science. His brother William Hunter, an eminent teacher of anatomy, became famous as an obstetrician. Male doctors were now attending women in childbirth, and the leading obstetrician in London was William Smellie. His well-known *Treatise on the Theory and Practice of Midwifery,* published in three volumes in 1752–64, contained the first systematic discussion on the safe use of obstetrical forceps, which have since saved countless lives. Smellie placed midwifery on a sound scientific footing and helped to establish obstetrics as a recognized medical discipline.

The science of modern pathology also had its beginnings in this century. Giovanni Battista Morgagni, of Padua, in 1761 published his massive work *De sedibus et causis morborum* (*The Seats and Causes of Diseases Investigated by Anatomy*), a description of the appearances found by postmortem examination of almost 700 cases, in which he attempted to correlate the findings after death with the clinical picture in life.

On the basis of work begun in the 18th century, René Laënnec, a native of Brittany, who practiced medicine in Paris, invented a simple stethoscope, or *cylindre,* as it was originally called. In 1819 he wrote a treatise, *De l'auscultation médiate* ("On Mediate Auscultation"), describing many of the curious sounds in the heart and lungs that are revealed by the instrument. Meanwhile a Viennese physician, Leopold Auenbrugger, discovered another method of investigating diseases of the chest, that of percussion. The son of an innkeeper, he is said to have conceived the idea of tapping with the fingers when he recalled that he had used this method to gauge the level of the fluid contents of his father's casks.

One highly significant medical advance, late in the century, was vaccination. Smallpox, disfiguring and often fatal, was widely prevalent. Inoculation, which had been practiced in the East, was popularized in England in 1721–22 by Lady Mary Wortley Montagu, who is best known for her letters. She observed the practice in Turkey, where it produced a mild form of the disease, thus securing immunity, although not without danger. The next step was taken by Edward Jenner, a country practitioner who had been a pupil of John Hunter. In 1796 Jenner began inoculations with material from cowpox (the bovine form of the disease); and when he later inoculated the same subject with smallpox, the disease did not appear. This procedure—vaccination—has been responsible for eradicating the disease.

Smallpox vaccination

Public health and hygiene were receiving more attention during the 18th century. Population statistics began to be kept, and suggestions arose concerning health legislation. Hospitals were established for a variety of purposes. In Paris, Philippe Pinel initiated bold reforms in the care of the mentally ill, releasing them from their chains and discarding the long-held notion that insanity was caused by demon possession.

Conditions improved for sailors and soldiers as well. James Lind, a British naval surgeon from Edinburgh, recommended fresh fruits and citrus juices to prevent scurvy, a remedy discovered by the Dutch in the 16th century. When the British navy adopted Lind's advice—decades later—this deficiency disease was eliminated. In 1752 a Scotsman, John Pringle, published his classic *Observations on the Diseases of the Army,* which contained numerous recommendations for the health and comfort of the troops. Serving with the British forces during the War of the Austrian Succession, he suggested in 1743 that military hospitals on both sides should be regarded as sanctuaries; this plan eventually led to the establishment of the Red Cross organization in 1864.

Two pseudoscientific doctrines relating to medicine emerged from Vienna in the latter part of the century and attained wide notoriety. Mesmerism, a belief in "animal magnetism" sponsored by Franz Anton Mesmer, probably owed any therapeutic value it had to suggestions given while the patient was under hypnosis. Phrenology, propounded by Franz Joseph Gall, held that the contours of the skull were a guide to an individual's mental faculties and character traits; this theory remained popular throughout the 19th century.

At the same time, sound scientific thinking was making steady progress, and advances in physics, chemistry, and the biological sciences were converging to form a rational scientific basis for every branch of clinical medicine. New knowledge disseminated thoughout Europe and traveled across the sea, where centres of medical excellence were being established in America.

The rise of scientific medicine in the 19th century

The historic portrayal of modern medicine becomes more difficult. Discoveries multiply, and the number of eminent doctors is so great that the history is apt to become a series of biographies. Nevertheless, it is possible to discern the leading trends in modern medical thought.

PHYSIOLOGY

By the beginning of the 19th century, the structure of the human body was almost fully known, due to new methods of microscopy and of injections. Even the body's microscopic structure was understood. But as important as anatomical knowledge was an understanding of physiological processes, which were rapidly being elucidated, especially in Germany. There, physiology became established as a distinct science under the guidance of Johannes Müller, who was a professor at Bonn and then at the University of Berlin. An energetic worker and an inspiring teacher, he described his discoveries in a famous textbook, *Handbuch der Physiologie des Menschen* ("Manual of Human Physiology"), published in the 1830s.

The work of Johannes Müller and his students

Among Müller's illustrious pupils were Hermann von Helmholtz, who made significant discoveries relating to sight and hearing and who invented the ophthalmoscope; and Rudolf Virchow, one of the century's great medical scientists, whose outstanding achievement was his conception of the cell as the centre of all pathological changes. Virchow's work *Die Cellularpathologie,* published in 1858, gave the deathblow to the outmoded view that disease is due to an imbalance of the four humours.

In France the most brilliant physiologist of the time was Claude Bernard, whose many important discoveries were the outcome of carefully planned experiments. His researches clarified the role of the pancreas in digestion, revealed the presence of glycogen in the liver, and explained how the contraction and expansion of the blood vessels are controlled by vasomotor nerves. He proposed the concept of the internal environment—the chemical balance in and around the cells—and the importance of its stability. His *Introduction à l'étude de la médecine expérimentale* (1865; *An Introduction to the Study of Experimental Medicine*) is still worthy of study by all who undertake research.

VERIFICATION OF THE GERM THEORY

Perhaps the overarching medical advance of the 19th century, certainly the most spectacular, was the conclusive demonstration that certain diseases, as well as the infection of surgical wounds, were directly caused by minute living organisms. This discovery changed the whole face of pathology and effected a complete revolution in the practice of surgery.

The idea that disease was caused by entry into the body of imperceptible particles was of ancient date. It had been expressed by the Roman encyclopaedist Varro as early as 100 BC, by Fracastoro in 1546, by Athanasius Kircher and Pierre Borel about a century later, and by Francesco Redi, who in 1684 wrote his *Osservazioni intorno agli animali viventi che si trovano negli animali viventi* ("Observations on Living Animals Which Are to Be Found Within Other Living Animals"), in which he sought to disprove the idea of spontaneous generation. Everything must have a parent, he wrote; only life produces life. A 19th-century pioneer in this field, regarded by some as founder of the parasitic theory of infection, was Agostino Bassi of Italy, who showed that a disease of silkworms was caused by a fungus that could be destroyed by chemical agents.

The main credit for establishing the science of bacteriology must be accorded to the French chemist Louis Pasteur. It was Pasteur who, by a brilliant series of experiments, proved that the fermentation of wine and the souring of milk are caused by living microorganisms. His work led to the pasteurization of milk and solved problems of agriculture and industry as well as those of animal and human diseases. He successfully employed inoculations to prevent anthrax in sheep and cattle, chicken cholera in fowl, and finally rabies in humans and dogs. The latter resulted in the widespread establishment of Pasteur institutes.

Influence of Pasteur

From Pasteur, Joseph Lister derived the concepts that enabled him to introduce the antiseptic principle into surgery. In 1865 Lister, a professor of surgery at Glasgow University, began placing an antiseptic barrier of carbolic acid between the wound and the germ-containing atmosphere. Infections and deaths fell dramatically, and his pioneering work led to more refined techniques of sterilizing the surgical environment.

Obstetrics had already been robbed of some of its terrors by Alexander Gordon at Aberdeen, Scot., Oliver Wendell Holmes at Boston, and Ignaz Semmelweis at Vienna and Pest (Budapest), who advocated disinfection of the hands and clothing of midwives and medical students who attended confinements. These measures produced a marked reduction in cases of puerperal fever, the bacterial scourge of women following childbirth.

Another pioneer in bacteriology was the German physician Robert Koch, who showed how bacteria could be cultivated, isolated, and examined in the laboratory. A meticulous investigator, Koch discovered the organisms of tuberculosis, in 1882, and cholera, in 1883. By the end of the century many other disease-producing microorganisms had been identified.

DISCOVERIES IN CLINICAL MEDICINE AND ANESTHESIA

There was perhaps some danger that in the search for bacteria other causes of disease would escape detection. Many physicians, however, were working along different lines in the 19th century. Among them were a group attached to Guy's Hospital, in London: Richard Bright, Thomas Addison, and Sir William Gull. Bright contributed significantly to the knowledge of kidney diseases, including Bright's disease, and Addison gave his name to disorders of the adrenal glands and the blood. Gull, a famous clinical teacher, left a legacy of pithy aphorisms that might well rank with those of Hippocrates.

The 19th century

In Dublin Robert Graves and William Stokes introduced new methods in clinical diagnosis and medical training; while in Paris a leading clinician, Pierre-Charles-Alexandre Louis, was attracting many students from America by the excellence of his teaching. By the early 19th century the United States was ready to send back the results of its own researches and breakthroughs. In 1809, in a small Kentucky town, Ephraim McDowell boldly operated on a woman—without anesthesia or antisepsis—and successfully removed a large ovarian tumour. William Beaumont, in treating a shotgun wound of the stomach, was led to make many original observations that were published in 1833 as *Experiments and Observations on the Gastric Juice and the Physiology of Digestion.*

The most famous contribution by the United States to medical progress at this period was undoubtedly the introduction of general anesthesia, a procedure that not only liberated the patient from the fearful pain of surgery but also enabled the surgeon to perform more extensive operations. The discovery was marred by controversy. Crawford Long, Gardner Colton, Horace Wells, and Charles Jackson are all claimants for priority; some used nitrous oxide gas, and others employed ether, which was less capricious. There is little doubt, however, that it was William Thomas Morton who, on Oct. 16, 1846, at Massachusetts General Hospital, in Boston, first demonstrated before a gathering of physicians the use of ether as a general anesthetic. The news quickly reached Europe, and general anesthesia soon became prevalent in surgery. At Edinburgh, the professor of midwifery, James Young Simpson, had been experimenting upon himself and his assistants, inhaling various vapours with the object of discovering an effective anesthetic. In November 1847 chloroform was tried with complete success, and soon it was preferred to ether and became the anesthetic of choice.

ADVANCES AT THE END OF THE CENTURY

While antisepsis and anesthesia placed surgery on an entirely new footing, similarly important work was carried out in other fields of study, such as parasitology and disease transmission. Patrick Manson, a British pioneer in tropical medicine, showed in China, in 1877, how insects can carry disease and how the embryos of the *Filaria*

Insect transmission of disease

worm, which can cause elephantiasis, are transmitted by the mosquito. Manson explained his views to a British army surgeon, Ronald Ross, then working on the problem of malaria, and Ross discovered the malarial parasite in the stomach of the *Anopheles* mosquito in 1897.

In Cuba, Carlos Finlay expressed the view, in 1881, that yellow fever is carried by the *Stegomyia* mosquito. Following his lead, the Americans Walter Reed, William Gorgas, and others were able to conquer the scourge of yellow fever in Panama and made possible the completion of the Panama Canal by reducing the death rate there from 176 per 1,000 to 6 per 1,000.

Other victories in preventive medicine ensued, because the maintenance of health was now becoming as important a concern as the cure of disease; and the 20th century was to witness the evolution and progress of national health services in a number of countries. In addition, spectacular advances in diagnosis and treatment followed the discovery of X rays by Wilhelm Conrad Röntgen, in 1895, and of radium by Pierre and Marie Curie in 1898. Before the turn of the century, too, the vast new field of psychiatry had been opened up by Sigmund Freud. The tremendous increase in scientific knowledge during the 19th century radically altered and expanded the practice of medicine. Concern for upholding the quality of services led to the establishment of public and professional bodies to govern the standards for medical training and practice.

(D.J.G./P.Rh.)

Medicine in the 20th century

The 20th century has produced such a plethora of discoveries and advances that in some ways the face of medicine has changed out of all recognition. In 1901, for instance, in the United Kingdom the expectation of life at birth, a primary indicator of the effect of health care on mortality (but also reflecting the state of health education, housing, and nutrition), was 48 years for males and 51.6 years for females. After steady increases, by the 1980s life expectancy had reached 71.4 years for males and 77.2 years for females. Other industrialized nations showed similar dramatic increases. Indeed, the outlook has so altered that, with the exception of diseases such as cancer and AIDS, attention has become focused on morbidity rather than mortality, and the emphasis has changed from keeping people alive to keeping them fit.

The rapid progress of medicine in this era was reinforced by enormous improvements in communication between scientists throughout the world. Through publications, conferences, and—later—computers and electronic media, they freely exchanged ideas and reported on their endeavours. No longer was it common for an individual to work in isolation. Although specialization increased, teamwork became the norm. It consequently has become more difficult to ascribe medical accomplishments to particular individuals.

In the first half of the century, emphasis continued to be placed on combating infection, and notable landmarks were also attained in endocrinology, nutrition, and other areas. In the years following World War II, insights derived from cell biology altered basic concepts of the disease process; new discoveries in biochemistry and physiology opened the way for more precise diagnostic tests and more effective therapies; and spectacular advances in biomedical engineering enabled the physician and surgeon to probe into the structures and functions of the body by noninvasive imaging techniques like ultrasound (sonar), computerized axial tomography (CAT), and nuclear magnetic resonance (NMR). With each new scientific development, medical practices of just a few years earlier became obsolete.

INFECTIOUS DISEASES AND CHEMOTHERAPY

Varieties of microorganisms

In the years following the turn of the century, ongoing research concentrated on the nature of infectious diseases and their means of transmission. Increasing numbers of pathogenic organisms were discovered and classified. Some, such as the rickettsias, which cause diseases like typhus, were smaller than bacteria; some were larger, such as the protozoans that engender malaria and other tropical diseases. The smallest to be identified were the viruses, producers of many diseases, among them mumps, measles, German measles, and poliomyelitis; and in 1910 Peyton Rous showed that a virus could also cause a malignant tumour, a sarcoma in chickens.

There was still little to be done for the victims of most infectious organisms beyond drainage, poultices, and ointments, in the case of local infections, and rest and nourishment for severe diseases. The search for treatments aimed at both vaccines and chemical remedies.

Ehrlich and arsphenamine. Germany was well to the forefront in medical progress. The scientific approach to medicine had been developed there long before it spread to other countries, and postgraduates flocked to German medical schools from all over the world. The opening decade of the 20th century has been well described as the golden age of German medicine. Outstanding among its leaders was Paul Ehrlich.

While still a student, Ehrlich carried out some work on lead poisoning from which he evolved the theory that was to guide much of his subsequent work—that certain tissues have a selective affinity for certain chemicals. He experimented with the effects of various chemical substances on disease organisms. In 1910, with his colleague Sahachiro Hata, he conducted tests on arsphenamine, once sold under the commercial name Salvarsan. Their success inaugurated the chemotherapeutic era, which was to revolutionize the treatment and control of infectious diseases. Salvarsan, a synthetic preparation containing arsenic, is lethal to the microorganism responsible for syphilis. Until the introduction of penicillin, Salvarsan or one of its modifications remained the standard treatment of syphilis and went far toward bringing this social and medical scourge under control.

Sulfonamide drugs. In 1932 the German bacteriologist Gerhard Domagk announced that the red dye Prontosil is active against streptococcal infections in mice and humans. Soon afterward French workers showed that its active antibacterial agent is sulfanilamide. In 1936 the English physician Leonard Colebrook and his colleagues provided overwhelming evidence of the efficacy of both Prontosil and sulfanilamide in streptococcal septicemia (bloodstream infection), thereby ushering in the sulfonamide era. New sulfonamides, which appeared with astonishing rapidity, had greater potency, wider antibacterial range, or lower toxicity. Some stood the test of time; others, like the original sulfanilamide and its immediate successor, sulfapyridine, were replaced by safer and more powerful successors.

Antibiotics. *Penicillin.* A dramatic episode in medical history occurred in 1928, when Alexander Fleming noticed the inhibitory action of a stray mold on a plate culture of staphylococcus bacteria in his laboratory at St. Mary's Hospital, London. Many other bacteriologists must have made the observation, but none had realized the possible implications. The mold was a strain of *Penicillium—P. notatum*—which gave its name to the now-famous drug penicillin. In spite of his conviction that penicillin was a potent antibacterial agent, Fleming was unable to carry his work to fruition, mainly because biochemists at the time were unable to isolate it in sufficient quantities or in a sufficiently pure form to allow its use on patients.

Ten years later Howard Florey, Ernst Chain, and their colleagues at Oxford University took up the problem again They isolated penicillin in a form that was fairly pure (by standards then current) and demonstrated its potency and relative lack of toxicity. By then World War II had begun, and techniques to facilitate commercial production were developed in the United States. By 1944 adequate amounts were available to meet the extraordinary needs of wartime.

Streptomycin, PAS, and isoniazid

Antituberculous drugs. While penicillin is the most useful and the safest antibiotic, it suffers from certain disadvantages. The most important of these is that it is not active against *Mycobacterium tuberculosis,* the bacillus of tuberculosis. In view of the importance of tuberculosis as a public health hazard, this is a serious defect. The position was rapidly rectified when, in 1944, Selman A. Waksman and his colleagues announced the discovery of

streptomycin from cultures of a soil organism, *Streptomyces griseus,* and stated that it was active against *M. tuberculosis.* Subsequent clinical trials amply confirmed this claim. Streptomycin suffers, however, from the great disadvantage that the tubercle bacillus tends to become resistant to it. Fortunately, other drugs became available to supplement it, the two most important being para-aminosalicylic acid (or PAS) and isoniazid. With a combination of two or more of these preparations, the outlook in tuberculosis improved immeasurably. The disease was not conquered, but it was brought well under control.

Other antibiotics. Penicillin is not effective over the entire field of microorganisms pathogenic to humans. During the 1950s the search for antibiotics to fill this gap resulted in a steady stream of them, some with a much wider antibacterial range than penicillin (the so-called broad-spectrum antibiotics) and some capable of coping with those microorganisms that are inherently resistant to penicillin or that have developed resistance through exposure to penicillin.

This tendency of microorganisms to develop resistance to penicillin at one time threatened to become almost as serious a problem as the development of resistance to streptomycin by the bacillus of tuberculosis. Fortunately, early appreciation of the problem by clinicians resulted in more discriminate use of penicillin. Scientists continued to look for means of obtaining new varieties of penicillin, and their researches produced the so-called semisynthetic antibiotics, some of which are active when taken by mouth, while others are effective against microorganisms that have developed resistance to the earlier form of penicillin.

IMMUNOLOGY

Dramatic though they undoubtedly were, the advances in chemotherapy still left one important area vulnerable, that of the viruses. It was in bringing viruses under control that advances in immunology—the study of immunity—played such a striking part. One of the paradoxes of medicine is that the first large-scale immunization against a viral disease was instituted and established long before viruses were discovered. When Edward Jenner introduced vaccination against the virus that causes smallpox, the identification of viruses was still 100 years in the future. It took almost another half century to discover an effective method of producing antiviral vaccines that were both safe and effective.

In the meantime, however, the process by which the body reacts against infectious organisms to generate immunity became better understood. In Paris, Élie Metchnikoff had already detected the role of white blood cells in the immune reaction, and Jules Bordet had identified antibodies in the blood serum. The mechanisms of antibody activity were used to devise diagnostic tests for a number of diseases. In 1906 August von Wassermann gave his name to the blood test for syphilis, and in 1908 the tuberculin test—the skin test for tuberculosis—came into use. At the same time, methods of producing effective substances for inoculation were improved, and immunization against bacterial diseases made rapid progress.

Wassermann and the tuberculin test

Antibacterial vaccination. *Typhoid.* In 1897 the English bacteriologist Almroth Wright introduced a vaccine prepared from killed typhoid bacilli as a preventive of typhoid. Preliminary trials in the Indian army produced excellent results, and typhoid vaccination was adopted for the use of British troops serving in the South African War. Unfortunately, the method of administration was inadequately controlled, and the government sanctioned inoculations only for soldiers that "voluntarily presented themselves for this purpose prior to their embarkation for the seat of war." The result was that, according to the official records, only 14,626 men volunteered out of a total strength of 328,244 who served during the three years of the war. Although later analysis showed that inoculation had had a beneficial effect, there were 57,684 cases of typhoid—approximately one in six of the British troops engaged—with 9,022 deaths.

A bitter controversy over the merits of the vaccine followed, but before the outbreak of World War I immunization had been officially adopted by the army. Comparative statistics would seem to provide striking confirmation of the value of antityphoid inoculation, even allowing for the better sanitary arrangements in the latter war. In the South African War the annual incidence of enteric infections (typhoid and paratyphoid) was 105 per 1,000 troops, and the annual death rate was 14.6 per 1,000; the comparable figures for World War I were 2.35 and 0.139, respectively.

It is perhaps a sign of the increasingly critical outlook that developed in medicine in the post-1945 era that experts continued to differ on some aspects of typhoid immunization. There was no question as to its fundamental efficacy, but there was considerable variation of opinion as to the best vaccine to use and the most effective way of administering it. Moreover, it was often difficult to decide to what extent the decline in typhoid was attributable to improved sanitary conditions and what to the greater use of the vaccine.

Tetanus. The other great hazard of war that was brought under control in World War I was tetanus. This was achieved by the prophylactic injection of tetanus antitoxin into all wounded men. The serum was originally prepared by the bacteriologists Emil von Behring and Shibasaburo Kitasato in 1890–92, and the results of this first large-scale trial amply confirmed its efficacy. (Tetanus antitoxin is a sterile solution of antibody globulins—a type of blood protein—from immunized horses or cattle.)

It was not until the 1930s, however, that an efficient vaccine, or toxoid, as it is known in the cases of tetanus and diphtheria, was produced against tetanus. (Tetanus toxoid is a preparation of the toxin—or poison—produced by the microorganism; injected into humans, it stimulates the body's own defenses against the disease, thus bringing about immunity.) Again, a war was to provide the opportunity for testing on a large scale, and experience with tetanus toxoid in World War II indicated that it gave a high degree of protection.

Diphtheria. The story of diphtheria is comparable to that of tetanus, though even more dramatic. First, as with tetanus antitoxin, came the preparation of diphtheria antitoxin by Behring and Kitasato in 1890. As the antitoxin came into general use for the treatment of cases, the death rate began to decline. There was no significant fall in the number of cases, however, until a toxin–antitoxin mixture, introduced by Behring in 1913, was used to immunize children. A more effective toxoid was introduced by the French bacteriologist Gaston Ramon in 1923, and with subsequent improvements this became one of the most effective vaccines available in medicine. Where mass immunization of children with the toxoid was practiced, as in the United States and Canada beginning in the late 1930s and in England and Wales in the early 1940s, cases of diphtheria and deaths from the disease became almost nonexistent. In England and Wales, for instance, the number of deaths fell from an annual average of 1,830 in 1940–44 to zero in 1969. Administration of a combined vaccine against diphtheria, pertussis (whooping cough), and tetanus (DPT) is recommended for young children. Although an increasing number of dangerous side effects from the DPT vaccine have been reported, it continues to be used in most countries because of the protection it affords.

Mass diphtheria immunization of children

BCG vaccine for tuberculosis. If, as is universally accepted, prevention is better than cure, immunization is the ideal way of dealing with diseases caused by microorganisms. An effective, safe vaccine protects the individual from disease, whereas chemotherapy merely copes with the infection once the individual has been affected. In spite of its undoubted value, however, immunization has been a recurring source of dispute. Like vaccination against typhoid (and against poliomyelitis later), tuberculosis immunization evoked widespread contention.

In 1908 Albert Calmette, a pupil of Pasteur, and Camille Guérin produced an avirulent (weakened) strain of the tubercle bacillus. About 13 years later, vaccination of children against tuberculosis was introduced, with a vaccine made from this avirulent strain and known as BCG (bacillus Calmette-Guérin) vaccine. Although it was adopted in France, Scandinavia, and elsewhere, British and U.S. authorities frowned upon its use on the grounds that it was

not safe and that the statistical evidence in its favour was not convincing.

One of the stumbling blocks in the way of its widespread adoption was what came to be known as the Lübeck disaster. In the spring of 1930, 249 infants were vaccinated with BCG vaccine in Lübeck, Ger.; by autumn, 73 of the 249 were dead. Criminal proceedings were instituted against those responsible for giving the vaccine. The final verdict was that the vaccine had been contaminated, and the BCG vaccine itself was exonerated from any responsibility for the deaths. A bitter controversy followed, but in the end the protagonists of the vaccine won when a further trial showed that the vaccine was safe and that it protected four out of five of those vaccinated.

Immunization against viral diseases. With the exception of smallpox, it was not until well into the 20th century that efficient viral vaccines became available. In fact, it was not until the 1930s that much began to be known about viruses. The two developments that contributed most to the rapid growth in knowledge after that time were the introduction of tissue culture as a means of growing viruses in the laboratory and the availability of the electron microscope. Once the virus could be cultivated with comparative ease in the laboratory, the research worker could study it with care and evolve methods for producing one of the two requirements for a safe and effective vaccine: either a virus that was so attenuated, or weakened, that it could not produce the disease for which it was responsible in its normally virulent form; or a killed virus that retained the faculty of inducing a protective antibody response in the vaccinated individual.

Poliomyelitis vaccine

The first of the viral vaccines to result from these advances was for yellow fever, developed by the microbiologist Max Theiler in the late 1930s. About 1945 the first relatively effective vaccine was produced for influenza; in 1954 the American physician Jonas E. Salk introduced a vaccine for poliomyelitis; and in 1960 an oral poliomyelitis vaccine, developed by the virologist Albert B. Sabin, came into wide use.

These vaccines went far toward bringing under control three of the major diseases of the time although, in the case of influenza, a major complication is the disturbing proclivity of the virus to change its character from one epidemic to another. Even so, sufficient progress has been made to ensure that a pandemic like the one that swept the world in 1918–19, killing more than 15,000,000 people, is unlikely to occur again. Centres are now equipped to monitor outbreaks of influenza throughout the world in order to establish the identity of the responsible viruses and, if necessary, take steps to produce appropriate vaccines.

During the 1960s effective vaccines came into use for measles and rubella (German measles). Both evoked a certain amount of controversy. In the case of measles in the Western world it was contended that, if acquired in childhood, it is not a particularly hazardous malady, and the naturally acquired disease evokes permanent immunity in the vast majority of cases. Conversely, the vaccine induces a certain number of adverse reactions, and the duration of the immunity it produces is problematical. In the end the official view was that universal measles vaccination is to be commended.

The situation with rubella vaccination was different. This is a fundamentally mild affliction, and the only cause for anxiety is its proclivity to induce congenital deformities if a pregnant woman should acquire the disease. Once an effective vaccine was available, the problem was the extent to which it should be used. Ultimately the consensus was reached that all girls who had not already had the disease should be vaccinated at about 12 years. In the United States children are routinely immunized against measles, mumps, and rubella at the age of 15 months.

The immune response. With advances in cell biology in the second half of the 20th century came a more profound understanding of both normal and abnormal conditions in the body. Electron microscopy enabled observers to peer more deeply into the structures of the cell, and chemical investigations revealed clues to their functions in the cell's intricate metabolism. The overriding importance of the nuclear genetic material DNA (deoxyribonucleic acid) in regulating the cell's protein and enzyme production lines became evident. A clearer comprehension also emerged of the ways in which the cells of the body defend themselves by modifying their chemical activities to produce antibodies against injurious agents.

Understanding the cell

Up until the turn of the century, immunity referred mostly to the means of resistance of an animal to invasion by a parasite or microorganism. Around mid-century there arose a growing realization that immunity and immunology cover a much wider field and are concerned with mechanisms for preserving the integrity of the individual. The introduction of organ transplantation, with its dreaded complication of tissue rejection, brought this broader concept of immunology to the fore.

At the same time, research workers and clinicians began to appreciate the far-reaching implications of immunity in relation to endocrinology, genetics, tumour biology, and the biology of a number of other maladies. The so-called autoimmune diseases are caused by an aberrant series of immune responses by which the body's own cells are attacked. Suspicion is growing that a number of major disorders such as diabetes, rheumatoid arthritis, and multiple sclerosis may be caused by similar mechanisms.

In some conditions viruses invade the genetic material of cells and distort their metabolic processes. Such viruses may lie dormant for many years before becoming active. This may be the underlying cause of many cancers, in which cells escape from the usual constraints imposed upon them by the normal body. The dreaded affliction of acquired immune deficiency syndrome (AIDS) is caused by a virus that has a long dormant period and then attacks the cells that produce antibodies. The result is that the affected person is not able to generate an immune response to infections or malignancies.

AIDS

ENDOCRINOLOGY

At the beginning of the 20th century, endocrinology was in its infancy. Indeed, it was not until 1905 that Ernest H. Starling, one of the many brilliant pupils of Edward Sharpey-Schafer, the dean of British physiology during the early decades of the century, introduced the term hormone for the internal secretions of the endocrine glands. In 1891 the English physician George Redmayne Murray achieved the first success in treating myxedema (the common form of hypothyroidism) with an extract of the thyroid gland. Three years later, Sharpey-Schafer and George Oliver demonstrated in extracts of the adrenal glands a substance that raised the blood pressure; and in 1901 Jokichi Takamine, a Japanese chemist working in the United States, isolated this active principle, known as epinephrine or adrenaline.

Insulin. During the first two decades of the century, steady progress was made in the isolation, identification, and study of the active principles of the various endocrine glands, but the outstanding event of the early years was the discovery of insulin by Frederick Banting, Charles H. Best, and J.J.R. Macleod in 1921. Almost overnight the lot of the diabetic patient changed from a sentence of almost certain death to a prospect not only of survival but of a long and healthy life.

For more than 30 years, some of the greatest minds in physiology had been seeking the cause of diabetes mellitus. In 1889 the German physicians Joseph von Mering and Oskar Minkowski had shown that removal of the pancreas in dogs produced the disease. In 1901 the American pathologist Eugene L. Opie described degenerative changes in the clumps of cells in the pancreas known as the islets of Langerhans, thus confirming the association between failure in the function of these cells and diabetes. Sharpey-Schafer concluded that the islets of Langerhans secrete a substance that controls the metabolism of carbohydrate. Then Banting, Best, and Macleod, working at the University of Toronto, succeeded in isolating the elusive hormone and gave it the name insulin.

Insulin was available in a variety of forms, but synthesis on a commercial scale was not achieved, and the only source of the hormone was the pancreas of animals. One of its practical disadvantages is that it has to be given by injection; consequently an intense search was conducted

for some alternative substance that would be active when taken by mouth. Various preparations—oral hypoglycemic agents, as they are known—appeared that were effective to a certain extent in controlling diabetes, but evidence indicated that these were only of value in relatively mild cases of the disease. For the person with advanced diabetes, a normal, healthy life remained dependent upon the continuing use of insulin injections.

Cortisone. Another major advance in endocrinology came from the Mayo Clinic, in Rochester, Minn. In 1949 Philip S. Hench and his colleagues announced that a substance isolated from the cortex of the adrenal gland had a dramatic effect upon rheumatoid arthritis. This was compound E, or cortisone, as it came to be known, which had been isolated by Edward C. Kendall in 1935. Cortisone and its many derivatives proved to be potent as anti-inflammatory agents. Although it is not a cure for rheumatoid arthritis, as a temporary measure cortisone can often control the acute exacerbation caused by the disease and can provide relief in other conditions, such as acute rheumatic fever, certain kidney diseases, certain serious diseases of the skin, and some allergic conditions, including acute exacerbations of asthma. Of even more long-term importance is the valuable role it has as a research tool.

Sex hormones. Not the least of the advances in endocrinology was the increasing knowledge and understanding of the sex hormones. This culminated in the application of this knowledge to the problem of birth control. After an initial stage of hesitancy, the contraceptive pill, with its basic rationale of preventing ovulation, was accepted by the vast majority of family-planning organizations and many gynecologists as the most satisfactory method of contraception. Its risks, practical and theoretical, introduced a note of caution, but this was not sufficient to detract from the wide appeal induced by its effectiveness and ease of use.

The contraceptive pill

VITAMINS

In the field of nutrition, the outstanding advance of the 20th century was the discovery and the appreciation of the importance to health of the "accessory food factors," or vitamins. Various workers had shown that animals did not thrive on a synthetic diet containing all the correct amounts of protein, fat, and carbohydrate; they even suggested that there must be some unknown ingredients in natural food that were essential for growth and the maintenance of health. But little progress was made in this field until the classical experiments of the English biologist F. Gowland Hopkins were published in 1912. These were so conclusive that there could be no doubt that what he termed "accessory substances" were essential for health and growth.

The name vitamine was suggested for these substances by the biochemist Casimir Funk in the belief that they were amines, certain compounds derived from ammonia. In due course, when it was realized that they were not amines, the term was altered to vitamin.

Once the concept of vitamins was established on a firm scientific basis it was not long before their identity began to be revealed. Soon there was a long series of vitamins, best known by the letters of the alphabet after which they were originally named when their chemical identity was still unknown. By supplementing the diet with foods containing particular vitamins, deficiency diseases such as rickets (due to deficiency of vitamin D) and scurvy (due to lack of vitamin C, or ascorbic acid) practically disappeared from Western countries, while deficiency diseases such as beriberi (caused by lack of vitamin B_1, or thiamine), which were endemic in Eastern countries, either disappeared or could be remedied with the greatest of ease.

The isolation of vitamin B_{12}, or cyanocobalamin, was of particular interest because it almost rounded off the fascinating story of how pernicious anemia was brought under control. Throughout the first two decades of the century, the diagnosis of pernicious anemia, like that of diabetes mellitus, was nearly equivalent to a death sentence. Unlike the more common form of so-called secondary anemia, it did not respond to the administration of suitable iron salts, and no other form of treatment touched it; hence, the grimly appropriate title of pernicious anemia.

In the early 1920s, George R. Minot, one of the many brilliant investigators that Harvard University has contributed to medical research, became interested in work being done by the American pathologist George H. Whipple on the beneficial effects of raw beef liver in severe experimental anemia. With a Harvard colleague, William P. Murphy, he decided to investigate the effect of raw liver in patients with pernicious anemia, and in 1926 they were able to announce that this form of therapy was successful. The validity of their findings was amply confirmed, and the fear of pernicious anemia came to an end.

As so often happens in medicine, many years were to pass before the rationale of liver therapy in pernicious anemia was fully understood. In 1948, however, almost simultaneously in the United States and Britain, the active principle, cyanocobalamin, was isolated from liver, and this vitamin became the standard treatment for pernicious anemia.

Liver therapy in pernicious anemia

MALIGNANT DISEASE

While progress was the hallmark of medicine after the beginning of the 20th century, there is one field in which a gloomier picture must be painted, that of malignant disease, or cancer. It is the second most common cause of death in most Western countries in the second half of the 20th century, being exceeded only by deaths from heart disease. Some progress, however, has been achieved. The causes of the various types of malignancies are not known, but many more methods are available for attacking the problem; surgery remains the principal therapeutic standby, but radiotherapy and chemotherapy are increasingly used.

Soon after the discovery of radium was announced, in 1898, its potentialities in treating cancer were realized; in due course it assumed an important role in therapy. Simultaneously, deep X-ray therapy was developed, and with the atomic age came the use of radioactive isotopes. (A radioactive isotope is an unstable variant of a substance that has a stable form; during the process of breaking down, the unstable form emits radiation.) High-voltage X-ray therapy and radioactive isotopes have largely replaced radium. Whereas irradiation long depended upon X rays generated at 250 kilovolts, machines that are capable of producing X rays generated at 8,000 kilovolts and betatrons of up to 22,000,000 electron volts (MeV) have come into clinical use.

The most effective of the isotopes is radioactive cobalt. Telecobalt machines (those that hold the cobalt at a distance from the body) are available containing 2,000 curies or more of the isotope, an amount equivalent to 3,000 grams of radium and sending out a beam equivalent to that from a 3,000-kilovolt X-ray machine.

Of even more significance have been the developments in the chemotherapy of cancer. Nothing remotely resembling a chemotherapeutic cure has been achieved, but in certain forms of malignant disease, such as leukemia, which cannot be treated by surgery, palliative effects have been achieved that prolong life and allow the patient in many instances to lead a comparatively normal existence.

Chemotherapy for cancer

Fundamentally, however, perhaps the most important advance of all in this field has been the increasing appreciation of the importance of prevention. The discovery of the relationship between cigarette smoking and lung cancer is the classic example. Less publicized, but of equal import, is the continuing supervision of new techniques in industry and food manufacture in an attempt to ensure that they do not involve the use of cancer-causing substances.

TROPICAL MEDICINE

The first half of the 20th century witnessed the virtual conquest of three of the major diseases of the tropics: malaria, yellow fever, and leprosy. At the turn of the century, as for the preceding two centuries, quinine was the only known drug to have any appreciable effect on malaria. With the increasing development of tropical countries and rising standards of public health, it became obvious that quinine was not completely satisfactory. Intensive research

between World Wars I and II indicated that several synthetic compounds were more effective. The first of these to become available, in 1934, was quinacrine (known as mepacrine, Atabrine, or Atebrin). In World War II it amply fulfilled the highest expectations and helped to reduce disease among Allied troops in Africa, Southeast Asia, and the Far East. A number of other effective antimalarial drugs subsequently became available.

An even brighter prospect—the virtual eradication of malaria—was opened up by the introduction, during World War II, of the insecticide DDT (1,1,1-trichloro-2,2,--bis[*p*-chlorophenyl]ethane, or dichlorodiphenyltrichloroethane). It had long been realized that the only effective way of controlling malaria was to eradicate the anopheline mosquitoes that transmit the disease. Older methods of mosquito control, however, were cumbersome and expensive. The lethal effect of DDT on the mosquito, its relative cheapness, and its ease of use on a widespread scale provided the answer. An intensive worldwide campaign, sponsored by the World Health Organization, was planned and went far toward bringing malaria under control.

The major problem encountered with respect to effectiveness was that the mosquitoes were able to develop a resistance to DDT; but the introduction of other insecticides, such as dieldrin and lindane (BHC), helped to overcome this difficulty. In recent years the use of these and other insecticides has been strongly criticized by ecologists, however.

Yellow fever is another mosquito-transmitted disease, and the prophylactic value of modern insecticides in its control was almost as great as in the case of malaria. The forest reservoirs of the virus present a more difficult problem, but the combined use of immunization and insecticides did much to bring this disease under control.

Until the 1940s the only drugs available for treating leprosy were the chaulmoogra oils and their derivatives. These, though helpful, were far from satisfactory. In the 1940s the group of drugs known as the sulfones appeared, and it soon became apparent that they were infinitely better than any other group of drugs in the treatment of leprosy. Several other drugs later proved promising. Although there is as yet no known cure—in the strict sense of the term—for leprosy, the outlook has so changed that there are good grounds for believing that this age-old scourge can be brought under control and the victims of the disease saved from those dreaded mutilations that have given leprosy such a fearsome reputation throughout the ages. (Wm.A.R.T./P.Rh.)

Control of leprosy

Surgery in the 20th century

THE OPENING PHASE

Three seemingly insuperable obstacles beset the surgeon in the years before the mid-19th century: pain, infection, and shock. Once these were overcome, the surgeon believed that he could burst the bonds of centuries and become the master of his craft. There is more, however, to anesthesia than putting the patient to sleep. Infection, despite first antisepsis (destruction of microorganisms present) and later asepsis (avoidance of contamination), is still an ever-present menace; and shock continues to perplex physicians. But in the 20th century, surgery has progressed farther, faster, and more dramatically than in all preceding ages.

The situation encountered. The shape of surgery that entered the new century was clearly recognizable as the forerunner of today's, blurred and hazy though the outlines may now seem. The operating theatre still retained an aura of the past, when the surgeon played to his audience and the patient was little more than a stage prop. In most hospitals it was a high room lit by a skylight, with tiers of benches rising above the narrow, wooden operating table. The instruments, kept in glazed or wooden cupboards around the walls, were of forged steel, unplated, and with handles of wood or ivory.

The means to combat infection hovered between antisepsis and asepsis. Instruments and dressings were mostly sterilized by soaking them in dilute carbolic acid (or other antiseptic), and the surgeon often endured a gown freshly wrung out in the same solution. Asepsis gained ground fast, however. It had been born in the Berlin clinic of Ernst von Bergmann where, in 1886, steam sterilization had been introduced. Gradually, this led to the complete aseptic ritual, which has as its basis the bacterial cleanliness (as opposed to social cleanliness) of everything that comes in contact with the wound. Hermann Kümmell, of Hamburg, devised the routine of "scrubbing up." In 1890 William Stewart Halsted, of Johns Hopkins University, had rubber gloves specially made for operating, and in 1896 Johannes von Mikulicz-Radecki, a Pole working at Breslau, Ger., invented the gauze mask.

Many surgeons, brought up in a confused misunderstanding of the antiseptic principle—believing that carbolic would cover a multitude of sins, many of which they were ignorant of committing—failed to grasp what asepsis was all about. Thomas Annandale, for example, blew through his catheters to make sure that they were clear, and many an instrument, dropped accidentally, was simply given a quick wipe and returned to use. Tradition died hard, and asepsis had an uphill struggle before it was fully accepted. "I believe firmly that more patients have died from the use of gloves than have ever been saved from infection by their use," wrote W.P. Carr, an American, in 1911. Over the years, however, a sound technique was evolved as the foundation for the growth of modern surgery.

Anesthesia, at the turn of the century, progressed slowly. Few physicians made a career of the subject, and frequently the patient was rendered unconscious by a student, a nurse, or a porter wielding a rag and bottle. Chloroform was overwhelmingly more popular than ether, on account of its ease of administration, despite the fact that it was liable to kill by stopping the heart.

Anesthesia at the turn of the century

Although by the end of the first decade, nitrous oxide (laughing gas) combined with ether had displaced—but by no means entirely—the use of chloroform, the surgical problems were far from ended. For years to come the abdominal surgeon besought the anesthetist to deepen the level of anesthesia and thus relax the abdominal muscles; the anesthetist responded to the best of his ability, acutely aware that the deeper he went, the closer the patient was to death. When other anesthetic agents were discovered, the anesthetist came into his own, and many advances in spheres such as brain and heart surgery would have been impossible without his skill.

The third obstacle, shock, is perhaps the most complex and the most difficult to define satisfactorily. The only major cause properly appreciated at the start of the 20th century was loss of blood, and once that had occurred nothing, in those days, could be done. And so, the study of shock—its causes, its effects on human physiology, and its prevention and treatment—became all-important to the progress of surgery.

In the latter part of the 19th century, then, surgeons had been liberated from the age-old bogies of pain, pus, and hospital gangrene. Hitherto, operations had been restricted to amputations, cutting for stone in the bladder, tying off arterial aneurysms (bulging and thinning of artery walls), repairing hernias, and a variety of procedures that could be done without going too deeply beneath the skin. But the anatomical knowledge, a crude skill derived from practice on dead bodies, and above all the enthusiasm, were there waiting. Largely ignoring the mass of problems they uncovered, surgeons launched forth into an exploration of the human body.

They acquired a reputation for showmanship; but much of their surgery, though speedy and spectacular, was rough and ready. There were a few who developed supreme skill and dexterity and could have undertaken a modern operation with but little practice; indeed, some devised the very operations still in use today. One such was Theodor Billroth, head of the surgical clinic at Vienna, who collected a formidable list of successful "first" operations. He represented the best of his generation—a surgical genius, an accomplished musician, and a kind, gentle man who brought the breath of humanity to his work. Moreover, the men he trained, including von Mikulicz, Vincenz Czerny, and Anton von Eiselsberg, consolidated the brilliant start that he had given to abdominal surgery in Europe.

Billroth's influence

Changes before World War I. The opening decade of the 20th century was a period of transition. Flamboyant exhibitionism was falling from favour as surgeons, through experience, learned the merits of painstaking, conscientious operation—treating the tissues gently and carefully controlling every bleeding point. The individualist was not submerged, however, and for many years the development of the various branches of surgery rested on the shoulders of a few clearly identifiable men. Teamwork on a large scale arrived only after World War II. The surgeon, at first, was undisputed master in his own wards and theatre. But as time went on and he found he could not solve his problems alone, he called for help from specialists in other fields of medicine and, even more significantly, from his colleagues in other scientific disciplines.

The increasing scope of surgery led to specialization. Admittedly, most general surgeons had a special interest, and for a long time there had been an element of specialization in such fields as ophthalmology, orthopedics, obstetrics, and gynecology; but before long it became apparent that, to achieve progress in certain areas, surgeons had to concentrate their attention on that particular subject.

Abdominal surgery. By the start of the 20th century, abdominal surgery, which provided the general surgeon with the bulk of his work, had grown beyond infancy, thanks largely to Billroth. In 1881 he had performed the first successful removal of part of the stomach for cancer. His next two cases were failures, and he was stoned in the streets of Vienna. Yet, he persisted and by 1891 had carried out 41 more of these operations with 16 deaths—a remarkable achievement for that era.

Peptic ulcers (gastric and duodenal) appeared on the surgical scene (perhaps as a new disease, but more probably because they had not been diagnosed previously), and in 1881 Ludwig Rydygier cured a young woman of her gastric ulcer by removing it. Bypass operations—gastroenterostomies—soon became more popular, however, and enjoyed a vogue that lasted into the 1930s, even though fresh ulcers at the site of the juncture were not uncommon.

The other end of the alimentary tract was also subjected to surgical intervention; cancers were removed from the large bowel and rectum with mortality rates that gradually fell from 80 to 60 to 20 to 12 percent as the surgeons developed their skill. In 1908 the British surgeon Ernest Miles carried out the first abdominoperineal resection for cancer of the rectum; that is, the cancer was attacked both from the abdomen and from below through the perineum (the area between the anus and the genitals), either by one surgeon, who actually did two operations, or by two working together. This technique formed the basis for all future developments.

Appendectomy

Much of the new surgery in the abdomen was for cancer, but not all. Appendectomy became the accepted treatment for appendicitis (in appropriate cases) in the United States before the close of the 19th century; but in Great Britain surgeons were reluctant to remove the organ until 1902, when King Edward VII's coronation was dramatically postponed on account of his appendicitis. The publicity attached to his operation caused the disease and its surgical treatment to become fashionable—despite the fact that the royal appendix remained in the King's abdomen; the surgeon, Frederic Treves, had merely drained the abscess.

Neurosurgery. Though probably the most demanding of all the surgical specialties, neurosurgery was nevertheless one of the first to emerge. The techniques and principles of general surgery were inadequate for work in such a delicate field. William Macewen, a Scottish general surgeon of outstanding versatility, and Victor Alexander Haden Horsley, the first British neurosurgeon, showed that the surgeon had much to offer in the treatment of disease of the brain and spinal cord. Macewen, in 1893, recorded 19 patients operated on for brain abscess, 18 of whom were cured; at that time most other surgeons had 100 percent mortality rates for the condition. His achievement remained unequaled until the discovery of penicillin.

An American, Harvey Williams Cushing, almost by himself consolidated neurosurgery as a specialty. From 1905 on, he advanced neurosurgery through a series of operations and through his writings. Tumours, epilepsy, trigeminal neuralgia, and pituitary disorders were among the conditions he treated successfully.

Radiology. In 1895 a development at the University of Würzburg had far-reaching effects on medicine and surgery, opening up an entirely fresh field of the diagnosis and study of disease and leading to a new form of treatment, radiation therapy. This was the discovery of X rays by Wilhelm Conrad Röntgen, a professor of physics. Within months of the discovery there was an extensive literature on the subject: Robert Jones, a British surgeon, had localized a bullet in a boy's wrist before operating; stones in the urinary bladder and gallbladder had been demonstrated; and fractures had been displayed.

X rays in research

Experiments began on introducing substances that are opaque to X rays into the body to reveal organs and formations, both normal and abnormal. Walter Cannon, a Boston physiologist, used X rays in 1898 in his studies of the alimentary tract. Friedrich Voelcker, of Heidelberg, devised retrograde pyelography (introduction of the radiopaque medium into the kidney pelvis by way of the ureter) for the study of the urinary tract in 1905; in Paris in 1921, Jean Sicard X-rayed the spinal canal with the help of an oily iodine substance, and the next year he did the same for the bronchial tree; and in 1924 Evarts Graham, of St. Louis, used a radiopaque contrast medium to view the gallbladder. Air was also used to provide contrast; in 1918, at Johns Hopkins, Walter Dandy injected air into the ventricles (liquid-filled cavities) of the brain.

The problems of injecting contrast media into the blood vessels took longer to solve, and it was not until 1927 that António Moniz, of Lisbon, succeeded in obtaining pictures of the arteries of the brain. Eleven years later, George Robb and Israel Steinberg of New York overcame some of the difficulties of cardiac catheterization (introduction of a small tube into the heart by way of veins or arteries) and were able to visualize the chambers of the heart on X-ray film. After much research, a further refinement came in 1962, when Frank Sones and Earl K. Shirey of Cleveland showed how to introduce the contrast medium into the coronary arteries.

WORLD WAR I

The battlefields of the 20th century stimulated the progress of surgery and taught the surgeon innumerable lessons, which were subsequently applied in civilian practice. Regrettably, though, the principles of military surgery and casualty evacuation, which can be traced back to the Napoleonic wars, had to be learned over again.

World War I broke, quite dramatically, the existing surgical hierarchy and rule of tradition. No longer did the European surgeon have to waste his best years in apprenticeship before seating himself in his master's chair. Suddenly, young surgeons in the armed forces began confronting problems that would have daunted their elders. Furthermore, their training had been in "clean" surgery performed under aseptic conditions. Now they found themselves faced with the need to treat large numbers of grossly contaminated wounds in improvised theatres. They rediscovered debridement (the surgical excision of dead and dying tissue and the removal of foreign matter).

Treatment of contaminated wounds

The older surgeons cried "back to Lister," but antiseptics, no matter how strong, were no match for putrefaction and gangrene. One method of antiseptic irrigation—devised by Alexis Carrel and Henry Dakin and called the Carrel–Dakin treatment—was, however, beneficial, but only after the wound had been adequately debrided. The scourges of tetanus and gas gangrene were controlled to a large extent by antitoxin and antiserum injections, yet surgical treatment of the wound remained an essential requirement.

Abdominal casualties fared badly for the first year of the war, because experience in the utterly different circumstances of the South African War had led to a belief that these men were better left alone surgically. Fortunately, the error of continuing with such a policy 15 years later was soon appreciated, and every effort was made to deliver the wounded men to a suitable surgical unit with all speed. Little progress was made with chest wounds beyond opening up the wound even further to drain pus from the pleural cavity between the chest wall and the lungs.

Perhaps the most worthwhile and enduring benefit to flow from World War I was rehabilitation. For almost the first time, surgeons realized that their work did not end with a healed wound. In 1915 Robert Jones set up special facilities for orthopedic patients, and at about the same time Harold Gillies founded British plastic surgery in a hut at Sidcup, Kent. In 1917 Gillies popularized the pedicle type of skin graft (the type of graft in which skin and subcutaneous tissue are left temporarily attached for nourishment to the site from which the graft was taken). Since then plastic surgery has given many techniques and principles to other branches of surgery.

BETWEEN THE WORLD WARS

The years between the two world wars may conveniently be regarded as the time when surgery consolidated its position. A surprising number of surgical firsts and an amazing amount of fundamental research had been achieved even in the late 19th century, but the knowledge and experience could not be converted to practical use because the human body could not survive the onslaught. In the years between World Wars I and II, it was realized that physiology—in its widest sense, including biochemistry and fluid and electrolyte balance—was of major importance along with anatomy, pathology, and surgical technique.

The problem of shock. The first problem to be tackled was shock, which was, in brief, found to be due to a decrease in the effective volume of the circulation. To combat shock, the volume had to be restored, and the obvious substance was blood itself. In 1901 Karl Landsteiner, then in Austria, discovered the ABO blood groups, and in 1914 sodium citrate was added to freshly drawn blood to prevent clotting. Blood was occasionally transfused during World War I, but three-quarters of a pint was considered a large amount. These transfusions were given by directly linking the vein of a donor with that of the recipient. The continuous drip method, in which blood flows from a flask, was introduced by Hugh Marriott and Alan Kekwick at the Middlesex Hospital, London, in 1935.

Blood transfusion

As blood transfusions increased in frequency and volume, blood banks were required. Although it took another world war before these were organized on a large scale, the first tentative steps were taken by Sergey Sergeyevich Yudin, of Moscow, who, in 1933, used cadaver blood, and by Bernard Fantus, of Chicago, who, four years later, used living donors as his source of supply. Saline solution, plasma, artificial plasma expanders, and other solutions are now also used in the appropriate circumstances.

Sometimes after operations (especially abdominal operations) the gut becomes paralyzed. It is distended, and quantities of fluid pour into it, dehydrating the body. In 1932 Owen Wangensteen, at the University of Minnesota, advised decompressing the bowel, and in 1934 two other Americans, Thomas Miller and William Abbott, of Philadelphia, invented apparatus for this purpose, a tube with an inflatable balloon on the end that could be passed into the small intestine. The fluid lost from the tissues was replaced by a continuous intravenous drip of saline solution on the principle described by Rudolph Matas, of New Orleans, in 1924. These techniques dramatically improved abdominal surgery, especially in cases of obstruction, peritonitis (inflammation of the abdominal membranes) and acute emergencies generally, since they made it possible to keep the bowel empty and at rest.

Anesthesia and thoracic surgery. The strides taken in anesthesia from the 1920s onward allowed surgeons much more freedom. Rectal anesthesia had never proved satisfactory, and the first improvement on the combination of nitrous oxide, oxygen, and ether was the introduction of the general anesthetic cyclopropane by Ralph Waters of Madison, Wis., in 1933. Soon afterward, intravenous anesthesia was introduced; John Lundy of the Mayo Clinic brought to a climax a long series of trials by many workers when he used Pentothal (thiopental sodium, a barbiturate) to put a patient peacefully to sleep. Then, in 1942, Harold Griffith and G. Enid Johnson, of Montreal, produced muscular paralysis by the injection of a purified preparation of curare. This was harmless since, by then, the anesthetist was able to control the patient's respiration.

If there was one person who was aided more than any other by the progress in anesthesia, it was the thoracic (chest) surgeon. What had bothered him previously was the collapse of the lung, which occurred whenever the pleural cavity was opened. Since the end of the 19th century, many and ingenious methods had been devised to prevent this from happening. The best known was the negative pressure cabinet of Ernst Sauerbruch, then at Mikulicz' clinic at Breslau; the cabinet was first used clinically in 1904 but was destined soon to become obsolete.

The solution lay in inhalational anesthesia administered under pressure. Indeed, when Théodore Tuffier, in 1891, successfully removed the apex of a lung for tuberculosis, this was the technique that he used; he even added an inflatable cuff around the tube inserted in the trachea to ensure a gas-tight fit. Tuffier was ahead of his time, however, and other surgeons and research workers wandered into confused and complex byways before Ivan Magill and Edgar Rowbotham, working at Gillies' plastic-surgery unit, found their way back to the simplicity of the endotracheal tube and positive pressure. In 1931 Ralph Waters showed that respiration could be controlled either by squeezing the anesthetic bag by hand or by using a small motor.

Preventing lung collapse

These advances allowed thoracic surgery to move into modern times. In the 1920s, operations had been performed mostly for infective conditions and as a last resort. The operations necessarily were unambitious and confined to collapse therapy, including thoracoplasty (removal of ribs), apicolysis (collapse of a lung apex and artificially filling the space), and phrenic crush (which paralyzed the diaphragm on the chosen side); to isolation of the area of lung to be removed by first creating pleural adhesions; and to drainage.

The technical problems of surgery within the chest were daunting until Harold Brunn of San Francisco reported six lobectomies (removals of lung lobes) for bronchiectasis with only one death. (In bronchiectasis one or more bronchi or bronchioles are chronically dilated and inflamed, with copious discharge of mucus mixed with pus.) The secret of Brunn's success was the use of intermittent suction after surgery to keep the cavity free of secretions until the remaining lobes of the lung could expand to fill the space. In 1931 Rudolf Nissen, in Berlin, removed an entire lung from a girl with bronchiectasis. She recovered to prove that the risks were not as bad as had been feared.

Cancer of the lung has become a major disease of the 20th century; perhaps it has genuinely increased, or perhaps modern techniques of diagnosis reveal it more often. As far back as 1913 a Welshman, Hugh Davies, removed a lower lobe for cancer, but a new era began when Evarts Graham removed a whole lung for cancer in 1933. The patient, a doctor, was still alive at the time of Graham's death in 1957.

The thoracic part of the esophagus is particularly difficult to reach, but in 1909 the British surgeon Arthur Evans successfully operated on it for cancer. But results were generally poor until, in 1944, John Garlock, of New York, showed that it is possible to excise the esophagus and to bring the stomach up through the chest and join it to the pharynx. Lengths of colon are also used as grafts to bridge the gap.

WORLD WAR II AND AFTER

Once the principles of military surgery were relearned and applied to modern warfare, instances of death, deformity, and loss of limb were reduced to levels previously unattainable. This was due largely to a thorough reorganization of the surgical services, adapting them to prevailing conditions, so that casualties received the appropriate treatment at the earliest possible moment. Evacuation by air (first used in World War I) helped greatly in this respect. Diagnostic facilities were improved, and progress in anesthesia kept pace with the surgeon's demands. Blood was transfused in adequate—and hitherto unthinkable—quantities, and the blood transfusion service as it is known today came into being.

Surgical specialization and teamwork reached new heights with the creation of units to deal with the special problems of injuries to different parts of the body. But the most

revolutionary change was in the approach to wound infections brought about by the use of sulfonamides and (after 1941) of penicillin. The fact that these drugs could never replace meticulous wound surgery was, however, another lesson learned only in the bitter school of experience.

When the war ended, surgeons returned to civilian life feeling that they were at the start of a completely new, exciting era; and indeed they were, for the intense stimulation of the war years had led to developments in many branches of science that could now be applied to surgery. Nevertheless, it must be remembered that these developments merely allowed surgeons to realize the dreams of their fathers and grandfathers; they opened up remarkably few original avenues. The two outstanding phenomena of the 1950s and 1960s—heart surgery and organ transplantation—both originated in a real and practical manner at the turn of the century.

Support from other technologies. At first, perhaps, the surgeon tried to do too much himself, but before long his failures taught him to share his problems with experts in other fields. This was especially so with respect to difficulties of biomedical engineering and the exploitation of new materials. The relative protection from infection given by antibiotics and chemotherapy allowed the surgeon to become far more adventurous than hitherto in repairing and replacing damaged or worn-out tissues with foreign materials. Much research was still needed to find the best material for a particular purpose and to make sure that it would be acceptable to the body.

Use of plastics in surgery

Plastics, in their seemingly infinite variety, have come to be used for almost everything from suture material to heart valves; for strengthening the repair of hernias; for replacement of the head of the femur (first done by the French surgeon Jean Judet and his brother Robert-Louis Judet in 1950); for replacement of the lens of the eye after extraction of the natural lens for cataract; for valves to drain fluid from the brain in patients with hydrocephalus; and for many other applications. This is a far cry, indeed, from the unsatisfactory use of celluloid to restore bony defects of the face by the German surgeon Fritz Berndt in the 1890s. Inert metals, such as vitallium, have also found a place in surgery, largely in orthopedics for the repair of fractures and the replacement of joints.

The scope of surgery was further expanded by the introduction of the operating microscope. This brought the benefit of magnification particularly to neurosurgery and to ear surgery. In the latter it opened up a whole field of operations on the eardrum and within the middle ear. The principles of these operations were stated in 1951 and 1952 by two German surgeons, Fritz Zöllner and Horst Wullstein; and in 1952 Samuel Rosen of New York mobilized the footplate of the stapes to restore hearing in otosclerosis—a procedure attempted by the German Jean Kessel in 1876.

Although surgeons aim to preserve as much of the body as disease permits, they are sometimes forced to take radical measures to save life; when, for instance, cancer affects the pelvic organs. Pelvic exenteration (surgical removal of the pelvic organs and nearby structures) in two stages was devised by Allen Whipple of New York City, in 1935, and in one stage by Alexander Brunschwig, of Chicago, in 1937. Then, in 1960, Charles S. Kennedy, of Detroit, after a long discussion with Brunschwig, put into practice an operation that he had been considering for 12 years: hemicorporectomy—surgical removal of the lower part of the body. The patient died on the 11th day. The first successful hemicorporectomy (at the level between the lowest lumbar vertebra and the sacrum) was performed 18 months later by J. Bradley Aust and Karel B. Absolon, of Minnesota. This operation would never have been possible without all the technical, supportive, and rehabilitative resources of modern medicine.

Heart surgery. The attitude of the medical profession toward heart surgery was for long overshadowed by doubt and disbelief. Wounds of the heart could be sutured (first done successfully by Ludwig Rehn, of Frankfurt am Main, in 1896); the pericardial cavity—the cavity formed by the sac enclosing the heart—could be drained in purulent infections (as had been done by Larrey in 1824); and the pericardium could be partially excised for constrictive pericarditis when it was inflamed and constricted the movement of the heart (this operation was performed by Rehn and Sauerbruch in 1913). But little beyond these procedures found acceptance.

Heart operations of early 20th century

Yet, in the first two decades of the 20th century, much experimental work had been carried out, notably by the French surgeons Théodore Tuffier and Alexis Carrel. Tuffier, in 1912, operated successfully on the aortic valve. In 1923 Elliott Cutler of Boston used a tenotome, a tendon-cutting instrument, to relieve a girl's mitral stenosis (a narrowing of the mitral valve between the upper and lower chambers of the left side of the heart) and in 1925, in London, Henry Souttar used a finger to dilate a mitral valve in a manner that was 25 years ahead of its time. Despite these achievements, there was too much experimental failure, and heart disease remained a medical, rather than surgical, matter.

Resistance began to crumble in 1938, when Robert Gross successfully tied off a persistent ductus arteriosus (a fetal blood vessel between the pulmonary artery and the aorta). It was finally swept aside in World War II by the remarkable record of Dwight Harken, who removed 134 missiles from the chest—13 in the heart chambers—without the loss of one patient.

After the war, advances came rapidly, with the initial emphasis on the correction or amelioration of congenital defects. Gordon Murray, of Toronto, made full use of his amazing technical ingenuity to devise and perform many pioneering operations. And Charles Bailey of Philadelphia, adopting a more orthodox approach, was responsible for establishing numerous basic principles in the growing specialty.

Until 1953, however, the techniques all had one great disadvantage: they were done "blind." The surgeon's dream was to stop the heart so that he could see what he was doing and be allowed more time in which to do it. In 1952 this dream began to come true when Floyd Lewis, of Minnesota, reduced the temperature of the body so as to lessen its need for oxygen while he closed a hole between the two upper heart chambers, the atria. The next year John Gibbon, Jr., of Philadelphia brought to fulfillment the research he had begun in 1937; he used his heart–lung machine to supply oxygen while he closed a hole in the septum between the atria.

Unfortunately, neither method alone was ideal, but intensive research and development led, in the early 1960s, to their being combined as extracorporeal cooling. That is, the blood circulated through a machine outside the body, which cooled it (and, after the operation, warmed it); the cooled blood lowered the temperature of the whole body. With the heart dry and motionless, the surgeon operated on the coronary arteries; he inserted plastic patches over holes; he sometimes almost remodeled the inside of the heart. But when it came to replacing valves destroyed by disease, he was faced with a difficult choice between human tissue and man-made valves, or even valves from animal sources.

Organ transplantation. In 1967 surgery arrived at a climax that made the whole world aware of its medico-surgical responsibilities when the South African surgeon Christiaan Barnard transplanted the first human heart. Reaction, both medical and lay, contained more than an element of hysteria. Yet, in 1964, James Hardy, of the University of Mississippi, had transplanted a chimpanzee's heart into a man; and in that year two prominent research workers, Richard Lower and Norman E. Shumway, had written: "Perhaps the cardiac surgeon should pause while society becomes accustomed to resurrection of the mythological chimera." Research had been remorselessly leading up to just such an operation ever since Charles Guthrie and Alexis Carrel, at the University of Chicago, perfected the suturing of blood vessels in 1905 and then carried out experiments in the transplantation of many organs, including the heart.

New developments in immunosuppression (the use of drugs to prevent organ rejection) have advanced the field of transplantation enormously. Kidney transplantation is now a routine procedure that is supplemented by dialysis

with an artificial kidney (invented by Willem Kolff in wartime Holland) before and after the operation; mortality has been reduced to about 10 percent per year. Rejection of the transplanted heart by the patient's immune system was overcome to some degree in the 1980s with the introduction of the immunosuppressant cyclosporine; records show that many patients have lived for five or more years after the transplant operation.

The complexity of the liver and the unavailability of supplemental therapies such as the artificial kidney have contributed to the slow progress in liver transplantation (first performed in 1963 by Thomas Starzl). An increasing number of patients, especially children, have undergone successful transplantation; however, a substantial number may require retransplantation due to the failure of the first graft.

Lung transplants (first performed by Hardy in 1963) are difficult procedures, and much progress is yet to be made in preventing rejection. A combined heart-lung transplant is still in the experimental stage, but it is being met with increasing success; two-thirds of those receiving transplants are surviving, although complications such as infection are still common. Transplantation of all or part of the pancreas is not completely successful, and further refinements of the procedures (first performed in 1966 by Richard Lillehei) are needed. (R.G.R./Ed.)

THE PRACTICE OF MODERN MEDICINE

Health care and its delivery

The World Health Organization at its 1978 international conference held in the Soviet Union produced the Alma-Ata Health Declaration, which was designed to serve governments as a basis for planning health care that would reach people at all levels of society. The declaration reaffirmed that "health, which is a state of complete physical, mental and social well-being, and not merely the absence of disease or infirmity, is a fundamental human right and that the attainment of the highest possible level of health is a most important world-wide social goal whose realization requires the action of many other social and economic sectors in addition to the health sector." In its widest form the practice of medicine, that is to say the promotion and care of health, is concerned with this ideal.

The Alma-Ata declaration

ORGANIZATION OF HEALTH SERVICES

It is generally the goal of most countries to have their health services organized in such a way to ensure that individuals, families, and communities obtain the maximum benefit from current knowledge and technology available for the promotion, maintenance, and restoration of health. In order to play their part in this process governments and other agencies are faced with numerous tasks, including the following: (1) They must obtain as much information as is possible on the size, extent, and urgency of their needs; without accurate information, planning can be misdirected. (2) These needs must then be revised against the resources likely to be available in terms of money, manpower, and materials; developing countries may well require external aid to supplement their own resources. (3) Based on their assessments, countries then need to determine realistic objectives and draw up plans. (4) Finally, a process of evaluation needs to be built into the program; the lack of reliable information and accurate assessment can lead to confusion, waste, and inefficiency.

Health services of any nature reflect a number of interrelated characteristics, among which the most obvious, but not necessarily the most important from a national point of view, is the curative function; that is to say, caring for those already ill. Others include special services that deal with particular groups (such as children or pregnant women) and with specific needs such as nutrition or immunization; preventive services, the protection of the health both of individuals and of communities; health education; and, as mentioned above, the collection and analysis of information.

Levels of health care. In the curative domain there are various forms of medical practice. They may be thought of generally as forming a pyramidal structure, with three tiers representing increasing degrees of specialization and technical sophistication but catering to diminishing numbers of patients as they are filtered out of the system at a lower level. Only those patients who require special attention either for diagnosis or treatment should reach the second (advisory) or third (specialized treatment) tiers where the cost per item of service becomes increasingly higher. The first level represents primary health care, or first contact care, at which patients have their initial contact with the health-care system.

Structure of health-care delivery

Primary health care is an integral part of a country's health maintenance system, of which it forms the largest and most important part. As described in the declaration of Alma-Ata primary health care should be "based on practical, scientifically sound and socially acceptable methods and technology made universally accessible to individuals and families in the community through their full participation and at a cost that the community and country can afford to maintain at every stage of their development." Primary health care in the developed countries is usually the province of a medically qualified physician; in the developing countries first contact care is often provided by nonmedically qualified personnel.

The vast majority of patients can be fully dealt with at the primary level. Those who cannot are referred to the second tier (secondary health care, or the referral services) for the opinion of a consultant with specialized knowledge or for X-ray examinations and special tests. Secondary health care often requires the technology offered by a local or regional hospital. Increasingly, however, the radiological and laboratory services provided by hospitals are available directly to the family doctor, thus improving his service to patients and increasing its range. The third tier of health care, employing specialist services, is offered by institutions such as teaching hospitals and units devoted to the care of particular groups—women, children, patients with mental disorders, and so on. The dramatic differences in the cost of treatment at the various levels is a matter of particular importance in developing countries, where the cost of treatment for patients at the primary health-care level is usually only a small fraction of that at the third level; medical costs at any level in such countries, however, are usually borne by the government.

Ideally, provision of health care at all levels will be available to all patients; such health care may be said to be universal. The well-off, both in relatively wealthy industrialized countries and in the poorer developing world, may be able to get medical attention from sources they prefer and can pay for in the private sector. The vast majority of people in most countries, however, are dependent in various ways upon health services provided by the state, to which they may contribute comparatively little or, in the case of poor countries, nothing at all.

Costs of health care. The costs to national economics of providing health care is considerable and has been growing at a rapidly increasing rate, especially in countries such as the United States, West Germany, and Sweden; the rise in Britain has been less rapid. This trend has been the cause of major concerns in both developed and developing countries. Some of this concern is based upon the lack of any consistent evidence to show that more spending on health care produces better health. There is a movement in developing countries to replace the type of organization of health-care services that evolved during European colonial times with some less expensive, and for them, more appropriate, health-care system.

Cost efficiency of health-care delivery

In the industrialized world the growing cost of health services has caused both private and public health-care delivery systems to question current policies and to seek more economical methods of achieving their goals. Despite expenditures health services are not always used

effectively by those who need them, and results can vary widely from community to community. In Britain, for example, between 1951 and 1971 the death rate fell by 24 percent in the wealthier sections of the population but by only half that in the most underprivileged sections of society. The achievement of good health is reliant upon more than just the quality of health care. Health entails such factors as good education, safe working conditions, a favourable environment, amenities in the home, well-integrated social services, and reasonable standards of living.

In the developing countries. The developing countries differ from one another culturally, socially, and economically, but what they have in common is a low average income per person, with large percentages of their populations living at or below the poverty level. Although most have a small elite class, living mainly in the cities, the largest part of their populations live in rural areas. Urban regions in developing and some developed countries in the mid- and late 20th century have developed pockets of slums, which are growing because of an influx of rural peoples. For lack of even the simplest measures, vast numbers of urban and rural poor die each year of preventable and curable diseases, often associated with poor hygiene and sanitation, impure water supplies, malnutrition, vitamin deficiencies, and chronic preventable infections. The effect of these and other deprivations is reflected by the finding that in the 1980s the life expectancy at birth for men and women was about one-third less in Africa than it was in Europe; similarly, infant mortality in Africa was about eight times greater than in Europe. The extension of primary health-care services is therefore a high priority in the developing countries.

The developing countries themselves, lacking the proper resources, have often been unable to generate or implement the plans necessary to provide required services at the village or urban poor level. It has, however, become clear that the system of health care that is appropriate for one country is often unsuitable for another. Research has established that effective health care is related to the special circumstances of the individual country, its people, culture, ideology, and economic and natural resources.

Relating health-care systems to different cultures

The rising costs of providing health care have influenced a trend, especially among the developing nations, to promote services that employ less highly trained primary health-care personnel who can be distributed more widely in order to reach the largest possible proportion of the community. The principal medical problems to be dealt with in the developing world include undernutrition, infection, gastrointestinal disorders, and respiratory complaints, which themselves may be the result of poverty, ignorance, and poor hygiene. For the most part, these are easy to identify and to treat. Furthermore, preventive measures are usually simple and cheap. Neither treatment nor prevention requires extensive professional training: in most cases they can be dealt with adequately by the "primary health worker," a term that includes all nonprofessional health personnel.

In the developed countries. Those concerned with providing health care in the developed countries face a different set of problems. The diseases so prevalent in the Third World have, for the most part, been eliminated or are readily treatable. Many of the adverse environmental conditions and public health hazards have been conquered. Social services of varying degrees of adequacy have been provided. Public funds can be called upon to support the cost of medical care, and there are a variety of private insurance plans available to the consumer. Nevertheless, the funds that a government can devote to health care are limited and the cost of modern medicine continues to increase, thus putting adequate medical services beyond the reach of many. Adding to the expense of modern medical practices is the increasing demand for greater funding of health education and preventive measures specifically directed toward the poor. (Ha.Sc.)

ADMINISTRATION OF PRIMARY HEALTH CARE

In many parts of the world, particularly in developing countries, people get their primary health care, or first-contact care, where available at all, from nonmedically qualified personnel; these cadres of medical auxiliaries are being trained in increasing numbers to meet overwhelming needs among rapidly growing populations. Even among the comparatively wealthy countries of the world, containing in all a much smaller percentage of the world's population, escalation in the costs of health services and in the cost of training a physician has precipitated some movement toward reappraisal of the role of the medical doctor in the delivery of first-contact care.

Medical auxiliary training

In advanced industrial countries, however, it is usually a trained physician who is called upon to provide the first-contact care. The patient seeking first-contact care can go either to a general practitioner or turn directly to a specialist. Which is the wisest choice has become a subject of some controversy. The general practitioner, however, is becoming rather rare in some developed countries. In countries where he does still exist, he is being increasingly observed as an obsolescent figure, because medicine covers an immense, rapidly changing, and complex field of which no physician can possibly master more than a small fraction. The very concept of the general practitioner, it is thus argued, may be absurd.

The obvious alternative to general practice is the direct access of a patient to a specialist. If a patient has problems with vision, he goes to an eye specialist, and if he has a pain in his chest (which he fears is due to his heart), he goes to a heart specialist. One objection to this plan is that the patient often cannot know which organ is responsible for his symptoms, and the most careful physician, after doing many investigations, may remain uncertain as to the cause. Breathlessness—a common symptom—may be due to heart disease, to lung disease, to anemia, or to emotional upset. Another common symptom is general malaise—feeling run-down or always tired; others are headache, chronic low backache, rheumatism, abdominal discomfort, poor appetite, and constipation. Some patients may also be overtly anxious or depressed. Among the most subtle medical skills is the ability to assess people with such symptoms and to distinguish between symptoms that are caused predominantly by emotional upset and those that are predominantly of bodily origin. A specialist may be capable of such a general assessment, but, often, with emphasis on his own subject, he fails at this point. The generalist with his broader training is often the better choice for a first diagnosis, with referral to a specialist as the next option.

It is often felt that there are also practical advantages for the patient in having his own doctor, who knows about his background, who has seen him through various illnesses, and who has often looked after his family as well. This personal physician, often a generalist, is in the best position to decide when the patient should be referred to a consultant.

The advantages of general practice and specialization are combined when the physician of first contact is a pediatrician. Although he sees only children and thus acquires a special knowledge of childhood maladies, he remains a generalist who looks at the whole patient. Another combination of general practice and specialization is represented by group practice, the members of which partially or fully specialize. One or more may be general practitioners, and one may be a surgeon, a second an obstetrician, a third a pediatrician, and a fourth an internist. In isolated communities group practice may be a satisfactory compromise, but in urban regions, where nearly everyone can be sent quickly to a hospital, the specialist surgeon working in a fully equipped hospital can usually provide better treatment than a general practitioner surgeon in a small clinic hospital.

MEDICAL PRACTICE IN DEVELOPED COUNTRIES

Britain. Before 1948, general practitioners in Britain settled where they could make a living. Patients fell into two main groups: weekly wage earners, who were compulsorily insured, were on a doctor's "panel" and were given free medical attention (for which the doctor was paid quarterly by the government); most of the remainder paid the doctor a fee for service at the time of the illness. In 1948 the National Health Service began operation.

National Health Service

Under its provisions, everyone is entitled to free medical attention with a general practitioner with whom he is registered. Though general practitioners in the National Health Service are not debarred from also having private patients, these must be people who are not registered with them under the National Health Service. Any physician is free to work as a general practitioner entirely independent of the National Health Service, though there are few who do so. Almost the entire population is registered with a National Health Service general practitioner, and the vast majority automatically sees this physician, or one of his partners, when they require medical attention. A few people, mostly wealthy, while registered with a National Health Service general practitioner, regularly see another physician privately; and a few may occasionally seek a private consultation because they are dissatisfied with their National Health Service physician.

A general practitioner under the National Health Service remains an independent contractor, paid by a capitation fee; that is, according to the number of people registered with him. He may work entirely from his own office, and he provides and pays his own receptionist, secretary, and other ancillary staff. Most general practitioners have one or more partners and work more and more in premises built for the purpose. Some of these structures are erected by the physicians themselves, but many are provided by the local authority, the physicians paying rent for using them. Health centres, in which groups of general practitioners work have become common.

In Britain only a small minority of general practitioners can admit patients to a hospital and look after them personally. Most of this minority are in country districts, where, before the days of the National Health Service, there were cottage hospitals run by general practitioners; many of these hospitals continued to function in a similar manner. All general practitioners use such hospital facilities as X-ray departments and laboratories, and many general practitioners work in hospitals in emergency rooms (casualty departments) or as clinical assistants to consultants, or specialists.

General practitioners are spread more evenly over the country than formerly, when there were many in the richer areas and few in the industrial towns. The maximum allowed list of National Health Service patients per doctor is 3,500; the average is about 2,500. Patients have free choice of the physician with whom they register, with the proviso that they cannot be accepted by one who already has a full list and that a physician can refuse to accept them (though such refusals are rare). In remote rural places there may be only one physician within a reasonable distance.

Home visitations

Until the mid-20th century it was not unusual for the doctor in Britain to visit patients in their own homes. A general practitioner might make 15 or 20 such house calls in a day, as well as seeing patients in his office or "surgery," often in the evenings. This enabled him to become a family doctor in fact as well as in name. In modern practice, however, a home visit is quite exceptional and is paid only to the severely disabled or seriously ill when other recourses are ruled out. All patients are normally required to go to the doctor.

It has also become unusual for a personal doctor to be available during weekends or holidays. His place may be taken by one of his partners in a group practice, a provision that is reasonably satisfactory. General practitioners, however, may now use one of several commercial deputizing services that employs young doctors to be on call. Although some of these young doctors may be well experienced, patients do not generally appreciate this kind of arrangement. (J.W.T./Ha.Sc.)

United States. Whereas in Britain the doctor of first contact is regularly a general practitioner, in the United States the nature of first-contact care is less consistent.

Decline of general practice

General practice in the United States has been in a state of decline in the second half of the 20th century, especially in metropolitan areas. The general practitioner, however, is being replaced to some degree by the growing field of family practice. In 1969 family practice was recognized as a medical specialty after the American Academy of General Practice (now the American Academy of Family Physicians) and the American Medical Association created the American Board of General (now Family) Practice. Since that time the field has become one of the larger medical specialties in the United States. The family physicians were the first group of medical specialists in the United States for whom recertification was required.

There is no national health service, as such, in the United States. Most physicians in the country are in some form of private practice whether they see patients in their own offices, clinics, medical centres, or another type of facility and regardless of the patients' income. Doctors are usually compensated by such state and federally supported agencies as Medicaid (for treating the poor) and Medicare (for treating the elderly); not all doctors, however, accept poor patients. There are also some state-supported clinics and hospitals where the poor and elderly may receive free or low-cost treatment, and some doctors devote a small percentage of their time to treatment of the indigent. Veterans may receive free treatment at Veterans Administration hospitals, and the federal government provides medical services to American Indians and Alaskan natives through its Indian Health Service, sometimes using trained auxiliaries for first-contact care.

In the rural United States first-contact care is likely to come from a generalist. The middle- and upper-income groups living in urban areas, however, have access to a larger number of primary medical care options. Children are often taken to pediatricians, who may oversee the child's health needs until adulthood. Adults frequently make their initial contact with an internist, whose field is mainly that of medical (as opposed to surgical) illnesses; he often becomes the family physician. Other adults choose to go directly to physicians with narrower specialties, including dermatologists, allergists, gynecologists, orthopedists, and ophthalmologists. Patients in the United States may also choose to be treated by doctors of osteopathy. These doctors are fully qualified, but they make up only a small percentage of the country's physicians. They may also branch off into specialties, but general practice is much more common in their group than among M.D.'s.

It has been common in the United States for physicians providing primary care to work independently, providing their own equipment and paying their own ancillary staff. In smaller cities they mostly had full hospital privileges, but in the larger cities these privileges were more likely to be restricted. Physicians, often sharing the same specialties, are increasingly entering into group associations, where the expenses of office space, staff, and equipment may be shared; such associations may work out of suites of offices, clinics, or medical centres. The increasing competition and risks of private practice have caused many physicians to join Health Maintenance Organizations (HMOs), which provide comprehensive medical care and hospital care on a prepaid basis. The cost savings to the patient is considerable, but they must use only the HMO doctors and facilities. HMOs stress preventive medicine and outpatient treatment as opposed to hospitalization as a means of reducing costs, a policy that has caused an increased number of empty U.S. hospital beds.

Health Maintenance Organizations

While the number of doctors per 100,000 population in the United States has been steadily increasing, there has been a trend among physicians toward the use of trained medical personnel to handle some of the basic services normally performed by the doctor. In the United States so-called physician extender services are commonly divided into nurse practitioners and physician's assistants, both of whom provide similar ancillary services for the general practitioner or specialist. Such personnel do not replace the doctor. Almost all U.S. physicians have systems for taking each other's calls when they become unavailable. House calls in the United States, as in Britain, have become exceedingly rare. (Ed.)

Soviet Union. In the Soviet Union general practitioners are most prevalent in the thinly populated rural areas. Pediatricians deal with children up to the age of 15. Internists look after the medical ills of adults, and occupational physicians deal with the workers, sharing care with internists. Patients have no free choice of physician, but, if

there is cause, it may be possible for a change to be made at the instigation of either. Attention is free, though some drugs have to be paid for.

Soviet polyclinics

Teams of physicians with varying specialties work from polyclinics or outpatient units, where many types of diseases are treated. Small towns usually have one polyclinic for all purposes. Large cities commonly have separate polyclinics for children and adults, as well as clinics for special purposes such as women's diseases, mental illnesses, and venereal diseases. Polyclinics usually have X-ray apparatus and facilities for examination of tissue specimens, facilities associated with the departments of the district hospital. Beginning in the late 1970s there was an emerging trend toward the development of more large, multipurpose treatment centres, first aid hospitals, and specialized all-union medicine and health-care centres.

Home visits are common, and much of the physician's time is spent in doing routine checkups for preventive purposes. Some patients in sparsely populated rural areas may be seen first by feldshers (auxiliary health workers), nurses, or midwives who work under the supervision of a polyclinic or hospital physician. The feldsher was once a lower grade physician in the army or peasant communities, but his work is now regarded as paramedical.

Japan. In Japan, with less rigid legal restriction of the sale of pharmaceuticals than in the West, there was formerly a strong tradition of self-medication and self-treatment. This was modified in 1961 by the institution of health insurance programs that covered a large proportion of the population; there was then a great increase in visits to the outpatient clinics of hospitals and to private clinics and individual physicians.

When Japan shifted from traditional Chinese medicine with the adoption of Western medical practices in the 1870s, Germany became the chief model. As a result of German influence and of their own traditions, Japanese physicians tended to prefer professorial status and scholarly research opportunities at the universities or positions in the national or prefectural hospitals to private practice. There were some pioneering physicians, however, who brought medical care to the ordinary people.

Urban–rural distribution of physicians

Physicians in Japan have tended to cluster in the urban areas. The Medical Service Law of 1963 was amended to empower the Ministry of Health and Welfare to control the planning and distribution of future public and nonprofit medical facilities, partly to redress the urban–rural imbalance. Meanwhile, mobile services were expanded.

The influx of patients into hospital and private clinics after the passage of the national health insurance acts of 1961 had, as one effect, a severe reduction in the amount of time available for any one patient. Perhaps in reaction to this situation, there has been a modest resurgence in the popularity of traditional Chinese medicine, with its leisurely interview, its dependence on herbal and other "natural" medicines, and its other traditional diagnostic and therapeutic practices. The rapid aging of the Japanese population as a result of the sharply decreasing death rate and birth rate has created an urgent need for expanded health-care services for the elderly. There has also been an increasing need for centres to treat health problems resulting from environmental causes.

Medical practice in Europe

Other developed countries. On the continent of Europe there are great differences both within single countries and between countries in the kinds of first-contact medical care. General practice, while declining in Europe as elsewhere, is still rather common even in some large cities, as well as in remote country areas.

In The Netherlands, departments of general practice are run by general practitioners in all the medical schools—an exceptional state of affairs—and general practice flourishes. In the larger cities of Denmark, general practice on an individual basis is usual and popular, because the physician works only during office hours, there being a duty doctor service for nights and weekends. In the cities of Sweden, primary care is given by specialists. In the remote regions of northern Sweden, district doctors act as general practitioners to patients spread over huge areas and delegate much of their home visiting to nurses.

In France there are still general practitioners, but their number is declining. Many medical practitioners advertise themselves directly to the public as specialists in internal medicine, ophthalmologists, gynecologists, and other kinds of specialists. Even when patients have a general practitioner, they may still go directly to a specialist. Attempts to stem the decline in general practice are being made by the development of group practice and of small rural hospitals equipped to deal with less serious illnesses, where general practitioners can look after their patients.

Although Israel has a high ratio of physicians to population, there is a shortage of general practitioners, and only in rural areas is general practice common. In the towns many people go directly to pediatricians, gynecologists, and other specialists, but there has been a reaction against this direct access to the specialist. More general practitioners have been trained, and the Israel Medical Association has recommended that no patient should be referred to a specialist except by the family physician or on instructions given by the family nurse. At Tel Aviv University there is a department of family medicine. In some newly developing areas, where the doctor shortage is greatest, there are medical centres at which all patients are initially interviewed by a nurse. The nurse may deal with many minor ailments, thus freeing the physician to treat the more seriously ill.

Nearly half the medical doctors in Australia are general practitioners—a far higher proportion than in most other advanced countries—though, as elsewhere, their numbers are declining. They tend to do far more for their patients than in Britain, many performing such operations as removal of the appendix, gallbladder, or uterus, operations that elsewhere would be carried out by a specialist surgeon. Group practices are common.

MEDICAL PRACTICE IN DEVELOPING COUNTRIES

China. Health services in China since the Cultural Revolution have been characterized by decentralization and dependence on personnel chosen locally and trained for short periods. Emphasis is given to selfless motivation, self-reliance, and to the involvement of everyone in the community. Campaigns stressing the importance of preventive measures and their implementation have served to create new social attitudes as well as to break down divisions between different categories of health workers. Health care is regarded as a local matter that should not require the intervention of any higher authority; it is based upon a highly organized and well-disciplined system that is egalitarian rather than hierarchical, as in Western societies, and which is well suited to the rural areas where about two-thirds of the population live. In the large and crowded cities an important constituent of the health-care system is the residents' committees, each for a population of 1,000 to 5,000 people. Care is provided by part-time personnel with periodic visits by a doctor. A number of residents' committees are grouped together into neighbourhoods of some 50,000 people where there are clinics and general hospitals staffed by doctors as well as health auxiliaries trained in both traditional and Westernized medicine. Specialized care is provided at the district level (over 100,000 people), in district hospitals and in epidemic and preventive medicine centres. In many rural districts people's communes have organized cooperative medical services that provide primary care for a small annual fee.

Residents' committees

Throughout China the value of traditional medicine is stressed, especially in the rural areas. All medical schools are encouraged to teach traditional medicine as part of their curriculum, and efforts are made to link colleges of Chinese medicine with Western-type medical schools. Medical education is of shorter duration than it is in Europe, and there is greater emphasis on practical work. Students spend part of their time away from the medical school working in factories or in communes; they are encouraged to question what they are taught and to participate in the educational process at all stages. One well-known form of traditional medicine is acupuncture, which is used as a therapeutic and pain-relieving technique; requiring the insertion of brass-handled needles at various points on the body, acupuncture has become quite prominent as a form of anesthesia.

Acupuncture

The vast number of nonmedically qualified health staff, upon whom the health-care system greatly depends, includes both full-time and part-time workers. The latter include so-called barefoot doctors, who work mainly in rural areas, worker doctors in factories, and medical workers in residential communities. None of these groups is medically qualified. They have had only a three-month period of formal training, part of which is done in a hospital, fairly evenly divided between theoretical and practical work. This is followed by a varying period of on-the-job experience under supervision.

Āyurvedic medicine

India. Āyurvedic medicine is an example of a well-organized system of traditional health care, both preventive and curative, that is widely practiced in parts of Asia. Āyurvedic medicine has a long tradition behind it, having originated in India perhaps as long as 3,000 years ago. It is still a favoured form of health care in large parts of the Eastern world, especially in India, where a large percentage of the population use this system exclusively or combined with modern medicine. The Indian Medical Council was set up in 1971 by the Indian government to establish maintenance of standards for undergraduate and postgraduate education. It establishes suitable qualifications in Indian medicine and recognizes various forms of traditional practice including Āyurvedic, Unani, and Siddha. Projects have been undertaken to integrate the indigenous Indian and Western forms of medicine. Most Āyurvedic practitioners work in rural areas, providing health care to at least 500,000,000 people in India alone. They therefore represent a major force for primary health care, and their training and deployment are important to the government of India.

Like scientific medicine, Āyurvedic medicine has both preventive and curative aspects. The preventive component emphasizes the need for a strict code of personal and social hygiene, the details of which depend upon individual, climatic, and environmental needs. Bodily exercises, the use of herbal preparations, and Yoga form a part of the remedial measures. The curative aspects of Āyurvedic medicine involves the use of herbal medicines, external preparations, physiotherapy, and diet. It is a principle of Āyurvedic medicine that the preventive and therapeutic measures be adapted to the personal requirements of each patient.

Other developing countries. A main goal of the World Health Organization (WHO), as expressed in the Alma-Ata Declaration of 1978, is to provide to all the citizens of the world a level of health that will allow them to lead socially and economically productive lives by the year 2000. By the late 1980s, however, vast disparities in health care still existed between the rich and poor countries of the world. In developing countries such as Ethiopia, Guinea, Mali, and Mozambique, for instance, governments in the late 1980s spent less than $5 per person per year on public health, while in most western European countries several hundred dollars per year was spent on each person. The disproportion of the number of physicians available between developing and developed countries is similarly wide.

Medical needs in developing countries

Along with the shortage of physicians, there is a shortage of everything else needed to provide medical care—of equipment, drugs, and suitable buildings, and of nurses, technicians, and all other grades of staff, whose presence is taken for granted in the affluent societies. Yet there are greater percentages of sick in the poor countries than in the rich countries. In the poor countries a high proportion of people are young, and all are liable to many infections, including tuberculosis, syphilis, typhoid, and cholera (which, with the possible exception of syphilis, are now rare in the rich countries), and also malaria, yaws, worm infestations, and many other conditions occurring primarily in the warmer climates. Nearly all of these infections respond to the antibiotics and other drugs that have been discovered since the 1920s. There is also much malnutrition and anemia, which can be cured if funding is available. There is a prevalence of disorders remediable by surgery. Preventive medicine can ensure clean water supplies, destroy insects that carry infections, teach hygiene, and show how to make the best use of resources.

In most poor countries there are a few people, usually living in the cities, who can afford to pay for medical care, and in a free market system the physicians tend to go where they can make the best living; this situation causes the doctor–patient ratio to be much higher in the towns than in country districts. A physician in Bombay or in Rio de Janeiro, for example, may have equipment as lavish as that of a physician in the United States and can earn an excellent income. The poor, both in town and country, however, can get medical attention only if it is paid for by the state, by some supranational body, or by a mission or other charitable organization. Often the quality of care may be low, and in remote regions it may be lacking altogether. In practice, hospitals run by a mission may cooperate closely with state-run health centres.

Because physicians are scarce, their skills must be used to best advantage, and much of the work normally done by physicians in the rich countries has to be delegated to auxiliaries or nurses, who have to diagnose the common conditions, give treatment, take blood samples, help with operations, supply simple posters containing health advice, and carry out other tasks. In such places the doctor has time only to perform major operations and deal with the more difficult medical problems. People are treated as far as possible on an outpatient basis from health centres housed in simple buildings; few can travel except on foot, and, if they are more than a few miles from a health centre, they tend not to go there. Health centres also may be used for health education.

Health care in Tanzania

Although primary health-care service differs from country to country, that developed in Tanzania is representative of many that have been devised in largely rural developing countries. The most important feature of the Tanzanian rural health service is the rural health centre, which, with its related dispensaries, is intended to provide comprehensive health services for the community. The staff is headed by the assistant medical officer and the medical assistant. The assistant medical officer has at least four years of experience, which is then followed by further training for 18 months. He is not a doctor but serves to bridge the gap between medical assistant and physician. The medical assistant has three years of general medical education. The work of the rural health centres and dispensaries is mainly of three kinds: diagnosis and treatment, maternal and child health, and environmental health. The main categories of primary health workers also include medical aids, maternal and child health aids, and health auxiliaries. Nurses and midwives form another category of worker. In the villages there are village health posts staffed by village medical helpers working under supervision from the rural health centre.

Primitive medicine

In some primitive elements of the societies of developing countries, and of some developed countries, there exists the belief that illness comes from the displeasure of ancestral gods and evil spirits, from the malign influence of evilly disposed persons, or from natural phenomena that can neither be forecast nor controlled. To deal with such causes there are many varieties of indigenous healers who practice elaborate rituals on behalf of both the physically ill and the mentally afflicted. If it is understood that such beliefs, and other forms of shamanism, may provide a basis upon which health care can be based, then primary health care may be said to exist almost everywhere. It is not only easily available but also readily acceptable, and often preferred, to more rational methods of diagnosis and treatment. Although such methods may sometimes be harmful, they may often be effective, especially where the cause is psychosomatic. Other patients, however, may suffer from a disease for which there is a cure in modern medicine.

In order to improve the coverage of primary health-care services and to spread more widely some of the benefits of Western medicine, attempts have sometimes been made to find a means of cooperation, or even integration, between traditional and modern medicine (see above *India*). In Africa, for example, some such attempts are officially sponsored by ministries of health, state governments, universities, and the like, and they have the approval of WHO, which often takes the lead in this activity. In view,

however, of the historical relationships between these two systems of medicine, their different basic concepts, and the fact that their methods cannot readily be combined, successful merging has been limited.

ALTERNATIVE OR COMPLEMENTARY MEDICINE

Persons dissatisfied with the methods of modern medicine or with its results sometimes seek help from those professing expertise in other, less conventional, and sometimes controversial, forms of health care. Such practitioners are not medically qualified unless they are combining such treatments with a regular (allopathic) practice, which includes osteopathy. In many countries the use of some forms, such as chiropractic, requires licensing and a degree from an approved college. The treatments afforded in these various practices are not always subjected to objective assessment, yet they provide services that are alternative, and sometimes complementary, to conventional practice. This group includes practitioners of homeopathy, naturopathy, acupuncture, hypnotism, and various meditative and quasi-religious forms. Numerous persons also seek out some form of faith healing to cure their ills, sometimes as a means of last resort. Religions commonly include some advents of miraculous curing within their scriptures. Belief in such curative powers has been in part responsible for the increasing popularity of the television, or "electronic," preacher in the United States, a phenomenon that involves millions of viewers. Millions of others annually visit religious shrines, such as the one at Lourdes in France, hoping to be miraculously healed.

Faith healing

SPECIAL PRACTICES AND FIELDS OF MEDICINE

Specialties in medicine. At the beginning of World War II it was possible to recognize a number of major medical specialties, including internal medicine, obstetrics and gynecology, pediatrics, pathology, anesthesiology, ophthalmology, surgery, orthopedic surgery, plastic surgery, psychiatry and neurology, radiology, and urology. Hematology was also an important field of study, and microbiology and biochemistry were important medically allied specialties. Since World War II, however, there has been an almost explosive increase of knowledge in the medical sciences as well as enormous advances in technology as applicable to medicine. These developments have led to more and more specialization. The knowledge of pathology has been greatly extended, mainly by the use of the electron microscope; similarly microbiology, which includes bacteriology, expanded with the growth of such other subfields as virology (the study of viruses) and mycology (the study of yeasts and fungi in medicine). Biochemistry, sometimes called clinical chemistry or chemical pathology, has contributed to the knowledge of disease, especially in the field of genetics where genetic engineering has become a key to curing some of the most difficult diseases. Hematology also expanded after World War II with the development of electron microscopy. Contributions to medicine have come from such fields as psychology and sociology especially in such areas as mental disorders and mental handicaps. Clinical pharmacology has led to the development of more effective drugs and to the identification of adverse reactions. More recently established medical specialties are those of preventive medicine, physical medicine and rehabilitation, family practice, and nuclear medicine. In the United States every medical specialist must be certified by a board composed of members of the specialty in which certification is sought. Some type of peer certification is required in most countries.

Certification

Expansion of knowledge both in depth and in range has encouraged the development of new forms of treatment that require high degrees of specialization, such as organ transplantation and exchange transfusion; the field of anesthesiology has grown increasingly complex as equipment and anesthetics have improved. New technologies have introduced microsurgery, laser beam surgery, and lens implantation (for cataract patients), all requiring the specialist's skill. Precision in diagnosis has markedly improved; advances in radiology, the use of ultrasound, computerized axial tomography (CAT scan), and nuclear magnetic resonance imaging are examples of the extension of technology requiring expertise in the field of medicine.

To provide more efficient service it is not uncommon for a specialist surgeon and a specialist physician to form a team working together in the field of, for example, heart disease. An advantage of this arrangement is that they can attract a highly trained group of nurses, technologists, operating room technicians, and so on, thus greatly improving the efficiency of the service to the patient. Such specialization is expensive, however, and has required an increasingly large proportion of the health budget of institutions, a situation that eventually has its financial effect on the individual citizen. The question therefore arises as to their cost-effectiveness. Governments of developing countries have usually found, for instance, that it is more cost-efficient to provide more people with basic care.

Teaching. In developed countries physicians frequently prefer posts in hospitals with medical schools. Newly qualified physicians want to work there because doing so will aid their future careers, though the actual experience may be wider and better in a hospital without a medical school. Senior physicians seek careers in hospitals with medical schools because consultant, specialist, or professorial posts there usually carry a high degree of prestige. When the posts are salaried, the salaries are sometimes, but not always, higher than in a nonteaching hospital. Usually a consultant who makes part of his living by private practice earns more when on the staff of a medical school.

Advantage of medical school association

In many medical schools, there are clinical professors in each of the major specialties, such as surgery, internal medicine, obstetrics and gynecology, psychiatry, and often of the smaller specialties as well. There are also professors of pathology, radiology, and radiotherapy. Whether professors or not, all doctors in teaching hospitals have the two functions of caring for the sick and educating students. They give lectures and seminars and are accompanied by students on ward rounds.

Industrial medicine. The Industrial Revolution greatly changed, and as a rule worsened, the health hazards caused by industry, while the numbers at risk vastly increased. In Britain the first small beginnings of efforts to ameliorate the lot of the workers in factories and mines began in 1802 with the passing of the first factory act, the Health and Morals of Apprentices Act. The factory act of 1838, however, was the first truly effective measure in the industrial field. It forbade night work for children and restricted their hours to 12 per day. Children under 13 were required to attend school. A factory inspectorate was established, the inspectors being given powers of entry into factories and power of prosecution of recalcitrant owners. Thereafter, there was a succession of acts with detailed regulations for safety and health in all industries. Industrial diseases were made notifiable, and those who developed any prescribed industrial disease were entitled to benefits.

The situation is similar in other developed countries. Physicians are bound by legal restrictions and must report industrial diseases. The industrial physician's most important function, however, is to prevent industrial diseases. Many of the measures to this end have become standard practice, but, especially in industries working with new substances, the physician has a purpose to determine if workers are being damaged and to suggest preventive measures. He advises management about industrial hygiene and may become involved in such matters as building design and the need for safety devices and protective clothing. The physician or health worker may also inform the worker of occupational health hazards.

Industrial physician's function

Modern factories usually have arrangements for giving first aid in case of accidents. Depending upon the size of the plant, the facilities may range from a simple first-aid station to a large suite of lavishly equipped rooms and may include a staff of qualified nurses and physiotherapists and one or perhaps more full-time physicians.

Periodic medical examination. Physicians in industry carry out medical examinations, especially on new employees and on those returning to work after sickness or injury. In addition, those liable to health hazards may be examined regularly in the hope of detecting evidence of incipient damage. In some organizations every employee may be offered a regular medical examination.

The industrial and the personal physician. When a worker also has a personal physician, there may be doubt, in some cases, as to which physician bears the main responsibility for his health. When someone has an accident or becomes acutely ill at work, the first aid is given or directed by the industrial physician. Subsequent treatment may either be given at the clinic at work or by the personal physician. Because of labour-management difficulties, workers sometimes tend not to trust the diagnosis of the management-hired physician.

Industrial health service in the Soviet Union

Industrial health services. In the Soviet Union and many other socialist countries industrial health service generally is more fully developed than in the capitalist countries. At the larger industrial establishments in the Soviet Union, polyclinics provide both occupational and general care to the workers and their families. Occupational physicians are responsible for the prevention of occupational diseases and injuries, health screening, immunization, and health education. He may be assisted by feldshers who have been specially trained in this work.

In the capitalist countries there is no fixed pattern of industrial health service. Legislation impinges upon health in various ways, including the provision of safety measures, the restriction of pollution, and the enforcement of minimum standards of lighting, ventilation, and space per person. In most of these countries there is found an infinite variety of schemes financed and run by individual firms or, equally, by huge industries. Labour unions have also done much to enforce health codes within their respective industries. In the developing countries there has been generally little advance in industrial medicine.

Family health care. In many societies special facilities are provided for the health care of pregnant women, mothers, and their young children. The health care needs of these three groups are generally recognized to be so closely related as to require a closely integrated service that includes prenatal care, the birth of the baby, the postnatal period, and the needs of the infant. Such a continuum should be followed by a service attentive to the needs of young children and then by a school health service. Family clinics are common in countries that have state-sponsored health services, such as those in the United Kingdom and elsewhere in Europe. Family health care in some developed countries, such as the United States, is provided for low-income groups by state-subsidized facilities, but other groups defer to private physicians or privately run clinics.

The prenatal clinic

Prenatal clinics provide a number of elements. There is, first, the care of the pregnant woman, especially if she is in a vulnerable group likely to develop some complication during the last few weeks of pregnancy and subsequent delivery. Many potential hazards, such as diabetes and high blood pressure, can be identified and measures taken to minimize their effects. In developing countries pregnant women are especially susceptible to many kinds of disorders, particularly infections such as malaria. Local conditions determine what special precautions should be taken to ensure a healthy child. Most pregnant women, in their concern to have a healthy child, are receptive to simple health education. The prenatal clinic provides an excellent opportunity to teach the mother how to look after herself during pregnancy, what to expect at delivery, and how to care for her baby. If the clinic is attended regularly, the woman's record will be available to the staff that will later supervise the delivery of the baby; this is particularly important for someone who has been determined to be at risk. The same clinical unit should be responsible for prenatal, natal, and postnatal care as well as for the care of the newborn infants.

Most pregnant women can be safely delivered in simple circumstances without an elaborately trained staff or sophisticated technical facilities, providing that these can be called upon in emergencies. In developed countries it was customary in premodern times for the delivery to take place in the woman's home supervised by a qualified midwife or by the family doctor. By the mid-20th century women, especially in urban areas, usually preferred to have their babies in a hospital, either in a general hospital or in a more specialized maternity hospital. In many developing countries traditional birth attendants supervise the delivery. They are women, for the most part without formal training, who have acquired skill by working with others and from their own experience. Normally they belong to the local community where they have the confidence of the family, where they are content to live and serve, and where their services are of great value. In many developing countries the better training of birth attendants has a high priority. In developed Western countries there has been a trend toward delivery by natural childbirth, including delivery in a hospital without anesthesia, and home delivery.

Postnatal health care

Postnatal care services are designed to supervise the return to normal of the mother. They are usually given by the staff of the same unit that was responsible for the delivery. Important considerations are the matter of breast- or artificial feeding and the care of the infant. Today the prospects for survival of babies born prematurely or after a difficult and complicated labour, as well as for neonates (recently born babies) with some physical abnormality, are vastly improved. This is due to technical advances, including those that can determine defects in the prenatal stage, as well as to the growth of neonatology as a specialty. A vital part of the family health-care service is the child welfare clinic, which undertakes the care of the newborn. The first step is the thorough physical examination of the child on one or more occasions to determine whether or not it is normal both physically and, if possible, mentally. Later periodic examinations serve to decide if the infant is growing satisfactorily. Arrangements can be made for the child to be protected from major hazards by, for example, immunization and dietary supplements. Any intercurrent condition, such as a chest infection or skin disorder, can be detected early and treated. Throughout the whole of this period mother and child are together, and particular attention is paid to the education of the mother for the care of the child.

A part of the health service available to children in the developed countries is that devoted to child guidance. This provides psychiatric guidance to maladjusted children usually through the cooperative work of a child psychiatrist, educational psychologist, and schoolteacher.

Geriatrics. Since the mid-20th century a change has occurred in the population structure in developed countries. The proportion of elderly people has been increasing. Since 1983, however, in most European countries the population growth of that group has leveled off, although it is expected to continue to grow more rapidly than the rest of the population in most countries through the first third of the 21st century. In the late 20th century Japan had the fastest growing elderly population.

Increasing need for elderly care

Geriatrics, the health care of the elderly, is therefore a considerable burden on health services. In the United Kingdom about one-third of all hospital beds are occupied by patients over 65; half of these are psychiatric patients. The physician's time is being spent more and more with the elderly, and since statistics show that women live longer than men, geriatric practice is becoming increasingly concerned with the treatment of women. Elderly people often have more than one disorder, many of which are chronic and incurable, and they need more attention from health-care services. In the United States there has been some movement toward making geriatrics a medical specialty, but it has not generally been recognized.

Support services for the elderly provided by private or state-subsidized sources include domestic help, delivery of meals, day-care centres, elderly residential homes or nursing homes, and hospital beds either in general medical wards or in specialized geriatric units. The degree of accessibility of these services is uneven from country to country and within countries. In the United States, for instance, although there are some federal programs, each state has its own elderly programs, which vary widely. However, as the elderly become an increasingly larger part of the population their voting rights are providing increased leverage for obtaining more federal and state benefits. The general practitioner or family physician working with visiting health and social workers and in conjunction with the patient's family often form a working team for elderly care.

In the developing world, countries are largely spared such geriatric problems, but not necessarily for positive

reasons. A principal cause, for instance, is that people do not live so long. Another major reason is that in the extended family concept, still prevalent among developing countries, most of the caretaking needs of the elderly are provided by the family.

Public health practice. The physician working in the field of public health is mainly concerned with the environmental causes of ill health and in their prevention. Bad drainage, polluted water and atmosphere, noise and smells, infected food, bad housing, and poverty in general are all his special concern. Perhaps the most descriptive title he can be given is that of community physician. In Britain he has been customarily known as the medical officer of health and, in the United States, as the health officer.

Public health measures of the 19th century

The spectacular improvement in the expectation of life in the affluent countries has been due far more to public health measures than to curative medicine. These public health measures began operation largely in the 19th century. At the beginning of that century, drainage and water supply systems were all more or less primitive; nearly all the cities of that time had poorer water and drainage systems than Rome had possessed 1,800 years previously. Infected water supplies caused outbreaks of typhoid, cholera, and other waterborne infections. By the end of the century, at least in the larger cities, water supplies were usually safe. Food-borne infections were also drastically reduced by the enforcement of laws concerned with the preparation, storage, and distribution of food. Insect-borne infections, such as malaria and yellow fever, which were common in tropical and semitropical climates, were eliminated by the destruction of the responsible insects. Fundamental to this improvement in health has been the diminution of poverty, for most public health measures are expensive. The peoples of the developing countries fall sick and sometimes die from infections that are virtually unknown in affluent countries.

Medical officer of health

Britain. Public health services in Britain are organized locally under the National Health Service. The medical officer of health is employed by the local council and is the adviser in health matters. The larger councils employ a number of mostly full-time medical officers; in some rural areas, a general practitioner may be employed part-time as medical officer of health.

The medical officer has various statutory powers conferred by acts of Parliament, regulations and orders, such as food and drugs acts, milk and dairies regulations, and factories acts. He supervises the work of sanitary inspectors in the control of health nuisances. The compulsorily notifiable infectious diseases are reported to him, and he takes appropriate action. Other concerns of the medical officer include those involved with the work of the district nurse, who carries out nursing duties in the home, and the health visitor, who gives advice on health matters, especially to the mothers of small babies. He has other duties in connection with infant welfare clinics, crèches, day and residential nurseries, the examination of schoolchildren, child guidance clinics, foster homes, factories, problem families, and the care of the aged and the handicapped.

United States. Federal, state, county, and city governments all have public health functions. Under the U.S. Department of Health and Human Services is the Public Health Service, headed by an assistant secretary for health and the surgeon general. State health departments are headed by a commissioner of health, usually a physician, who is often in the governor's cabinet. He usually has a board of health that adopts health regulations and holds hearings on their alleged violations. A state's public health code is the foundation on which all county and city health regulations must be based. A city health department may be independent of its surrounding county health department, or there may be a combined city-county health department. The physicians of the local health departments are usually called health officers, though occasionally people with this title are not physicians. The larger departments may have a public health director, a district health director, or a regional health director.

The minimal complement of a local health department is a health officer, a public health nurse, a sanitation expert, and a clerk who is also a registrar of vital statistics. There may also be sanitation personnel, nutritionists, social workers, laboratory technicians, and others.

Japan. Japan's Ministry of Health and Welfare directs public health programs at the national level, maintaining close coordination among the fields of preventive medicine, medical care, and welfare and health insurance. The departments of health of the prefectures and of the largest municipalities operate health centres. The integrated community health programs of the centres encompass maternal and child health, communicable disease control, health education, family planning, health statistics, food inspection, and environmental sanitation. Private physicians, through their local medical associations, help to formulate and execute particular public health programs needed by their localities.

Numerous laws are administered through the ministry's bureaus and agencies, which range from public health, environmental sanitation, and medical affairs to the children and families bureau. The various categories of institutions run by the ministry, in addition to the national hospitals, include research centres for cancer and leprosy, homes for the blind, rehabilitation centres for the physically handicapped, and port quarantine services.

Soviet Union. The public health services for the U.S.S.R. as a whole are directed by the Ministry of Health under a minister who is a member of the Council of Ministers. The ministry, through the 15 union republic ministries of health, directs all medical institutions within its competence as well as the public health authorities and services throughout the country.

The administration is centralized, with little local autonomy. Each of the 15 republics has its own ministry of health, which is responsible for carrying out the plans and decisions established by the U.S.S.R. Ministry of Health. Each republic is divided into *oblasti,* or provinces, which have departments of health directly responsible to the republic ministry of health. Each *oblast,* in turn, has *rayony* (municipalities), which have their own health departments accountable to the *oblast* health department. Finally, each *rayon* is subdivided into smaller *uchastoki* (districts).

In most rural *rayony* the responsibility for public health lies with the chief physician, who is also medical director of the central *rayon* hospital. This system ensures unity of public health administration and implementation of the principle of planned development. In the *uchastoki,* there are three types of salaried physicians: general practitioners or internists; pediatricians; and public health officers. There are also numbers of other health personnel, including nurses, feldshers, and midwives.

For more information on the history, organization, and progress of public health, see below.

Military practice. The medical services of armies, navies, and air forces are geared to war. During campaigns the first requirement is the prevention of sickness. In all wars before the 20th century, many more combatants died of disease than of wounds. And even in World War II, and wars thereafter, although few died of disease, vast numbers became casualties from disease.

Military prevention of sickness

The main means of preventing sickness are the provision of adequate food and pure water, thus eliminating starvation, avitaminosis, and dysentery and other bowel infections, which used to be particular scourges of armies; the provision of proper clothing and other means of protection from the weather; the elimination from the service of those likely to fall sick; the use of vaccination and suppressive drugs to prevent various infections, such as typhoid and malaria; and education in hygiene and in the prevention of venereal disease, a particular problem in the services. In addition, the maintenance of high morale has a striking effect on casualty rates, for, when morale is poor, men are likely to suffer psychiatric breakdowns, and malingering is more prevalent.

The medical branch may provide advice about disease prevention, but the actual execution of this advice is through the ordinary chains of command. It is the duty of the military, not of the medical, officer to ensure that the troops obey orders not to drink infected water and to take tablets to suppress malaria.

Army medical organization. The medical doctor of first

contact to the soldier in the armies of developed countries is usually an officer in the medical corps. In peacetime, he sees the sick and has functions similar to those of the general practitioner. He prescribes drugs and dressings and may have a sick bay where slightly sick men can remain for a few days. He is usually assisted by trained nurses and corpsmen. If a further medical opinion is required, the patient can be referred to a specialist at a military or a civilian hospital.

Dealing with the wounded

In a war zone, the medical officer has an aid post where, with the help of one or more corpsmen, he applies first aid to the walking wounded and to the more seriously wounded who are brought in. The casualties are evacuated as quickly as possible by field ambulances or helicopters. At a company station, medical officers and medical corpsmen may provide further treatment before patients are evacuated to the main dressing station at the field ambulance headquarters, where a surgeon may perform emergency operations. Thereafter, evacuation may be to casualty clearing stations, to advanced hospitals, or to base hospitals. Air evacuation is widely used.

In peacetime most of the intermediate medical units exist only in skeleton form; the active units are at the battalion and hospital level. When a physician joins the medical corps, he may join with specialist qualifications, or he may obtain such qualifications while in the army. A feature of army medicine is promotion to administrative positions. The commanding officer of a hospital and the medical officer at headquarters may have no contacts with actual patients.

Although the medical officer in peacetime has some choice of the kind of work he will do, he is in a chain of command and is subject to military discipline. When dealing with patients, however, he is in a special position; he cannot be ordered by a superior officer to give some treatment or take other action that he believes is wrong. The medical officer also does not bear or use arms unless his patients are being attacked.

Naval and air force medicine. Naval medical services are run on similar lines to those of the army. Junior

Medical organization in the navy

medical officers are attached to ships or to shore stations and deal with most cases of sickness in their units. When at sea, the medical officer has an exceptional degree of responsibility in that he is working alone, unless he is on a very large ship. In peacetime, only the larger ships carry a medical officer; in wartime, destroyers and other small craft may also carry medical officers. Serious cases go to either a shore-based hospital or a hospital ship.

Flying has many medical repercussions. Cold, lack of oxygen, and changes of direction at high speed all have important effects on bodily and mental functions. Armies and air forces may share the same medical services.

A developing field is aerospace medicine. This involves medical problems that were not experienced before spaceflight, for the main reason that humans in space are not under the influence of gravity, a condition that has profound physiological effects.

CLINICAL RESEARCH

The remarkable developments in medicine that have been brought about in the 20th century, especially since World War II, have been based on research either in the basic sciences related to medicine or in the clinical field. Advances in the use of radiations, nuclear energy, and space research have played an important part in this progress. Some laypeople often think of research as taking place only in sophisticated laboratories or highly specialized institutions where work is devoted to scientific advances that may, or may not, be applicable to medical practice. This notion, however, ignores the clinical research that takes place on a day-to-day basis in hospitals and doctors' offices.

Historical notes. Although the most spectacular changes in the medical scene during the 20th century, and the most widely heralded, have been the development of potent

Abandonment of remedies of the past

drugs and elaborate operations, another striking change has been the abandonment of most of the remedies of the past. In the mid-19th century, persons ill with numerous maladies were starved (partially or completely), bled, purged, cupped (by applying a tight-fitting vessel filled with steam to some part and then cooling the vessel), and rested, perhaps for months or even years. Much more recently they were prescribed various restricted diets and were routinely kept in bed for weeks after abdominal operations, for many weeks or months when their hearts were thought to be affected, and for many months or years with tuberculosis. The abandonment of these measures may not be thought of as involving research, but the physician who first encouraged persons who had peptic ulcers to eat normally (rather than to live on the customary bland foods) and the physician who first got his patients out of bed a week or two after they had had minor coronary thrombosis (rather than insisting on a minimum of six weeks of strict bed rest) were as much doing research as is the physician who first tries out a new drug on a patient. This research, by observing what happens when remedies are abandoned, has been of inestimable value, and the need for it has not passed.

Clinical observation. Much of the investigative clinical field work undertaken in the present day requires only relatively simple laboratory facilities because it is observational rather than experimental in character. A feature of much contemporary medical research is that it requires the collaboration of a number of persons, perhaps not all of them doctors. Despite the advancing technology, there is much to be learned simply from the observation and analysis of the natural history of disease processes as they begin to affect patients, pursue their course, and end, either in their resolution or by the death of the patient. Such studies may be suitably undertaken by physicians working in their offices who are in a better position than doctors working only in hospitals to observe the whole course of an illness. Disease rarely begins in a hospital and usually does not end there. It is notable, however, that observational research is subject to many limitations and pitfalls of interpretation, even when it is carefully planned and meticulously carried out.

Drug research. The administration of any medicament, especially a new drug, to a patient is fundamentally an experiment: so is a surgical operation, particularly if it involves a modification to an established technique or a completely new procedure. Concern for the patient, careful observation, accurate recording, and a detached mind are the keys to this kind of investigation, as indeed to all forms of clinical study. Because patients are individuals reacting to a situation in their own different ways, the data obtained in groups of patients may well require statistical analysis for their evaluation and validation.

New drug development

One of the striking characteristics in the medical field in the 20th century has been the development of new drugs, usually by pharmaceutical companies. Until the end of the 19th century, the discovery of new drugs was largely a matter of chance. It was in that period that Paul Ehrlich, the German scientist, began to lay down the principles for modern pharmaceutical research that made possible the development of a vast array of safe and effective drugs. Such benefits, however, bring with them their own disadvantages: it is estimated that as many as 30 percent of patients in, or admitted to, hospitals suffer from the adverse effect of drugs prescribed by a physician for their treatment (iatrogenic disease). Sometimes it is extremely difficult to determine whether a drug has been responsible for some disorder. An example of the difficulty is provided by the thalidomide disaster between 1959 and 1962. Only after numerous deformed babies had been born throughout the world did it become clear that thalidomide taken by the mother as a sedative had been responsible.

Hospital ethical committees

In hospitals where clinical research is carried out, ethical committees often consider each research project. If the committee believes that the risks are not justified, the project is rejected.

After a potentially useful chemical compound has been identified in the laboratory, it is extensively tested in animals, usually for a period of months or even years. Few drugs make it beyond this point. If the tests are satisfactory, however, the decision may be made for testing the drug in humans. It is this activity that forms the basis of much clinical research. In most countries the first step is the study of its effects in a small number of health vol-

unteers. The response, effect on metabolism, and possible toxicity are carefully monitored and have to be completely satisfactory before the drug can be passed for further studies, namely with patients who have the disorder for which the drug is to be used. Tests are administered at first to a limited number of these patients to determine effectiveness, proper dosage, and possible adverse reactions. These searching studies are scrupulously controlled under stringent conditions. Larger groups of patients are subsequently involved to gain a wider sampling of the information. Finally, a full-scale clinical trial is set up. If the regulatory authority is satisfied about the drug's quality, safety, and efficacy, it receives a license to be produced. As the drug becomes more widely used, it eventually finds its proper place in therapeutic practice, a process that may take years.

The controlled clinical trial

An important step forward in clinical research was taken in the mid-20th century with the development of the controlled clinical trial. This sets out to compare two groups of patients, one of which has had some form of treatment that the other group has not. The testing of a new drug is a case in point: one group receives the drug, the other a product identical in appearance, but which is known to be inert—a so-called placebo. At the end of the trial, the results of which can be assessed in various ways, it can be determined whether or not the drug is effective and safe. By the same technique two treatments can be compared, for example a new drug against a more familiar one. Because individuals differ physiologically and psychologically, the allocation of patients between the two groups must be made in a random fashion; some method independent of human choice must be used so that such differences are distributed equally between the two groups.

In order to reduce bias and make the trial as objective as possible the double-blind technique is sometimes used. In this procedure, neither the doctor nor the patients know which of two treatments is being given. Despite such precautions the results of such trials can be prejudiced, so that rigorous statistical analysis is required. It is obvious that many ethical, not to say legal, considerations arise, and it is essential that all patients have given their informed consent to be included. Difficulties arise when patients are unconscious, mentally confused, or otherwise unable to give their informed consent. Children present a special difficulty because not all laws agree that parents can legally commit a child to an experimental procedure. Trials, and indeed all forms of clinical research that involve patients, must often be submitted to a committee set up locally to scrutinize each proposal.

Surgery. In drug research the essential steps are taken by the chemists who synthesize or isolate new drugs in the laboratory; clinicians play only a subsidiary part. In developing new surgical operations clinicians play a more important role, though laboratory scientists and others in the background may also contribute largely. Many new operations have been made possible by advances in anesthesia, and these in turn depend upon engineers who have devised machines and chemists who have produced new drugs. Other operations are made possible by new materials, such as the alloys and plastics that are used to make artificial hip and knee joints.

Whenever practicable, new operations are tried on animals before they are tried on patients. This practice is particularly relevant to organ transplants. Surgeons themselves—not experimental physiologists—transplanted kidneys, livers, and hearts in animals before attempting these procedures on patients. Experiments on animals are of limited value, however, because animals do not suffer from all of the same maladies as do humans.

Animal experimentation

Many other developments in modern surgical treatment rest on a firm basis of experimentation, often first in animals but also in humans; among them are renal dialysis (the artificial kidney), arterial bypass operations, embryo implantation, and exchange transfusions. These treatments are but a few of the more dramatic of a large range of therapeutic measures that have not only provided patients with new therapies but also have led to the acquisition of new knowledge of how the body works. Among the research projects of the late 20th century is that of gene transplantation, which has the potential of providing cures for cancer and other diseases.

SCREENING PROCEDURES

Developments in modern medical science have made it possible to detect morbid conditions before a person actually feels the effects of the condition. Examples are many: they include certain forms of cancer; high blood pressure; heart and lung disease; various familial and congenital conditions; disorders of metabolism, like diabetes; and acquired immune deficiency syndrome (AIDS). The consideration to be made in screening is whether or not such potential patients should be identified by periodic examinations. To do so is to imply that the subjects should be made aware of their condition and, second, that there are effective measures that can be taken to prevent their condition, if they test positive, from worsening. Such so-called specific screening procedures are costly since they involve large numbers of people. Screening may lead to a change in the life-style of many persons, but not all such moves have been shown in the long run to be fully effective. Although screening clinics may not be run by doctors, they are a factor of increasing importance in the preventive health service.

Periodic general medical examination of various sections of the population, business executives for example, is another way of identifying risk factors that, if not corrected, can lead to the development of overt disease.

(J.W.T./Ha.Sc.)

Medical education

Goals of medical education

Medical education is directed toward imparting to persons seeking to become physicians the knowledge and skills required for the prevention and treatment of disease and also for developing the methods and objectives appropriate to the study of the still unknown factors that produce disease or favour well-being. Among the goals of medical education is the production of physicians sensitive to the health needs of their country, capable of ministering to those needs, and aware of the necessity of continuing their own education. It therefore follows that the plan of education, the medical curriculum, should not be the same in all countries. Although there may be basic elements common to all, the details should vary from place to place and from time to time. Whatever form the curriculum takes, ideally it will be flexible enough to allow modification as circumstances alter, medical knowledge grows, and needs change.

Attention in this article is focused primarily on general medical education.

HISTORY OF MEDICAL EDUCATION

Although it is difficult to identify the origin of medical education, authorities usually consider that it began with the ancient Greeks' method of rational inquiry, which introduced the practice of observation and reasoning regarding disease. Rational interpretation and discussion, it is theorized, led to teaching and thus to the formation of schools such as that at Cos, where the Greek physician Hippocrates is said to have taught in the 5th century BC and originated the oath that became a credo for practitioners through the ages.

Later, the Christian religion greatly contributed to both the learning and the teaching of medicine in the West because it favoured not only the protection and care of the sick but also the establishment of institutions where collections of sick people encouraged observation, analysis, and discussion among physicians by furnishing opportunities for comparison. Apprenticeship training in monastic infirmaries and hospitals dominated medical education during the early Middle Ages. A medical school in anything like its present form, however, did not evolve until the establishment of the one at Salerno in southern Italy between the 9th and 11th centuries. Even there teaching was by the apprentice system, but an attempt was made at systemization of the knowledge of the time, a series of health precepts was drawn up, and a form of registration to practice was approved by the Holy Roman emperor

Frederick II. During the same period, medicine and medical education were flourishing in the Muslim world at such centres as Baghdad, Cairo, and Córdoba.

With the rise of the universities in Italy and later in Cracow, Prague, Paris, Oxford, and elsewhere in western Europe, the teachers of medicine were in some measure drawn away from the life of the hospitals and were offered the attractions and prestige of university professorships and lectureships. As a result, the study of medicine led more often to a familiarity with theories about disease than with actual sick persons. However, the establishment in 1518 of the Royal College of Physicians of London, which came about largely through the energies of Thomas Linacre, produced a system that called for examination of medical practitioners. The discovery of the circulation of the blood by William Harvey provided a stimulus to the scientific study of the processes of the body, bringing some deemphasis to the tradition of theory and doctrine.

The Royal College of Physicians

Gradually, in the 17th and 18th centuries, the value of hospital experience and the training of the students' sight, hearing, and touch in studying disease were reasserted. In Europe, medical education began slowly to assume its modern character in the application of an increasing knowledge of natural science to the actual care of patients. There was also encouragement of the systematic study of anatomy, botany, and chemistry, sciences at that time considered to be the basis of medicine. The return to the bedside aided the hospitals in their long evolution from dwelling places of the poor, the diseased, and the infirm, maintained by charity and staffed usually by religious orders, into relatively well-equipped, well-staffed, efficient establishments that became available to the entire community and were maintained by private or public expense.

It was not until the mid-19th century, however, that an ordered pattern of science-oriented teaching was established. This pattern, the traditional medical curriculum, was generally adopted by Western medical schools. It was based upon teaching, where the student mostly listens, rather than learning, where the student is more investigative. The clinical component, largely confined to hospitals (charitable institutions staffed by unpaid consultants), was not well organized. The new direction in medical education was aided in Britain by the passage of the Medical Act of 1858, which has been termed the most important event in British medicine. It established the General Medical Council, which thenceforth controlled admission to the medical register and thus had great powers over medical education and examinations. Further interest in medicine grew from these advances, which opened the way for the discoveries of Louis Pasteur, which showed the relation of microorganisms to certain diseases, Joseph Lister's application of Pasteur's concepts to surgery, and the studies of Rudolf Virchow and Robert Koch in cellular pathology and bacteriology.

Advances at Johns Hopkins Medical School

In the United States, medical education was greatly influenced by the example set in 1893 by the Johns Hopkins Medical School in Baltimore. It admitted only college graduates with a year's training in the natural sciences. Its clinical work was superior because the school was supplemented by the Johns Hopkins Hospital, created expressly for teaching and research carried on by members of the medical faculty. The adequacy of medical schools in the United States was improved after the Carnegie Foundation for the Advancement of Teaching published in 1910 a report by the educator Abraham Flexner. In the report, which had an immediate impact, he pointed out that medical education actually is a form of education rather than a mysterious process of professional initiation or apprenticeship. As such, it needs an academic staff, working full-time in their departments, whose whole responsibility is to their professed subject and to the students studying it. Medical education, the report further stated, needs laboratories, libraries, teaching rooms, and ready access to a large hospital, the administration of which should reflect the presence and influence of the academic staff. Thus the nature of the teaching hospital was also influenced. Aided by the General Education Board, the Rockefeller Foundation, and a large number of private donors, U.S. and Canadian medical education was characterized by substantial improvements from 1913 to 1929 in such matters as were stressed in the Flexner report.

MODERN PATTERNS OF MEDICAL EDUCATION

As medical education developed after the Flexner report was published, the distinctive feature was the thoroughness with which theoretical and scientific knowledge were fused with what experience teaches in the practical responsibility of taking care of human beings. Medical education eventually developed into a process that involved four generally recognized stages: premedical, undergraduate, postgraduate, and continuing education.

Premedical education and admission to medical school. In the United States, Britain, and the Commonwealth countries, generally, medical schools are inclined to limit the number of students admitted so as to increase the opportunities for each student. In western Europe, South America, and most other countries, no exact limitation of numbers of students is in effect, though there is a trend toward such limitation in some of the western European schools. Some medical schools in North America have developed ratios of teaching staff to students as high as 1 to 1 or 1 to 2, in contrast with 1 teacher to 20 or even 100 students in certain universities in other countries. The number of students applying to medical school greatly exceeds the number finally selected in most countries.

Limitation of number of students

Requirements to enter medical school, of course, vary from country to country, and in some countries, such as the United States, from university to university. Generally speaking, in Western universities, there is a requirement for a specified number of years of undergraduate work and passing of a test, possibly state regulated, and a transcript of grades. In the United States entry into medical school is highly competitive, especially in the more prestigious universities. Stanford University, for instance, accepts only about 5 percent of its applicants. Most U.S. schools require the applicant to take the Medical College Admission Test, which measures aptitude in medically related subjects. Other requirements may include letters of recommendation and a personal interview. Many U.S. institutions require a bachelor's degree or its equivalent from an undergraduate school. A specific minimum grade point average is not required, but most students entering medical school have between an A and a B average.

The premedical courses required in most countries emphasize physics, chemistry, and biology. These are required in order to make it possible to present subsequently courses in anatomy, physiology, biochemistry, and pharmacology with precision and economy of time to students prepared in scientific method and content. Each of the required courses includes laboratory periods throughout the full academic year. Student familiarity with the use of instruments and laboratory procedures tends to vary widely from country to country, however.

Undergraduate education. The medical curriculum also varies from country to country. Most U.S. curriculums cover four years; in Britain five years is normal. The early part of the medical school program is sometimes called the preclinical phase. Medical schools usually begin their work with the study of the structure of the body and its formation: anatomy, histology, and embryology. Concurrently, or soon thereafter, come studies related to function—*i.e.*, physiology, biochemistry, pharmacology, and, in many schools, biophysics. After the microscopic study of normal tissues (histology) has begun, the student is usually introduced to pathological anatomy, bacteriology, immunology, parasitology—in short, to the agents of disease and the changes that they cause in the structure and function of the tissues. Courses in medical psychology, biostatistics, public health, alcoholism, biomedical engineering, emergency medicine, ethical problems, and other less traditional courses are becoming more common in the first years of the medical curriculum.

The two or more clinical years of an effective curriculum are characterized by active student participation in small group conferences and discussions, a decrease in the number of formal lectures, and an increase in the amount of contact with patients in teaching hospitals and clinics.

The clinical years

Clinical work begins with general medicine and surgery

and goes on to include the major clinical specialties, including obstetrics and gynecology, pediatrics, disorders of the eye, ear, nose, throat, and skin, and psychiatry. The student works in the hospital's outpatient and emergency departments, the radiology department, diagnostic laboratories, and the surgical theatres. He also studies sciences closely related to medicine, such as pathology, microbiology, hematology, immunology, and clinical chemistry and becomes familiar with epidemiology and the methods of community medicine. Some knowledge of forensic (legal) medicine is also expected. During the clinical curriculum many students have an opportunity to pursue a particular interest of their own or to enlarge their clinical experience by working in a different environment, perhaps even in a foreign country—the so-called elective period. Most students find clinical work demanding, usually requiring long hours of continuous duty and personal commitment.

In the United States after satisfactory completion of a course of study in an accredited medical school the degree of doctor of medicine (M.D.) or doctor of osteopathy (D.O.) is conferred. In Britain and some of the other Commonwealth countries the academic degree conferred after undergraduate studies are completed is bachelor of medicine and of surgery (or chirurgery), M.B., B.S. or M.B., CHb. Only after further study is the M.D. degree given. Similar degrees are conferred in other countries, although they are not always of the same status.

Postgraduate education. On completion of medical school, the physician usually seeks graduate training and experience in a hospital under the supervision of competent clinicians and other teachers. In Britain a year of resident hospital work is required after qualification and before admission to the medical register. In North America, the first year of such training has been known as an internship, but it is no longer distinguished in most hospitals from the total postgraduate period, called residency. After the first year physicians usually seek further graduate education and training to qualify themselves as specialists or to fulfill requirements for a higher academic degree. Physicians seeking special postgraduate degrees are sometimes called fellows.

Internship and residency

Continuing education. The process by which physicians keep themselves up-to-date is called continuing education. It consists of courses and training opportunities of from a few days to several months in duration, designed to enable physicians to learn of new developments within their special areas of concern. Physicians also attend medical and scientific meetings, national and international conferences, discussion groups, and clinical meetings, and they read medical journals and other materials, all of which serve to keep them aware of progress in their chosen field. Although continuing education is not a formal process, organizations designed to promote continuing education have become common. In the United States the Accreditation Council for Continuing Medical Education was formed in 1985, and some certifying boards of medical specialties have stringent requirements for continuing education.

The quality of medical education is supervised in many countries by councils appointed by the profession as a whole. In the United States these include the Council on Medical Education and the Liaison Committee on Medical Education, both affiliates of the American Medical Association, and the American Osteopathic Association. In Britain the statutory body is the General Medical Council, most of whose members are from the profession, although only a minority of the members are appointed by it. In other countries medical education may be regulated by an office or ministry of public instruction with, in some cases, the help of special professional councils.

Medical school faculty. As applied to clinical teachers the term full-time originally implied an educational ideal: that a clinician's salary from a university should be large enough to relieve him of any reason for seeing private patients for the sake of supplementing his salary by professional fees. Full-time came to be applied, however, to a variety of modifications; it could mean that a clinical professor might supplement his salary as a teacher up to a defined maximum, might see private patients only at his hospital office, or might see such patients only a certain number of hours per week. The intent of full-time has always been to place the teacher's capacities and strength entirely at the service of his students and the patients entrusted to his care as a teacher and investigator.

Courses in the medical sciences have commonly followed the formula of three hours of lectures and six to nine hours of laboratory work per week for a three-, six-, or nine-month course. Instruction in clinical subjects, though retaining the formal lecture, have tended to diminish the time and emphasis allowed to lectures in favour of experience with and attendance on patients. Nonetheless, the level of lecturing and formal presentation remains high in some countries.

REQUIREMENTS FOR PRACTICE

Graduation from medical school and postgraduate work does not always allow the physician to practice. In the United States, licensure to practice medicine is controlled by boards of licensure in each state. The boards set and conduct examinations of applicants to practice within the state, and they examine the credentials of applicants who want licenses earned in other states to be accepted in lieu of examination. The National Board of Medical Examiners holds examinations leading to a degree that is acceptable to most state boards. National laws regulating professional practice cannot be enacted in the United States. In Canada, the Medical Council of Canada conducts examinations and enrolls successful candidates on the Canadian medical register, which the provincial governments accept as the main requirement for licensure. In Britain, the medical register is kept by the General Medical Council, which supervises the licensing bodies; unregistered practice, however, is not illegal. In some European countries graduation from a state-controlled university or medical school in effect serves as a license to practice; the same is true for the Soviet Union and Japan. (For further details of licensure, see below *Legal aspects of medicine.*)

Licensure to practice

ECONOMIC ASPECTS

The income of a medical school is derived from four principal sources: (1) tuition and fees; (2) endowment income or appropriation from the government (taxation); (3) gifts from private sources; and (4) donation of teachers' services. Tuition or student fees are large in most English-speaking countries (except in U.S. state universities) and relatively small throughout the rest of the world. Tuition in most U.S. schools, however, rarely makes up more than a small part of total operating expenses. The total cost of maintaining a medical school, if prorated among the students, would produce a figure many times greater than the tuition or other charges paid by each student. The costs of operating medical schools in the United States increased by about 30 times between the late 1950s and the mid-1980s.

Sources of medical school income

The expenses of medical education fall into two groups: those of the instruction given in the medical sciences and those connected with hospital teaching. In the medical sciences, costs of building maintenance, laboratory equipment and supplies, research expenses, salaries of teachers, and wages of employees are heavy but comparable to those in other departments of a university. In the clinical subjects all expenses in connection with the care of patients usually are considered as hospital expenses and are not carried on the medical school budget, which is normally reserved for the expenses of teaching and research. Here the heavy expenses are salaries of clinical teachers and the cost of studying cases of illness with a thoroughness appropriate to their use as teaching material.

To a considerable degree in free-market countries, the cost of securing an adequate medical education has tended to exclude the student whose family cannot contribute a large share of his tuition and living expenses for four to 10 years. This difficulty is offset in some medical schools by loan funds and scholarships, but these aids are commonly offered only in the second or subsequent years. In Britain, scholarships and maintenance grants are available through state and local educational authority funds, so that an individual can secure a medical education even though the parents may not be able to afford its cost.

SCIENTIFIC AND INTERNATIONAL ASPECTS

Medical education has the double task of passing on to students what is known and of attacking what is still unknown. The cost of medical research is borne by only a few; the benefits are shared by many. There are countries whose citizens are too poor to support physicians or to use them; countries that can support a few physicians but are too poor to maintain a good medical school; countries that can maintain medical schools where what is known can be taught but where no research can be carried out; and a few countries in which teaching and research in medicine can be carried on to the great advantage of the world at large.

A medical school having close geographical as well as administrative relationships with the rest of the university of which it forms a part usually profits by this intimate and easy contact. Medicine cannot wisely be separated from the biological sciences, and it continues to gain immensely from chemistry, physics, mathematics, and psychology, as well as from modern technology. The social sciences contribute by making physicians aware of the need for better distribution of medical care. Contact with teachers and the advancing knowledge in other faculties also may have a corollary effect in advancing medicine.

WHO

With the development of the World Health Organization (WHO) and the World Medical Association after World War II, there has been increasing international interest in medical education. WHO conducts a regular program for aiding countries in the development and expansion of their educational facilities. World War II showed the advantages and economy derived from satisfactory systems of medical education: defects and diseases were more widely and accurately detected among recruits than ever before, health and morale were effectively maintained among combatants, and disease and battle injuries were effectively treated. (Al.Gr./E.L.T./Ha.Sc.)

MAJOR MEDICAL INSTITUTIONS

Hospitals

A hospital is an institution that is built, staffed, and equipped for the identification (diagnosis) of disease; for the treatment, both medical and surgical, of the sick and the injured; and for their housing during this process. The modern hospital also often serves as a centre for investigation and for teaching. To better serve the wide-ranging needs of the community, the modern hospital has often developed outpatient facilities, as well as emergency, psychiatric, and rehabilitation services.

Hospitals have long existed in every civilized country. The developing countries, which contain a large proportion of the world's population, do not have enough hospitals, equipment, and trained staff, and, by the standards of the industrialized countries, the hospitals that do exist are poorly equipped to handle the volume of persons who need care. These persons, then, do not always receive the benefits of modern medicine, public-health measures, or hospital care, and they generally have lower life expectancies.

In the developed countries, the hospital as an institution is becoming more complex as modern technology increases the range of diagnostic capabilities and expands the possibilities for treatment. As a result of the greater range of services and the more involved treatment and surgery available, the ratio of staff to patient has increased and a more highly trained staff is required. During recent years a combination of medicine and engineering has produced a vast array of new instrumentation, much of which requires a hospital setting for its use. Hospitals thus are becoming more expensive to run and health service administrators are increasingly concerned with the question of cost-effectiveness.

HISTORY OF HOSPITALS

Hospitals in antiquity

As early as 4000 BC religions identified certain of their deities with healing. The temples of Saturn, and later of Asclepius in Asia Minor, were recognized as healing centres. Brahmanic hospitals were established in Sri Lanka as early as 431 BC, and King Aśoka established a chain of hospitals in Hindustān about 230 BC. Around 100 BC the Romans established hospitals (*valetudinaria*) for the treatment of their sick and injured soldiers; their care was important because it was upon the integrity of the legions that the power of Rome was based.

It can be said, however, that the modern concept of a hospital dates from AD 331 when Constantine, having been converted to Christianity, abolished all pagan hospitals and thus created the opportunity for a new start. Until that time disease had isolated the sufferer from the community. The Christian tradition emphasized the close relationship of the sufferer to his fellow man, upon whom rested the obligation for care. Illness thus became a matter for the Christian church.

Around AD 370 St. Basil of Caesarea established a religious foundation in Cappadocia that included a hospital, an isolation unit for those suffering from leprosy, and buildings to house the poor, the elderly, and the sick. Following this example similar hospitals were later built in the eastern part of the Roman Empire. Another notable foundation was that of St. Benedict at Monte Cassino, founded early in the 6th century, where the care of the sick was placed above and before every other Christian duty. It was from this beginning that one of the first medical schools in Europe ultimately grew at Salerno and was of high repute by the 11th century. This example led to the establishment of similar monastic infirmaries in the western part of the empire.

The Hôtel-Dieu of Lyon was opened in 542 and the Hôtel-Dieu of Paris in 660. In these hospitals more attention was given to the well-being of the patient's soul than to curing bodily ailments. The manner in which monks cared for their own sick became a model for the laity. The monasteries had an *infirmitorium,* a place to which their sick were taken for treatment. The monasteries had a pharmacy and frequently a garden with medicinal plants. In addition to caring for sick monks, the monasteries opened their doors to pilgrims and to other travelers.

Religion continued to be the dominant influence in the establishment of hospitals during the Middle Ages. The growth of hospitals accelerated during the Crusades, which began at the end of the 11th century. Pestilence and disease were more potent enemies than the Saracens in defeating the crusaders. Military hospitals came into being along the traveled routes; the Knights Hospitalers of the Order of St. John in 1099 established in the Holy Land a hospital that could care for some 2,000 patients. It is said to have been especially concerned with eye disease, and may have been the first of the specialized hospitals. This order has survived through the centuries as the St. John's Ambulance Corps.

Throughout the Middle Ages, but notably in the 12th century, the number of hospitals grew rapidly in Europe. The Arabs established hospitals in Baghdad and Damascus and in Córdoba in Spain. Arab hospitals were notable for the fact that they admitted patients regardless of religious belief, race, or social order. The Hospital of the Holy Ghost, founded in 1145 at Montpellier in France, established a high reputation and later became one of the most important centres in Europe for the training of doctors. By far the greater number of hospitals established during the Middle Ages, however, were monastic institutions under the Benedictines, who are credited with having founded more than 2,000.

The Middle Ages also saw the beginnings of support for hospital-like institutions by secular authorities. Toward the end of the 15th century many cities and towns supported some kind of institutional health care: it has been said that in England there were no less than 200 such establishments that met a growing social need. This gradual transfer of responsibility for institutional health care

from the church to civil authorities continued in Europe after the dissolution of the monasteries in 1540 by Henry VIII, which put an end to hospital building in England for some 200 years.

The loss of monastic hospitals in England caused the secular authorities to provide for the sick, the injured, and the handicapped, thus laying the foundation for the voluntary hospital movement. The first voluntary hospital in England was probably established in 1718 by Huguenots from France and was closely followed by the foundation of such London hospitals as the Westminster Hospital in 1719, Guy's Hospital in 1724, and the London Hospital in 1740. Between 1736 and 1787 hospitals were established outside London in at least 18 cities. The initiative spread to Scotland where the first voluntary hospital, the Little Hospital, was opened in Edinburgh in 1729.

Early hospitals in North America

The first hospital in North America was built in Mexico City in 1524 by Cortés; the structure still stands. The French established a hospital in Canada in 1639 at Quebec city, the Hôtel-Dieu du Précieux Sang, which is still in operation although not at its original location. In 1644 Jeanne Mance, a French noblewoman, built a hospital of ax-hewn logs on the island of Montreal; this was the beginning of the Hôtel-Dieu de St. Joseph, out of which grew the order of the Sisters of St. Joseph, now considered to be the oldest nursing group organized in North America. The first hospital in the territory of the present-day United States is said to have been a hospital for soldiers on Manhattan Island, established in 1663.

The early hospitals were primarily almshouses, one of the first of which was established by William Penn in Philadelphia in 1713. The first incorporated hospital in America was the Pennsylvania Hospital, in Philadelphia, which obtained a charter from the crown in 1751.

THE MODERN HOSPITAL

Hospitals may be compared and classified in various ways; by ownership and control; by type of service rendered; by length of stay; by size; or by facilities and organization provided. Terms in general use include the general as distinct from the special hospital; the short-stay hospital; and the long-term hospital.

Indexes of comparison

Comparison. Hospitals may be compared by the number of beds they contain. Modern hospitals tend to be small, rarely exceeding 800 beds, which is thought to be the largest number that can be governed satisfactorily from a single administrative unit, yet not too large to retain a corporate unity.

Another index is the average bed-occupancy, that is, the percentage of available beds actually occupied per day or per month. In Europe bed-occupancy may be higher in the cold winter months, which bring more respiratory disease. In developing countries the bed-occupancy is often more than 100 percent: that is to say there are more patients in the hospital than there are beds for them.

The amount of time that a patient spends in a hospital bed, the bed-stay, is another important index and depends on the nature of the hospital. In an acute-care hospital the bed-occupancy will be low. In hospitals catering to the more chronically ill, the average bed-stay probably will be higher. There are even significant variations between units in the same hospital doing the same kind of work. In hospitals in developing countries, the average bed-stay is much shorter than in Europe.

Ownership, control, and financing. *Ownership and control.* In most countries outside of North America nearly all hospitals are owned and operated by the government. In Great Britain, except for a small number run by religious orders or serving special groups, most hospitals are within the National Health Service. The local hospital management committee answers directly to the regional hospital board and ultimately to the Department of Health and Social Security. In the United States and Canada most hospitals are nonprofit and are neither owned nor operated by governmental agencies. Many of them are associated with universities; others were founded by religious groups or by public-spirited individuals. Mental hospitals traditionally have been the responsibility of the state governments, while military and veterans hospitals have been provided by the federal government. In addition, there are a number of municipal and county general hospitals.

Financing. Almost universally, hospital construction costs are met at least in some part by governmental contributions. Operating costs are taken care of in a variety of ways. Ultimately, a substantial portion of the expenses not covered by private endowments or gifts is met by contributions from the general funds of some unit of government or out of funds collected by insurance carriers from subscribers. In countries in which hospital insurance is not universal or its coverage complete, some of the operating costs are met by charges on uninsured or inadequately insured patients.

Carriers of hospital insurance

The carriers of the hospital insurance in a particular country may be governmental agencies, private corporations or agencies, or both. In Britain, for example, under the National Insurance Act, the government is the carrier. All persons who have reached the minimum school-leaving age and are not full-time students, beyond the age of retirement, in prison, or receivers of benefits from the insurance and who do not have less than a certain minimum income are contributors under the plan whether employed by others, self-employed, or nonemployed. Employers also contribute.

In the United States persons who are employed by others or are self-employed make compulsory contributions toward a form of national hospital insurance, Medicare, which pays a large portion of the hospital costs of persons at the age of 65 or over. Employers then make matching payments. A majority of the persons ineligible, by reason of age, for benefits under the Medicare program are enrolled in some other form of hospital insurance, such as one of the plans offered by the commercial insurance companies or one of the independent plans, including community and community-controlled plans and those operated by unions, employers, welfare funds, and private medical clinics.

In the United States, even with federal participation under Medicare and Medicaid (a program for persons under 65 who are unable to pay), the payment for health care services on an insurance basis, either voluntary or governmental, is considerably less advanced than it is in many other parts of the world. In Europe, particularly, the financial support of services in hospitals tends to be much more collectivized. Less than 10 percent of the costs of hospital operation in Europe is covered by payments made directly by patients. Details vary somewhat from country to country: in the Soviet Union, for example, the program is based entirely on funds from a single source, while in Great Britain the funds for total hospital operation are appropriated by the Ministry of Social Security to each regional hospital board, which in turn distributes them to the local hospital groups. In Sweden approximately 90 percent of hospital operating costs are provided by local or provincial units of government from public revenue; the remaining 10 percent comes from payments made by insurance funds on behalf of the patient. In general, in the majority of European countries, hospital operating costs are paid out of insurance funds; such is the case in France, West Germany, Italy, The Netherlands, Norway, and elsewhere.

The community general hospital

The general hospital. General hospitals are general in the sense that they admit all types of medical and surgical cases and they concentrate on patients with acute illness needing relatively short-term care. A community general hospital with about 200 beds has an organized medical staff, a professional nursing staff, and diagnostic equipment. In addition to the essential services relating to patient care, it has a pharmacy, a laboratory, X-ray and physical therapy departments, possibly a maternity division (ordinarily including a nursery and a delivery room), operating rooms, recovery rooms, an outpatient department, and an emergency department.

In a somewhat larger hospital there may be additional facilities: dental services; a nursery for premature infants; an organ bank for use in transplantation; a department of renal dialysis (removal of wastes from the blood by passing it through semipermeable membranes, as in the artificial kidney); equipment for inhalation therapy; an intensive-

care unit; a volunteer-services department; and, possibly, a home-care program. The complexity of the general hospital, then, reflects the advances made after World War II, including the use of antibiotics, a vast new array of laboratory procedures, new surgical techniques, new materials and equipment for radiation therapy, and an increased emphasis on physical therapy and rehabilitation.

The legally constituted governing body of the hospital, with full responsibility for the conduct and efficient management of the hospital, is usually a hospital board. The board establishes policy and, on the advice of a medical advisory board, appoints a medical staff and an administrator. It exercises control over expenditures and has the responsibility for maintaining professional standards.

The administrator is the chief executive officer and is responsible to the board. In a large hospital there are many separate departments, each of which is controlled by a department head. The largest department in any hospital is nursing, followed by the dietary department and housekeeping. Other departments that are important to the functioning of the hospital include laundry, engineering, stores, purchasing, accounting, pharmacy, physical therapy, social service, pathology, X-ray, and medical records.

Organization of medical staff

The medical staff is organized into such departments as surgery, medicine, obstetrics, and pediatrics. The degree of departmentalization of the medical staff depends on the specialization of its members and not primarily on the size of the hospital, although there is usually some correlation between the two. The chiefs of the medical-staff departments, along with the radiologist and the pathologist, make up the medical advisory board, which usually holds monthly meetings on medical–administrative matters. The professional work of the individual staff members is reviewed by such medical-staff committees. In a large hospital the committees may report to the medical advisory board; in a smaller hospital, to the medical staff directly, at regular staff meetings.

General hospitals often also have a formal or informal role as teaching institutions. When formally designed as such, teaching hospitals are affiliated with undergraduate and postgraduate medical education at a university, and they provide up-to-date and often specialized therapeutic measures and facilities unavailable elsewhere in the region. As teaching hospitals have become more specialized, general hospitals have become more involved in providing general clinical training to medical students.

Specialized health- and medical-care facilities. Hospitals that specialize in one type of illness or one type of patient can be found in Europe and in North America, although, except in large university centres where postgraduate teaching is carried out on a large scale, the special hospital is increasingly becoming a department of the general hospital. Changing conditions or modes of treatment have lessened the need or reduced the number of some types of specialized institutions; this may be seen in the cases of tuberculosis, leprosy, and mental hospitals.

Tuberculosis and leprosy hospitals. Between 1880 and 1940 tuberculosis hospitals provided rest, relaxation, special diets, and fresh air, and even if the tuberculosis was in an early stage, a stay of more than two years was thought necessary to effect a healing of the disease; a permanent cure was not considered entirely feasible. Today the use of antibiotics, along with advances in chest surgery and routine X-ray programs, has meant that the treatment of tuberculosis need not be carried out in a specialized facility.

Leprosy has been known for centuries to be contagious. Lazar houses were established throughout Europe in the Middle Ages to isolate those with leprosy, at that time a common disease, from the community. In the 14th century there may have been some 7,000 leper houses in France alone, and some of the earliest hospitals in England were established for lepers. Although it is now rare in Europe, leprosy is still common in many parts of the world. The purpose of the modern leprosarium is not so much isolation as it is treatment. The chronic form of the disease is treated by surgical correction of deformities, occupational therapy, rehabilitation, and sheltered living in associated villages. Acute leprosy is treated in general hospitals, clinics, and dispensaries.

Mental hospitals. Psychiatric patients traditionally have been housed in long-stay mental hospitals, formerly called asylums, although the majority of large general hospitals now have a psychiatric unit. In the past few decades, the hospital stay of many persons with chronic mental disease has been shortened by modern medication and better understanding on the part of the public. Mental patients often may participate in many activities, first within the hospital setting and later in the community, either with trial visits at home or with placement of selected patients in foster homes. Every effort is now made with appropriate medication and the judicious use of the support services to get the patient home and in the care of the family. Even those mentally handicapped persons who require custodial care are no longer isolated from contact with their relatives and friends.

Changes in treatment of mental patients

Long-stay institutions. Historically the long-stay institution was a place for the elderly, the infirm, and those with chronic irreversible and disabling disorders, especially if the patients were indigent. Medical and nursing care was minimal. Today the long-stay hospital has a more active role in health care. Many are well staffed and well equipped to prepare a patient as much as possible to live at home or with a member of the family. Long-stay hospitals represent a significant extension of the hospital health care system, helping to conserve expensive facilities for the acutely ill and improving the prospects of the chronically disabled.

Private hospitals. Throughout Europe, as well as in North America, Australia, and New Zealand, there are small private hospitals, often called nursing homes in Britain, many of which until recently were able to provide little more than accommodation and simple nursing, the patient being under the care of his general practitioner or of a visiting consultant. Medical practice in the towns of developing countries is characterized by a proliferation of many small private hospitals, usually owned by doctors, that have developed to meet the widespread need for hospital care not otherwise available.

Another method of providing health care in a hospital for those able to pay for it, both in industrial and developing countries, is the provision of a limited number of beds for private patients within a large general hospital otherwise financed to some degree by public funds. In the United Kingdom and, for example, in West Africa, these so-called amenity beds usually form part of the ward unit, the patient being required to pay for certain amenities such as a measure of privacy, unrestricted visiting, attractively served food, and a more liberal routine. Alternatively, many large general hospitals are able to offer much more costly accommodations in so-called private blocks, that is to say in a part of the hospital specially designed and equipped for private patients. Patients in a private block pay a large portion of the total cost of their medical care, including that of surgery.

Private patients

More recent is the development of the wholly independent private hospital run by a company or business consortium. Many of these privately funded hospitals are able to provide most or all of the services of a general hospital, including constant medical care and a first-rate nursing service, although such facilities are costly.

The hospice. Historically, a hospice was a guest house intended for pilgrims, which normally would be closely connected with a monastery and supervised by monks. From the beginning it had a strong religious connection and exemplified the Christian insistence on compassion and care for the aged, the infirm, the needy, and the ill. In modern Britain, the hospice movement has developed gradually from its beginning in 1905 when the Sisters of Charity founded the St. James Hospice in London. The St. Christopher Hospice, also in London, founded in 1967, soon became known for its peaceful environment and expert medical and nursing care. The hospice movement has spread throughout Britain, North America, Australia, and Europe.

Mission hospitals. The spread of Western medicine and the establishment and founding of hospitals in the developing countries can be attributed in large part to the influence of the medical missionary. The establishment

of mission hospitals gained momentum gradually in the second half of the 19th century. By the second half of the 20th century, however, this steady growth was already dwindling since all but a few of the hospitals and dispensaries founded during that hundred years have been absorbed into the native health care system. The Christian missionaries had a great influence on creating centres of Western medicine in many developing countries and in promulgating the concept of a hospital in which health care would be centralized and organized for the benefit of the ill and injured, many of whom would not otherwise have survived. The medical missionaries also promoted the idea and the ideals of nursing as a profession for native men and women.

Apart from its religious associations, a mission hospital functions as a general hospital in the sense that it admits all who need hospital care. A number of mission hospitals, however, have been devoted to specific diseases; for example, leprosy and diseases of the eyes. Perhaps the most important contribution made by mission hospitals is in the enormous numbers of persons, particularly women and children, who have been treated as outpatients.

Extended health care. With the advance in medical science and the ever-increasing cost of hospital operations, the progressive-care concept is more attractive, both for outpatient and inpatient care. Progressive care can be divided into five categories: (1) intensive care; (2) intermediate care; (3) self-care; (4) long-term care; and (5) organized home-care programs. Two of the categories, self-care and home-care programs, are relatively new departures from past practice and deserve special attention.

Categories of progressive care

Self-care facilities are organized into a separate unit in which ambulatory patients who require only diagnostic or convalescent care are given accommodations similar to those of a hotel. The patients are free to wear street clothes and to go to the hospital cafeteria. Such a ward or wing of a general hospital requires much less costly equipment than the intensive- or intermediate-care units and can be staffed with far fewer nurses and aides.

Home-care programs are for patients who need some care but not all of the treatment facilities of a hospital. The patients are provided with a range of individualized medical, nursing, social, and rehabilitative services in their own homes, coordinated through one central agency. Patients can be considered ready for home care when the following criteria are met: (1) diagnosis and a plan for treatment have been established; (2) inpatient hospital facilities are no longer required for proper care; (3) no more than two visits per week by physicians are required; (4) the nursing service has found that the physical environment of the home is such that the patient receives adequate care; (5) the patient is too ill to visit an outpatient clinic but does not need hospital care; (6) the family environment would have a therapeutic effect, and family members or others can be taught to provide the necessary care; and (7) the family and the patient prefer that he be cared for at home. Even though home care conserves expensive acute-care beds, and although most patients on home care do as well as or better than expected, home-care programs have not been widely adopted.

Regional planning. Sweden and the Soviet Union provide examples of advanced planning in the integration of hospital networks into coordinated health services. In both nations the government has assumed the responsibility of providing health care to all citizens. In Sweden financing is in part by compulsory health insurance.

Sweden is divided into health-service regions; each region includes several counties and has a central hospital. Each county within a region has a county hospital with up to 1,000 beds and with specialized and outpatient facilities to serve a population of about 300,000. The counties, in turn, are divided into districts, each of which has a population of about 75,000 and is served by a district hospital, which usually has 300 or more beds. Smaller communities have health centres or ambulatory service centres that are not administered as part of the hospital system.

The Soviet Union takes a somewhat different approach. In its thinly populated rural areas, general hospitals, called *uchastok* hospitals, may serve populations as small as 2,000 to 15,000 persons. These 15- to 100-bed general hospitals occupy the same premises and employ the same staff as general clinics (polyclinics) that provide general and specialized care. The hospital–clinic staff includes a general physician, a surgeon, and a dental surgeon. Larger *uchastok* centres may also have a radiologist and a pathologist.

The next larger hospitals, the district hospitals, have 250–500 beds and usually have divisions for surgical, medical, obstetric, and pediatric services and provide care for infectious diseases; they may also include departments for eye, ear, nose, and throat disorders and for orthopedic surgery. Patients who cannot be treated adequately in the district hospitals are referred to the next higher level, the regional hospital, which serves a population of 1,000,000–5,000,000 people and contains up to 1,250 beds.

The republic hospital occupies the highest level in the Soviet system. Such a hospital, or complex of hospitals, serves as a referral centre and has the responsibility of undergraduate medical education. It may also be associated with one or more research institutes.

Regional planning in North America is less advanced. One regional pattern is a satellite system, centred on a metropolis and applying the principle of progressive patient care. The system is focused on the efficient provision of comprehensive health care to the residents of the region. Less serious cases are handled in the outer, more accessible health facilities of the system; the more serious are referred to the inner hospitals of the ring or to the research and teaching hospital at the core.

The term metropolitan planning council is often used to denote an advisory planning group that coordinates services among member hospitals in a metropolitan area and decides such questions as where more beds are to be added. In North America, however, most hospitals are not government operated, and it is difficult to achieve close cooperation among voluntary groups. (W.D.P./Ha.Sc.)

Metropolitan planning council

Public health services

Public health has been defined as the art and science of preventing disease, prolonging life, and promoting physical and mental health, sanitation, personal hygiene, control of infection, and organization of health services. There has emerged a recognition of the importance of community action in the promotion of health and the prevention and treatment of disease. This is expressed in the concept of public health.

Comparable terms for public health medicine are social medicine and community medicine; the latter has been widely adopted in the United Kingdom, and the practitioners are called community physicians. The practice of public health draws heavily on medical science and philosophy and concentrates especially on manipulating and controlling the environment for the benefit of the public. It is concerned therefore with housing, water supplies, and food. Noxious agents can be introduced into these through farming, fertilizers, inadequate sewage disposal and drainage, construction, defective heating and ventilating systems, machinery, and toxic chemicals. Public health medicine is part of the greater enterprise of preserving and improving the public health. Community physicians cooperate with such diverse groups as architects, builders, sanitary and heating and ventilating engineers, factory and food inspectors, psychologists and sociologists, chemists, physicists, and toxicologists. Occupational medicine is concerned with the health, safety, and welfare of persons in the workplace. It may be viewed as a specialized part of public health medicine since its aim is to reduce the risks in the environment in which persons work.

Concerns

The venture of preserving, maintaining, and actively promoting the health of the public requires special methods of information-gathering (epidemiology) and corporate arrangements to act upon significant findings and put them into practice. Statistics collected by epidemiologists attempt to describe and explain the occurrence of disease in a population by correlating factors such as diet, environment, radiation, or cigarette smoking with the incidence and prevalence of disease. The government, through laws

and regulations, is able to oversee and formally inspect such things as water supplies, food processing, sewage treatment, drains, air contamination, and pollution. Governments also are concerned with the control of epidemic infections by enforced quarantine and isolation; for example, the health control that takes place at sea- and airports in an attempt to assure that infectious diseases are not brought into a country.

This section traces the historical development of public health, beginning in ancient times and emphasizing how various public health concepts have evolved. It outlines the organizational and administrative methods of handling these problems in the developed and the developing countries of the world. Special attention is given to the developing countries and to how the health problems, limitations of resources, education of health personnel, and other factors must be taken into account in designing health-service systems. Finally, there are descriptions of the most recent developments in public health, together with some indications of the problems still to be solved.

HISTORY OF PUBLIC HEALTH

Beginnings in antiquity. Most of the world's primitive people have practiced cleanliness and personal hygiene, often for religious reasons, and, apparently, a wish to be pure in the eyes of their gods. The Old Testament, for example, has many adjurations and prohibitions about clean and unclean living. Religion, law, and custom were inextricably interwoven. For thousands of years primitive societies looked upon epidemics as divine judgments on the wickedness of mankind. The idea that pestilence is due to natural causes, such as climate and physical environment, however, gradually developed. This great advance in thought took place in Greece during the 5th and 4th centuries BC and represented the first attempt at a rational, scientific theory of disease causation. The association between malaria and swamps, for example, was established very early (503–403 BC), even though the reasons for the association were obscure. In the book *Airs, Waters, and Places,* thought to have been written by Hippocrates in the 5th or 4th century BC, the first systematic attempt was made to set forth a causal relationship between human diseases and the environment. Until the new sciences of bacteriology and immunology emerged well into the 19th century, this book provided a theoretical basis for the comprehension of endemic disease (that persisting in a particular locality) and epidemic disease (that affecting a number of people within a relatively short period).

Greek theories of disease causation

The Middle Ages. In terms of disease, the Middle Ages can be regarded as beginning with the plague of 542 and ending with the Black Death (bubonic plague) of 1348. Diseases in epidemic proportions included leprosy, bubonic plague, smallpox, tuberculosis, scabies, erysipelas, anthrax, trachoma, sweating sickness, and dancing mania (see INFECTIOUS DISEASES). The isolation of persons with communicable diseases first arose in response to the spread of leprosy. This disease became a serious problem in the Middle Ages and particularly in the 13th and 14th centuries.

The Black Death

The Black Death reached the shores of southern Europe from the Middle East in 1348 and in three years swept throughout Europe. The chief method of combating plague was to isolate known or suspected cases as well as persons who had been in contact with them. The period of isolation at first was about 14 days and gradually was increased to 40 days. Stirred by the Black Death, public officials created a system of sanitary control to combat contagious diseases, using observation stations, isolation hospitals, and disinfection procedures. Major efforts to improve sanitation included the development of pure water supplies, garbage and sewage disposal, and food inspection. These efforts were especially important in the cities, where people lived in crowded conditions in a rural manner with many animals around their homes.

During the Middle Ages a number of first steps in public health were made: attempts to cope with the unsanitary conditions of the cities and, by means of quarantine, to limit the spread of disease; the establishment of hospitals; and provision of medical care and social assistance.

The Renaissance. Centuries of technological advance culminated in the 16th and 17th centuries in a number of scientific accomplishments. Educated leaders of the time recognized that the political and economic strength of the state required that the population maintain good health. No national health policies were developed in England or on the Continent, however, because the government lacked the knowledge and administrative machinery to carry out such policies. As a result, public health problems continued to be handled on a local community basis, as they had been in medieval times.

Scientific advances of the 16th and 17th centuries laid the foundations of anatomy and physiology. Observation and classification made possible the more precise recognition of diseases. The idea that microscopic organisms might cause communicable diseases had begun to take shape.

Among the early pioneers in public health medicine was John Graunt, who in 1662 published a book of statistics, which had been compiled by parish and municipal councils, that gave numbers for deaths and sometimes suggested their causes. Inevitably the numbers were inaccurate but a start was made in epidemiology.

National developments in the 18th and 19th centuries. Nineteenth-century movements to improve sanitation occurred simultaneously in several European countries and were built upon foundations laid in the period between 1750 and 1830. From about 1750 the population of Europe increased rapidly, and with this increase came a heightened awareness of the large numbers of infant deaths and of the unsavoury conditions in prisons and in mental institutions.

This period also witnessed the beginning and the rapid growth of hospitals. Hospitals founded in Britain, as the result of voluntary efforts by private citizens, helped to create a pattern that was to become familiar in public health services. First, a social evil is recognized and studies or experiments are undertaken through individual initiative. These efforts mold public opinion and attract governmental attention. Finally, such agitation leads to governmental action.

This era was also characterized by efforts to educate people in health matters. In 1852 Sir John Pringle published a book that discussed ventilation in barracks and the provision of latrines. Two years earlier he had written about jail fever (now thought to be typhus), and again he emphasized the same needs as well as personal hygiene. In 1754 James Lind published a treatise on scurvy, a disease caused by a lack of vitamin C.

The health of industrial workers

As the Industrial Revolution developed, the health and welfare of the workers deteriorated. In England, where the Industrial Revolution and its bad effects on health were first experienced, there arose in the 19th century a movement toward sanitary reform that finally led to the establishment of public health institutions. Between 1801 and 1841 the population of London doubled; that of Leeds nearly tripled. With such growth there also came rising death rates. Between 1831 and 1844 the death rate per thousand increased in Birmingham from 14.6 to 27.2; in Bristol, from 16.9 to 31; and in Liverpool, from 21 to 34.8. These figures were the result of an increase in the urban population that far exceeded available housing and of the subsequent development of conditions that led to widespread disease and poor health.

Around the beginning of the 19th century humanitarians and philanthropists in England worked to educate the population and the government on problems associated with population growth, poverty, and epidemics. Thomas Malthus wrote in 1798 about population growth, its dependence on food supply, and the control of breeding by contraceptive methods. The utilitarian philosopher Jeremy Bentham propounded the idea of the greatest good of the greatest number as a yardstick against which the morality of certain actions might be judged. Thomas Southwood Smith founded the Health of Towns Association in 1839, and by 1848 he served as a member of the new government department, then called the General Board of Health. He published reports on quarantine, cholera, yellow fever, and the benefits of sanitary improvements.

The Poor Law Commission, created in 1834, explored

problems of community health and suggested means for solving them. Its report, in 1838, argued that "the expenditures necessary to the adoption and maintenance of measures of prevention would ultimately amount to less than the cost of the disease now constantly engendered." Sanitary surveys proved that a relationship exists between communicable disease and filth in the environment, and it was said that safeguarding public health is the province of the engineer rather than of the physician.

The Public Health Act of 1848 established a General Board of Health to furnish guidance and aid in sanitary matters to local authorities, whose earlier efforts had been impeded by lack of a central authority. The board had authority to establish local boards of health and to investigate sanitary conditions in particular districts. Since this time several public health acts have been passed to regulate sewage and refuse disposal, the housing of animals, the water supply, prevention and control of disease, registration and inspection of private nursing homes and hospitals, the notification of births, and the provision of maternity and child welfare services.

Advances in public health in England had a strong influence in the United States, where one of the basic problems, as in England, was the need to create effective administrative mechanisms for the supervision and regulation of community health. In America recurrent epidemics of yellow fever, cholera, smallpox, typhoid, and typhus made the need for effective public health administration a matter of urgency. The so-called Shattuck report, published in 1850 by the Massachusetts Sanitary Commission, reviewed the serious health problems and grossly unsatisfactory living conditions in Boston. Its recommendations included an outline for a sound public health organization based on a state health department and local boards of health in each town. In New York City (in 1866) such an organization was created for the first time in the United States.

Developments in Germany and France in the 19th century

Nineteenth-century developments in Germany and France pointed the way for future public health action. France was preeminent in the areas of political and social theory. As a result the public health movement in France was deeply influenced by a spirit of public reform. The French contributed significantly to the application of scientific methods for the identification, treatment, and control of communicable disease.

Although many public health trends in Germany resembled those of England and France, the absence of a centralized government until after the Franco-German War did cause significant differences. After the end of that war and the formation of the Second Reich, a centralized public health unit was formed. Another development was the emergence of hygiene as an experimental laboratory science. In 1865 the creation at Munich of the first chair in experimental hygiene signaled the entrance of science into the field of public health.

There were other advances. The use of statistical analysis in handling health problems emerged. The forerunner of the United States Public Health Service came into being, in 1878, with the establishment in the United States of port quarantine on a national basis and with assignment of enforcement of the quarantine to the Surgeon General of the Marine Hospital Service. (Port quarantine was the isolation of a ship at port for a limited period to allow time for the manifestation of disease.)

Developments from 1875. The work of an Italian bacteriologist, Agostino Bassi, with silkworm infections early in the 19th century prepared the way for the later demonstration that specific organisms cause a number of diseases. Some questions, however, were still unanswered. These included problems related to variations in transmissibility of organisms and in susceptibility of individuals to disease. Light was thrown on these questions by discoveries of human and animal carriers of infectious diseases.

19th-century developments in bacteriology

In the last decades of the 19th century the French chemist Louis Pasteur, the Germans Ferdinand Julius Cohn and Robert Koch, and others developed methods for isolating and characterizing bacteria; the English surgeon Joseph Lister developed concepts of antiseptic surgery; the English physician Ronald Ross identified the mosquito as the carrier of malaria; a French epidemiologist, Paul-Louis Simond, provided evidence that plague is primarily a disease of rats spread by rat fleas; and two Americans, Walter Reed and James Carroll, demonstrated that yellow fever is caused by a filterable virus carried by mosquitoes. Thus, modern public health and preventive medicine owe much to the early medical entomologists and bacteriologists. A further debt is owed bacteriology because of its offshoot, immunology.

In 1881 Pasteur established the principle of protective vaccines and thus stimulated an interest in the mechanisms of immunity. The development of microbiology and immunology had immense consequences for community health. In the 19th century the efforts of health departments to control contagious disease consisted in attempts to improve environmental conditions. As bacteriologists identified the microorganisms that cause specific diseases, progress was made toward the rational control of specific infectious diseases.

In the United States the diagnostic bacteriologic laboratory was developed—a practical application of the theory of bacteriology, which evolved largely in Europe. These laboratories, established in many cities to protect and improve the health of the community, were a practical outgrowth of the study of microorganisms, just as the establishment of health departments was an outgrowth of an earlier movement toward sanitary reform. And just as the health department was the administrative mechanism for dealing with community health problems, the public health laboratory was the tool for the implementation of the public health program. Evidence of the effectiveness of this new phase of public health may be seen in statistics of immunization against diphtheria—in New York City the mortality rate due to diphtheria fell from 785 per 100,000 in 1894 to 1.1 per 100,000 in 1940.

Poverty and disease

While improvements in environmental sanitation during the first decade of the 20th century were valuable in dealing with some problems, they were of only limited usefulness in solving the many health problems found among the poor. In the slums of England and the United States malnutrition, venereal disease, alcoholism, and other diseases were widespread. Nineteenth-century economic liberalism held that increased production of goods would eventually bring an end to scarcity, poverty, and suffering. By the turn of the century, it seemed clear that deliberate and positive intervention by reform-minded groups, including the state, also would be necessary. For this reason many physicians, clergymen, social workers, public-spirited citizens, and government officials promoted social action. Organized efforts were undertaken to prevent tuberculosis, lessen occupational hazards, and improve children's health.

The first half of the 20th century saw further advances in community health care, particularly in the welfare of mothers and children and the health of schoolchildren, the emergence of the public health nurse, and the development of voluntary health agencies, health education programs, and occupational health programs.

In the second half of the 19th century two significant attempts were made to provide medical care for large populations. One was by Russia, and took the form of a system of medical services in rural districts; after the Communist Revolution, this was expanded to include complete government-supported medical and public health services for everyone. Similar programs have since been adopted by a number of European and Asian countries. The other attempt was prepayment for medical care, a form of social insurance first adopted toward the close of the 19th century in Germany, where prepayment for medical care had long been familiar. A number of other European countries adopted similar insurance programs.

In the United Kingdom, a royal-commission examination of the Poor Law in 1909 led to a proposal for a unified state medical service. This service was the forerunner of the 1946 National Health Service Act, which represented an attempt by a modern industrialized country to provide services to all people.

In recent years prenatal care has made a substantial contribution to preventive medicine, for it is hoped that through the education of mothers the physical and psy-

chological health of families may be influenced and passed on to succeeding generations. Prenatal care provides the opportunity to educate the mother in personal hygiene, diet, exercise, the damaging effects of smoking, the careful use of alcohol, and the dangers of drug abuse.

Public health interests also have turned to such disorders as cancer, cardiac disease, thrombosis, lung disease, and arthritis, among others. There is increasing evidence that several of these disorders are caused by factors in the environment; for example, the association of cigarette smoking with certain lung and cardiovascular diseases. Theoretically, they are preventable if the environment can be altered. Health education is of great importance and is a responsibility of national and local government agencies as well as voluntary bodies. Life expectancy has increased in almost every country, except where public health standards are low.

MODERN ORGANIZATIONAL AND ADMINISTRATIVE PATTERNS

International organizations. Since ancient times, the spread of epidemic disease demonstrated the need for international cooperation for health protection. Early efforts toward international control of disease appeared in national quarantines in Europe and the Middle East. The first formal international health conference, held in Paris in 1851, was followed by a series of similar conferences aimed at drafting international quarantine regulations. A permanent international health organization was established in Paris in 1907 to receive notification of serious communicable diseases from participating nations, to transmit this information to the member nations, and to study and develop sanitary conventions and quarantine regulations on shipping and train travel. This organization was ultimately absorbed by the World Health Organization (WHO) in 1948.

In the Americas, the organization of international health probably began with a regional health conference in Rio de Janeiro in 1887. From 1889 onward there were several conferences of American countries, which led ultimately to the establishment of the Pan-American Sanitary Bureau; this was made a regional office of WHO in 1949, when it became known as the Pan-American Health Organization.

The rise and decline of health organizations has been influenced by wars and their aftermaths. After World War I, a Health Section of the League of Nations was established and functioned until World War II. After the war, the United Nations Relief and Rehabilitation Administration (UNRRA) was set up; it processed displaced persons in such a way as to prevent the spread of disease.

Establishment of WHO

It was responsible for the planning steps that led to the establishment in 1948 of the World Health Organization as a special agency of the United Nations. WHO is concerned with physical, mental, and social well-being and not merely with the absence of disease.

The work of WHO is carried out under the direction of the World Health Assembly, which has representatives from the member states. The first assembly gave consideration to diseases and problems that exist in large areas of the world and that lend themselves to international action. Malaria, tuberculosis, venereal disease, the promotion of health, environmental conditions responsible for a significant proportion of deaths, and nutrition were given priority. Other areas of need have been included since.

Among important functions of the organization are the advisory services offered to governments through its regional staff. Regional offices in a number of countries, both industrialized and developing, as well as local representatives in many developing countries, help WHO maintain contact with needs and sources of financial aid. In specialized fields, a number of expert committees consider specific questions.

WHO maintains close relationships with other United Nations agencies, particularly the United Nations Children's Fund (UNICEF) and the Food and Agriculture Organization (FAO), and with international labour organizations. From its inception in 1946, UNICEF focused its aid on maternal and child health services and the control of infections, especially in children. Priority has been given to the production of vaccines, the institution of environmental sanitation, the provision of clean water, and the training of local personnel in their own countries (especially in rural areas). Aid is channeled through organized health services in developing countries. Recent efforts have concentrated on persuading governments to undertake national surveys to identify the basic needs of their children and to devise appropriate national policies.

The work of WHO includes three main categories of activities. First, it is a clearinghouse for information about disease throughout the world, and it has developed a uniform system for reporting diseases and causes of death. It has established internationally accepted standards for drugs and drawn up a list of "essential" (effective, cheap, and reliable) drugs. It has sponsored and financed many research projects all over the world. Second, WHO has promoted mass campaigns to control epidemic and endemic diseases, many of which have been very successful. Third, WHO attempts to strengthen and expand the public health administration and services of member nations by providing technical advice, teams of experts to carry out surveys and demonstrate projects, and aid in support of regional and national health development projects.

Developed nations. Methods of health administration vary from country to country. Major health functions are frequently grouped in a department that is responsible for health and for related functions. In the United Kingdom they are carried out by the Department of Health and Social Security; in the United States the Department of Health and Human Services controls the programs covered by national legislation.

Departments of health

Few central departments of health are all-embracing; other departments also operate medical programs of some sort. No country places the health services of its military forces under the central health agency. Because unity of control at the centre is impracticable, coordination is important. Central administration is further complicated in federal systems. In the United States there are 50 states, no two of which have the same patterns of health organization.

Patterns shared. The official responsible for the administration of national health affairs is usually a member of the Cabinet. Advisory councils are frequently used to bring the ideas of leading scientists, health experts, and community leaders to bear on national health problems.

An organization that provides basic community health services under the direction of a medical officer is called a local health unit. It is usually governed by a local authority. Its programs may include maternal and child health, communicable-disease control, environmental sanitation, maintenance of records for statistical purposes, health education of the public, public health nursing, medical care, and, often, school health services. The local health unit can provide the administrative framework for a wider range of community health services, including the care of the aged, of the physically handicapped, and of the chronically ill, and mental health services. Although social welfare services may be provided by a separate agency, there are advantages in amalgamating health and welfare services, because a family's health and social problems tend to be interrelated. In England, welfare and public health are often integrated at the local level, whereas in the United States they are almost always separate.

The population served by a local health unit may be only a few thousand or several hundred thousand. There are substantially different problems involved in administering health services for a large rural area that is sparsely populated and a municipality with a population of one or two million.

One problem of administering local health services is the question of whether they should be run by independent local authorities or organized regionally to ensure coordination and effective referral and to avoid duplication of services.

Types of public medical service

Medical care is provided as a public service to some degree in most countries. It may be limited to hospitalization of persons with certain ailments, such as mental disease, tuberculosis, chronic illness, and acute infections. Comprehensive health services may be provided for certain

population groups, as in Canada and the United States, where the federal government provides care for Indians and Eskimos. Many countries have compulsory medical insurance, and some combine the socialization of hospitals with medical insurance covering general medical care, as in Denmark. Full-scale socialization of health services exists in a few countries, including the United Kingdom, the Soviet Union, and New Zealand.

In countries such as The Netherlands and the United States, where voluntary and nonprofit organizations support a considerable share of the health services and operate most of the general hospitals, there is pluralism in health administration. This makes coordination difficult, but voluntary effort has the advantages of involving citizens directly in the development of health services and of promoting experimentation in administration.

Levels of health care

There is a trend toward regional planning of comprehensive health services for defined populations. In an idealized plan, the first level of contact between the population and the system, which can be called primary care, is provided by health personnel who work in community health centres and who reach beyond the health centres into the communities and homes with preventive, promotive, and educational services. At the next level of care, specialists in community hospitals provide secondary care for patients referred from the primary-care centres. Finally, tertiary, or superspecialty, care is provided by a major medical centre. The various levels of this regional scheme are linked by a two-way flow of medical records, patients, and health personnel. Regionalization has been most fully achieved in Europe and least so in North America, where voluntary hospitals provide most of the short-term general services and retain autonomy in their administration.

Variations. Among the developed nations, there is substantial variation in the organization and administration of health services. The United Kingdom has a National Health Service with substantial autonomy given to local government for implementation; the Soviet Union has health services that are the responsibility of the national government, and administration is strongly centralized; and the United States has a pluralistic approach to health services, in which local, state, and national governments have varying areas of responsibility, with the private sector playing a prominent role.

During the first half of the 20th century in Britain, the emphasis shifted gradually from environmental toward personal public health. A succession of statutes, of which the Maternity and Child Welfare Act (1918) was probably the most important, placed responsibility for most of the work on county governments. National health insurance (1911) gave benefits to 16,000,000 workers and marked the beginning of a process upon which the National Health Service Act (1946) was built.

The National Health Service Act provided comprehensive coverage for most of the health services, including hospitals, general practice, and public health. The service remained at the periphery, however, in three types of care: (1) Primary medical care is given by family physicians or general practitioners. This service is organized locally by an executive council. Each general practitioner has a list of people registered with him for whom he is responsible for providing primary care. (2) Specialist consultation and outpatient and inpatient treatment are provided in hospitals under the direction of regional authorities. A later concept makes each district general hospital responsible for providing hospital services for a defined population. (3) Services, such as health visiting, home nursing, home helps, domiciliary midwifery, the prevention of illness, and the provision of health centres are the responsibility of local authorities.

Health services in the Soviet Union

In the Soviet Union the protection and promotion of public health is the responsibility of the state. There is free public access to all forms of medical care. The principles of the health services are complete integration of curative and preventive services; medicine as a social service; preventive programs; health centres or polyclinics (clinics in which a variety of diseases are handled); and community participation.

The public health services for the Soviet Union are directed by the Ministry of Health. Each of 15 republics of the union has its own ministry. Each republic is divided into *oblasti* (provinces), which in turn are divided into *rayony* (municipalities), and finally into *uchastoki* (districts). Each subdivision has its own health department accountable to the next highest division (see above *The practice of public health*).

There are well-established referral procedures, from the polyclinics and smaller hospitals in the *uchastoki* to the larger *rayon* hospitals, and from feldshers (paramedical personnel trained in medical care) and other paramedical personnel to internists and pediatricians and, when necessary, to more highly specialized personnel.

The health services of the United States can be considered at three levels: local, state, and federal.

Levels of service

Locally in cities or counties, there is substantial autonomy within broad guidelines developed by the state. The size and scope of local programs vary, but some of their functions are control of communicable diseases; clinics for mothers and children, particularly for certain preventive and diagnostic services; public health nursing services; environmental health services; health education; vital statistics; community health centres, hospitals, and other medical care facilities; community health planning and coordination.

At the state level, a department of health is charged with overall responsibility for health, though a number of agencies may actually be involved. The state department of health usually has five functions: public health and preventive programs; medical and custodial care such as the operation of hospitals for mental illness; expansion and improvement of hospitals, medical facilities, and health centres; licensure for health purposes of individuals, agencies, and enterprises serving the public; and financial and technical assistance to local governments for conducting health programs.

At the federal, or national, level, the Public Health Service of the Department of Health and Human Services is the principal health agency, but several other departments have health interests and responsibilities. Federal health agencies accept responsibility for improving state and local services, for controlling interstate health hazards, and for working with other countries on international health matters. The federal government also has the following specific responsibilities: (1) protecting the United States from communicable diseases from abroad; (2) providing for the medical needs of military personnel, veterans, merchant seamen, and American Indians; (3) protecting consumers against impure or misbranded foods, drugs, and cosmetics; and (4) regulating production of biological products, such as vaccines. In addition, the federal government promotes and supports medical research, health services, and educational programs throughout the country.

Voluntary effort is a significant part of health work in the United States. There are more than 100,000 voluntary agencies in the health field functioning mostly at the local level but also at state and national levels. Supported largely through private sources, these agencies contribute to programs related to education, research, and health services.

Channels for medical care

Medical care is provided and paid for through many channels, including public institutions, such as municipal, county, state, and federal health centres, hospitals, and medical care programs, and through private hospitals and private practitioners working either alone or, increasingly, in groups. Generally, medical care is financed by public funds, voluntary health insurance, or personal payment. There is a trend away from the traditional fee-for-service payment to individual practitioners toward prepaid-care systems including health teams working at community, health centre, and hospital levels.

Thus, in the United States there is great variety in the content, scope, and quality of health services. These services are provided by several independent agencies. In effect, however, they constitute a working partnership for the protection and promotion of human health.

Recently, two factors have contributed to rapid change in the orientation of health services in the United States. One of these is an increasing awareness that, while the existing system of health services provides high quality

care for many, there are others for whom the care is either lacking or unsatisfactory. The second factor is that of steeply rising costs of medical care. These two issues have led to reconsideration of the entire system of personal medical care and proposals for new systems of providing and financing health care.

Developing nations. *Patterns shared.* Developing countries have sometimes been influenced in their approaches to health care problems by the developed countries that have had a role in their history. The countries in Africa and Asia that were once colonies of Britain have educational programs and health-care systems that reflect British patterns, though there have been adaptations to local needs. Similar effects may be observed in countries influenced by France, The Netherlands, and Belgium.

Despite variations from country to country, a common, if somewhat idealized, administrative pattern may be drawn for developing countries. All health services, except for a small amount of private practice, are under a ministry of health, in which there are about five bureaus, or departments—hospital services, health services, education and training, personnel, and research and planning. Hospital and health services are distributed throughout the country. At the periphery of the system are dispensaries, or health outposts, often manned by one or two persons with limited training. The dispensaries are often of limited effectiveness and are upgraded to full health centres when possible. Health centres and their activities are the foundation of the system. Health centres are usually staffed by auxiliaries who have four to 10 years of basic education plus one to four years of technical training. The staff may include a midwife, an auxiliary nurse, a sanitarian, and a medical assistant. The assistants, trained in the diagnosis and treatment of sickness, refer to a physician the problems that are beyond their own competence. Together, these auxiliaries provide comprehensive care for a population of 10,000 to 25,000. Several health centres together with a district hospital serve a district of about 100,000 to 200,000 people. All health services are under the responsibility of the district medical officer, who, assisted by other professional and auxiliary personnel, integrates the health efforts into a comprehensive program.

The typical administrative pattern

Of central importance is the distribution of responsibilities between auxiliaries and professionals. The auxiliaries, by handling the large number of relatively simple problems, allow the professionals to look after only the more complex problems, to supervise and teach the auxiliaries, and to plan and manage the programs.

The district hospital is dependent on a regional hospital, to which patients with complex problems can be referred for more specialized services. Administrative direction of both regional health services and regional hospital services can be combined at this level under a regional medical officer. The central administration of the ministry of health provides policies and guidance for an entire health service and, in some instances, also provides a central planning unit.

Problems of transportation and communication over great distances, shortages of staff and other resources, and inadequacies in staff preparation and motivation often lead to malfunctions in the system. Nonetheless, the public health services developed in African and Asian countries have generally provided a sound basis for future development within the framework of national development.

Variations. The organization of public health services in Latin-American countries differs substantially from those of Africa and Asia; these differences are an expression of their different historical backgrounds. The Latin-American countries are generally more affluent than those of Asia and Africa. Private practice is more widespread, and private or voluntary agencies are more prominent. Health services are provided largely by local and national governments. Many Latin-American countries also have systems of clinics and hospitals for workers financed by employers and workers. The distribution of health services, with health centres, hospitals, and preventive services, is roughly similar to Africa and Asia. The Latin-American countries, however, have used auxiliaries less than African and Asian countries. Latin America has pioneered in the development of health-planning methods. Chile has one of the most advanced approaches to health planning in the world.

Differences between Latin America and Africa

Thailand was never colonized and therefore has no historical influence favouring any particular pattern of health services. The Thai Ministry of Health has a well-developed system of hospitals and health centres distributed across the country to serve rural as well as urban people. It differs from the pattern described in the previous section in that, despite the extreme shortages of physicians and nurses in rural areas, the nation has been reluctant to use auxiliaries for medical care. It does, however, use auxiliary midwives and sanitarians. Hospital services and public health services have separate administration. Within the public health services, there are a number of separate divisions—*e.g.*, for tuberculosis, venereal disease, and nutrition—each with its own staff, budget, and facilities. The trend elsewhere has been away from relatively independent, disease-oriented approaches and toward integrated systems in which the same network of health services handles most problems.

Health problems and obstacles. The difficulties of providing health services for the people of the developing nations involve a cluster of interrelated problems. These arise from the nature of the diseases and hazards to health, insufficient and maldistributed resources, the design of health services systems, and the education of health personnel in those systems. Woven through the health programs of the developing nations and complicating them at both family and national levels are the pressures associated with rapidly growing populations.

There are differences not only in the kinds of diseases of different countries but also in the rates at which they occur and in the age groups involved. Life expectancy in some countries is less than half that in others, principally because of high death rates among small children in the developing countries. In much of Southeast Asia, for example, 40 percent of children die by their fourth year, a death rate not reached until age 60 in North America. The infant (under one year of age) mortality rate in Central and South America is two to four times that in North America, and the death rate in children one to four years of age is as much as 25 times greater. The differences for Central Africa are even more striking: infant mortality in some areas has been 12 times that in the United States, and the mortality in preschool children has been more than 60 times the United States figure.

Infant mortality rates

The principal causes of sickness and death among small children in the developing world are diarrhea, respiratory infections, and malnutrition, all of which are intimately related to culture, custom, and economic status. Malnutrition is as often due to food customs as it is to the lack of food—taboos and simple oversight lead to deprivation of children too small to help themselves. Gastroenteritis (inflammation of the lining of the stomach and intestines, usually with accompanying diarrhea) and respiratory infections are often due to infectious organisms that are not susceptible to antibiotics. The interrelationships of these diseases increase the complexity of dealing with them. Malnutrition is often the underlying culprit; not only does it cause damage itself, such as retardation of physical and mental development, but it also seems to set the stage for other illnesses. A malnourished child develops gastroenteritis, inability to eat, further weakness, and then dehydration. Weakened in this way, the child is susceptible to a lethal infection, such as pneumonia. Or, to complete the vicious circle, infection can affect protein metabolism in ways that contribute to malnutrition.

Another factor that contributes to this is family size. Malnutrition, with associated death and disability, occurs most often in children born into large and poorly spaced families. The resulting high death rate among small children often reinforces the tendency of parents to have more children. People are not inclined to limit the size of their families until it is apparent that their children have a reasonable chance of survival. Thus, there is a fertility–mortality cycle in which high fertility, reflected in large numbers of small children crowded into a poor home, leads to high childhood mortality, which, in turn,

encourages high fertility. This is the basis of the belief that population-control programs should include effective means of reducing unnecessary deaths among children.

Ratios of medical personnel to population

Among limitations of resources, shortages of trained personnel are among the most important; ratios of population to physicians, nurses, and beds provide an indication of the seriousness of these deficiencies and also of the great differences from country to country. Thus, the proportion of population to physicians in developing countries varies drastically.

Money is a crucial factor in health care—it determines how many health personnel can be trained, how many can be maintained in the field, and the resources that they will have to work with when they are there. Governmental expenditures on health care vary greatly from country to country.

As it attempts to provide health care for its people, a nation, on the one hand, must meet the urgent and complex problems, such as obstetric and surgical emergencies for which hospital care is essential. On the other hand, it must reach into the communities and homes to find those who need care but do not seek it and must discover the causes of such diseases as malnutrition and gastroenteritis.

Education of health personnel. In the education of health personnel, a particular set of problems emerges. Educational programs for auxiliaries are suited to the local situation, perhaps because they were not established in the more developed nations. Medical and nursing education, on the other hand, is similar to that of the more advanced countries, and it prepares students better for working in industrialized nations than in their own. This misfit between education and the jobs to be done has probably contributed substantially both to the ineffectiveness of health services systems and to the migration of professional personnel to the more developed countries.

PROGRESS IN PUBLIC HEALTH

Developed nations. Among the more developed nations the following trends are apparent.

Increasing interest of national governments. Formerly, governments were chiefly concerned with basic health problems, such as environmental sanitation, medical care of the poor, quarantine, and the control of communicable diseases. Gradually, they have extended their activities into the field of medical care services in the home, clinic, and hospital, so as to provide comprehensive health care for entire communities. Three factors have influenced this trend: (1) the nongovernmental voluntary agencies have been unable to meet the rising cost of medical care; (2) there is an increasing appreciation of the economic loss to a country from sickness; and (3) there is an increasing public interest in social services. Health and social welfare are now recognized as complementary, and social legislation tends to cover both areas. There is an administrative trend toward a close cooperation between health and social welfare services.

Factors in the increase of government involvement

Changing concepts of preventable disease. Until recently, the term preventable disease referred to a circumscribed group of infectious diseases. The term has acquired a broader meaning, however, as epidemiological methods are applied to other conditions. Preventive health services now deal with a wide range of health hazards, such as malignant tumours, rheumatism, cardiovascular diseases, other chronic and degenerative diseases, and even accidents.

Integration of preventive and medical care services. Medical care had its origin in the humanitarian motive of caring for the sick, while preventive services sprang from the need to protect a healthy environment from epidemic diseases. They grew apart, but recently the trend has been to integrate them within a comprehensive health service. Such an integration is the fundamental principle of public health in the U.S.S.R. and in some other eastern European countries in which all local health services are centred in the district hospital under one administration. In other European countries, especially in rural areas, the two branches are brought together by the local medical practitioner. The focal point of many discussions is the role that the hospital should play in health services. Many feel that its influence at present is too restricted and that it should spread beyond its walls to health centres and homes.

Comprehensive health service

Provisions directed toward better mental health. Mental health now has a place in the preventive services. Improvements in arrangements for mental health include the provision of outpatient clinics and inpatient accommodations at general hospitals for early mental cases, an increase in child-guidance and marriage-guidance clinics, and schemes for the care of alcoholics and drug addicts. There have also been significant developments in the treatment of maladjusted members of society. Gains in understanding of psychoneuroses by general practitioners and the development of research facilities are also noteworthy.

Growing emphasis on health education. Many countries have expanded their commitment to health education, usually in cooperation with voluntary agencies. The most effective work is carried out at the local level, especially in schools. The trend is toward an expansion of health education as an essential preventive health service.

The biostatistical, epidemiological approach. A statistical service is essential in planning, administering, and evaluating health services. The interest of public authorities in medical-care schemes has increased the importance of statistics on the incidence of diseases and other problems, as well as the epidemiology necessary to combat them. Both are vital in the planning, organization, and evaluation of medical-care schemes. Traditionally, the epidemiological method was used for infectious diseases, but it is now being used increasingly for noninfectious diseases and the problems of medical care.

Changes resulting from an aging population. In more affluent nations, an increase in older age groups brings about the need for public health facilities to provide special services for them. Health care of the elderly includes measures to prevent premature aging and the chronic and degenerative diseases and to confront the psychological problems resulting from loneliness and inactivity. Geriatric clinics have been set up to meet these needs and to conduct research into the process of senescence.

Concern regarding the quality of the environment. There is widespread concern about environmental deterioration. Controlled atomic radiation has created new hazards to health, such as the potential pollution of air or water by radioactive discharges, the possible effects from radioactive fallout on the public generally, and the dangers to workers in atomic installations in industry. A growing population requires an increase in industrial and commercial activities, which add to the volume of pollutants that threaten the atmosphere, rivers, lakes, and oceans and have destructive effects on natural ecology. Individual countries have taken steps toward the control of environmental deterioration, and means of international regulation have also been proposed.

New hazards to health

Developing nations. In view of the large numbers of serious health problems in the developing nations and their limited resources for dealing with them, it is understandable that along with substantial progress there would be some stagnation, or even regression.

Communicable-disease control. Smallpox and malaria are examples of diseases that have been brought under closer control throughout the world. For other diseases, such as hepatitis (liver inflammation), rabies, leprosy, and sleeping sickness, there have been important growths in understanding that may contribute to their eventual control.

Disease problems that await solution. El Tor cholera, which has appeared in epidemic form in previously uninvolved areas, represents one of the most serious challenges to public health. Venereal disease, an old problem, has increased in incidence. Certain parasitic diseases have spread as humans have brought about changes in their environment—the increase in schistosomiasis (infestation with blood fluke by means of snails as the intermediate hosts) in irrigation and man-made lake areas is an example. Widespread malnutrition, particularly protein–calorie malnutrition in small children, remains a problem. Protein-rich food supplements and more effective educational programs are being developed to combat it.

Family health. The problems of rapidly growing populations have important consequences at both the family

Population growth

and the national level. Problems of maternal and child health, human reproduction, and human genetics, including family planning, are now seen as aspects of the greater problem of the health of the whole family as a single and fundamental social unit. Accordingly, family health is a matter deserving high priority among the public health services.

Health manpower. There is widespread recognition of inadequacies in both number and education of health personnel. The trend is toward coordinating the education of health personnel with the particular health service in which they will function. This trend requires close relationships between educational institutions and the agencies responsible for health services.

Comprehensive community health services. The fragmentation of earlier health service organizations, such as programs concerned with only a single disease and the separation of curative and preventive services, is giving way to more comprehensive organizational patterns. Health promotion, disease prevention, and the curing and rehabilitation of the ill are brought together into one network of integrated services that reaches to the community level.

National health planning. Complex decision making is involved in allocating limited health service resources to large numbers of people. As a result, there has been an increasing emphasis on the health-planning process and on the design of more effective health-service systems. A number of countries have established health-planning units in the ministry of health or the national planning organization. An important aspect of national health planning is the close coordination between planning, budgeting, implementing, and evaluating programs.

(J.H.Br./P.Rh.)

Clinics

A clinic is an organized medical service offering diagnostic, therapeutic, or preventive treatment to ambulatory patients. Often in Europe and occasionally in the United States the term covers the entire teaching centre, including the hospital and the ambulatory-patient facilities. The medical care offered by a clinic may or may not be connected with a hospital.

The term clinic may be used to designate all the activities of a general clinic or it may be used to designate only a particular division of the work; *e.g.,* the psychiatric clinic, neurology clinic, surgery clinic, etc. The entire activity when connected with a hospital is called the outpatient department, and the specific subdivisions are referred to as clinics.

The London Dispensary

The first clinic in the English-speaking world, the London Dispensary, was founded in 1696 as a central means of dispensing medicines to the sick poor whom the physicians were treating in the patients' homes. The New York City, Philadelphia, and Boston dispensaries, founded in 1771, 1786, and 1796, respectively, had the same objective. Later, for the sake of convenience, physicians began to treat their free patients at the dispensary. The number of such clinics did not increase rapidly, and as late as 1890 only 132 were operating in the United States. The impetus for the mushroomlike growth that has occurred since that time came with the rapid growth of hospitals and also from the public health movement.

HOSPITAL CLINICS

During the late 1800s the modern concept of a hospital began to take shape. During this period some of the hospitals connected with medical schools inaugurated outpatient departments for the purposes of teaching and charity. The advantages of providing ambulatory care close to the facilities of a hospital became apparent, and such hospital clinics multiplied rapidly.

The organization of a hospital clinic in general follows that of the inpatient facilities. Hospital clinics are primarily concerned with acute diseases, and the physicians in the clinics are usually the same physicians who treat inpatients in the hospital.

In many hospital clinics, especially those in countries that do not have national health insurance programs, care is made available only to the medically indigent, and no professional fee is charged. Practically all such clinics, however, charge a small registration fee if the patient is financially able to pay; income from such fees helps pay operating costs. A number of successful attempts have been made to extend hospital clinic care to paying patients. Most of this effort has been in the area of lower income groups although in a few hospitals no limit is placed on income in determining eligibility for care. The hospitals of the University of Chicago, for example, began operating a clinic on such a basis in 1928.

PUBLIC HEALTH CLINICS

The public health movement has been mainly concerned with preventive medicine, child and maternal health, and other medical problems affecting broad segments of the population. The first public health clinics were established in the late 19th century. In 1890 A. Pinard set up a maternal dispensary or antenatal clinic at the Maternité Baudelocque in Paris. Milk distribution centres were set up in France by J. Comby (1890) and in Britain by F.D. Harris (1899). Infant welfare clinics were established in Barcelona (1890); and clinics for older children were founded in St. Pancras, London, by J.F.J. Sykes (1907).

The first clinics

Unlike the hospital clinics, which have had their greatest growth in the cities, the public health clinics are located chiefly in smaller towns and villages. In the United States the first great movement in creating public health clinics resulted in the founding of the National Association for the Study and Prevention of Tuberculosis in 1904. It was the association's goal to study and prevent tuberculosis by making clinic facilities available for free examination and treatment in every city and country. Other nationwide, private health agencies in specialized medical fields quickly adopted this method to improve the quality and extend the quantity of medical service in their fields. Local governmental health units operate similar clinics in the medical specialties related to the prevention of communicable disease and long-term illness. Such clinics are generally concerned with one particular medical interest, usually one of the following: tuberculosis, venereal disease, prenatal care, well-baby care, teeth, tonsils, eyes, crippled children, and mental health. There is a tendency toward the establishment of traveling clinics, such as dental clinics for schoolchildren. Often no charge is made for service in public health clinics, and for many medical conditions no income restrictions are imposed. A few are operated in connection with hospitals, but most such clinics use public buildings or space furnished by welfare and other social agencies. Financial support is received mostly from the same sources.

Traveling clinics

PRIVATE CLINICS

The advantages of group medical service, with facilities and technical personnel beyond the means of an individual practitioner plus the benefit of group consultation, have encouraged the establishment of pay or private clinics. Such a clinic is essentially a voluntary association of physicians engaged in the practice of medicine on an organized group basis. Common administration and facilities are used, and the resulting expense and income are shared according to a predetermined plan. To be classified as a group clinic the relationship between each physician and the organization must be defined in a legal agreement. The relationship usually takes the form of a partnership. Several of these, such as the Mayo Clinic, in Rochester, Minn., have achieved a national reputation and attract patients from a wide area. Most of these organized group clinics are general clinics—*i.e.,* they have several of the different medical specialties represented on their staffs. A number of private clinics, however, limit their work to one medical specialty. An enterprise of special interest is the London Clinic, established in 1936 by a group of prominent consultant surgeons and physicians who wished to make available to their private patients a place where the comforts and privacy of a nursing home could be combined with facilities for diagnosis and therapy such as exist only in the larger general hospitals.

Usually the group is organized independently of any hos-

pital or other agency, but in some instances such clinics own and operate their own hospital facilities. In other instances the clinic is a part of a prepaid health service plan. This latter pattern has received impetus in recent years as labour unions have set up medical clinics supported by welfare benefits contributed by employers. The United Mine Workers, for example, has established a system of such clinics in hospitals constructed by the union in the coal-mining areas of Virginia and West Virginia.

HEALTH CENTRES

In 1910 the first health centres, or multiple clinics, were established in Pittsburgh and Wilkes-Barre, Pa. Others were opened in 1913 in New York City and in 1916 in Boston and Philadelphia. In 1920 in Britain a consultative council on medical and allied services (Dawson Committee) described a health centre as an institution wherein various medical services, preventive and curative, are brought together. Under Section 21 of the National Health Service Act, 1946, local health authorities provide, equip, maintain, and staff health centres to offer facilities for all or any of the following services: general medical and dental; pharmaceutical; hospital outpatient and health education; and care of expectant and nursing mothers. The prenatal and postnatal clinics include periodic medical and dental examinations, classes in parenthood and relaxation, and welfare foods. Activities in child welfare clinics comprise education in all aspects of motherhood; periodic medical and dental examinations; advice on mental health problems; immunization and vaccination; and distribution of welfare foods.

Services

POLYCLINICS

In the U.S.S.R. the polyclinic (*poliklinika*) is designed to combine the function of a hospital outpatient department with that of a general practitioner health centre. Factory workers and their families can attend general polyclinics attached to major factories, and there are many special children's polyclinics in the towns. A typical urban polyclinic, which is usually associated with a hospital, includes reception and waiting rooms, registration desk, and consulting and treatment rooms of the following: internist, pediatrician, gynecologist, surgeon, eye specialist, ear, nose, and throat specialist, neurologist, urologist, and dentist. There are often rooms for first aid, reception of infectious cases, and women and children's welfare, as well as a dental department, drugstore, pathological laboratory, X-ray department, gymnasium, and lecture hall. There are always fairly elaborate physiotherapy departments and usually large and small operating theatres. (Ed.)

FAMILY PLANNING CLINICS

The main purposes of the family planning service are to encourage parents to make responsible decisions about pregnancy that take into account the best interests of the family; to provide guidance to couples who wish to limit the size of their families; and to advise on the technical methods that are available for doing so. There are marked differences in attitude toward the desirability of a reduction in family size as between developed and developing countries. This difference is dominated by high infant and child mortality in most developing countries as compared with developed countries. (Ha.Sc.)

RELATED FIELDS

Nursing

The need for nursing is universal. The International Council of Nurses states that the function of nursing is fourfold—to promote health, prevent illness, restore health, and alleviate suffering—and that inherent in nursing is "the respect for life, dignity and the rights of man. It is unrestricted by considerations of nationality, race, creed, colour, age, sex, politics or social status. Nurses render health services to individuals, the family and the community and co-ordinate their services with those of related groups." Nurses form the largest group of health service workers throughout the world. As medical science advances, the health needs of a given population change and expectations for care rise, so that the potential for nursing service increases. The World Health Organization (WHO) stated in 1986 that "nursing and midwifery personnel have an even greater potential contribution to make to health for all than they are already making. Nurses' potential lies in their role as providers of primary health care services in community settings, clinics, schools and industries as well as in hospitals."

HISTORY OF NURSING

Contribution of Christianity

When Christianity developed, with its teachings of duty, love, and brotherhood, caring for the sick gained new impetus. Among the early converts to Christianity were highborn Roman matrons. Phoebe, a friend of St. Paul, "succored many"; Marcella made her palace into a monastery for women; Fabiola turned her home into a hospital. These women and others like them exemplified new intellectual freedom for women and their participation in religious and social action at a high level.

For several centuries thereafter the spread of health work moved slowly. Monastic orders grew, and from them some health services radiated; but large areas of the world were untouched, and the monastic undertaking of health work seemed to diminish. Later, military and chivalric orders were founded, supporting the Crusades and combining the making of war with charitable and hospital work. Some established companion orders for women. Monastic and chivalric orders established in various forms among health workers the beginnings of hierarchical organization, amateur participation that provided the roots of volunteerism, and a sense of calling or vocation.

Nursing can be seen to rise from two wellsprings, one scientific, the other religious and social. Acceleration of scientific advancement in health began with the 16th century. During the 19th century the "germ theory" of disease was developed; ways of treating and preventing infectious diseases, then the largest cause of death, were introduced. Anesthesia was also discovered. It is often said that research produced more medical and health knowledge in the decades after World War II than in all previous centuries combined. The growing mass of new knowledge to be applied in health services by health workers challenged the educational system for physicians, nurses, and others and strained the system of distribution of services to an awakened public.

Paralleling these scientific advances were those made through the years in social and religious action. The work of St. Vincent de Paul and, later, of John Howard typify this kind of action. After making thorough studies of conditions, they recommended marked changes in hospitals and prisons. A new dimension, planning based on facts, thus was added. St. Vincent, with St. Louise de Marillac, founded the Daughters of Charity (1633). In the following century Howard revealed deplorable conditions in nursing—filth, stealing, and ill treatment of patients. Theodor Fliedner, a 19th-century clergyman, later influenced reforms and emphasized citizens' responsibility for the health and welfare of people everywhere.

The work of Florence Nightingale

During the 19th century the movement for reform in nursing was led by Florence Nightingale, a woman of intellectual and moral power. Family contacts with humanitarian leaders and an education that included training in science, mathematics, and political economy were her preparation. She critically studied nursing as it was practiced in several countries, formulated her ideas, and wrote extensively.

In 1854 Florence Nightingale was asked by the British secretary of state at war to go to Scutari in Turkey, where absence of sewers and of laundering facilities, lack of supplies, poor food, disorganized medical service, and absence of nursing led to a death rate of more than 50 percent among wounded soldiers. Her work and that of

the nurses whom she recruited brought sufficient improvement to lower the death rate to 2.2 percent. A gift of £45,000 was raised by popular subscription, and Florence Nightingale used it to establish schools of nursing at St. Thomas's Hospital in London and elsewhere.

Florence Nightingale believed nursing to be suitable as an independent career for capable, trained women, that nursing services should be administered by those with special preparation, and that relationships between physicians and nurses should be professional. She maintained that schools of nursing should be administered by nurses with physicians as part of the hospital labour force. She believed that there was a substantial body of knowledge and skills to be learned in nursing. Nurses were to be prepared for hospital nursing and care of the sick at home, and they were to teach good health practices to patients and families.

By the end of the 19th century, the idea that a nurse needed to be educated and trained had spread to much of the Western world. In England, Scandinavia, America, and much of the British Empire, schools of nursing were generally based on training hospitals, and more nurses had become independent of religious institutions. On the Continent, however, the Motherhouse system and religious organizations often persisted. With the foundation of the Red Cross in 1863, some national Red Cross societies undertook the training of nurses and provided instruction for auxiliaries to help in time of war or emergency.

With advances in medicine, nurses followed doctors into specialities, including pediatrics, surgery, orthopedics, ophthalmology, psychiatry, and public health. The development of physicians' assistants has sometimes overlapped with the development of nurse specialists, though efforts have been made to differentiate the two categories. Pediatric nurse-practitioners, for example, now undertake functions previously performed only by physicians—taking histories, performing physical examinations on children, and running clinics for mothers and babies. There has also been experimentation with shifts in functions from medicine to nursing.

Planning is aimed at bringing nursing resources into balance with the needs for nursing personnel to implement health programs. WHO has developed a guide for such planning, published in several languages; the guide helps countries to establish goals and avoid disadvantages of haphazard growth. Coordinated planning among the several health professions and occupations is now considered to be highly desirable.

THE PRACTICE OF NURSING

Methods of classifying nurses

Kinds of nursing. Nursing personnel and the kinds of nursing in which they engage may be classified in a variety of ways: by legal designation—registered nurses, licensed practical nurses, and, in addition, unlicensed nurses' aides; by the kind of educational preparation—vocational, technical, and professional, including graduate education—that they have received; by the kind of work they perform—institutional, community, educational, occupational, research, or journalistic; by the level of responsibility that they assume—that of staff nurse, teacher, supervisor, administrator, or consultant; and by the place of employment—hospital, physician's office, public health agency, school (school health nurses), industry, a school of nursing, including undergraduate and graduate programs. Nurses may also be classified as either generalists or specialists. Most descriptions of nursing, including the following, combine these classifications.

Hospital nursing. Hospital nursing occupies approximately two-thirds of the total nursing force in most countries. Nurses giving direct care to patients in hospitals of the United States, for example, include registered nurses, some of whom may be specialists; licensed practical nurses; and nurses' aides. Aides and sometimes practical nurses, if there are any, are called auxiliaries in many other developed countries. In large, highly organized hospitals, administrators of nursing services need special preparation and occupy demanding executive positions.

Certain areas of hospitals are highly specialized—surgical suites (operating rooms), recovery and intensive-care units, coronary-care units, and drug addiction clinics. In these, nurses giving direct patient care and their supervisors must have the additional preparation afforded by graduate study or by continuing education or staff-development (in-service-education) programs.

The rapid development of hospitals in response to advances in science and new community responsibilities requires new kinds of nursing. Some hospitals use nurse-midwives in maternity departments. Nurse specialists work in a number of other areas as well—medical, surgical, obstetric, pediatric, psychiatric, and rehabilitative; these nurses frequently cross over to other fields as consultants. For example, public health nurses are employed to ease patients' adjustment to returning home and to assure planned continuity of health care. A few nurses occupy ombudsman, or patient-advocate, positions. Some others perform research aimed at solving problems and at devising improved ways of caring for patients.

New health problems bring additional responsibilities to nurses in hospitals. Drug abuse, family planning, and terminal illness, for example, are areas that call for specialized knowledge on the part of nurses, as well as for decisions about the organization of the services in hospitals.

Out-patient and home care

Outpatient and home care are sometimes organized by hospitals to reduce costly hospitalization for patients. Clinics in outpatient departments may be operated almost exclusively by nurses—clinics for hypertension (high blood pressure), chronic heart disease, and poststroke and maternity patients are examples. Satellite clinics bring services close to patients' homes. These expanded services require that nurses learn assessment methods—simple physical examination and history taking, for example—understand community problems and the organization of nursing services.

Efforts to return many patients of mental hospitals to their communities have met with some success. In some of these hospitals, as well as in community mental health centres, nurses with graduate degrees share in the treatment of patients and conduct group therapy.

Public health nursing. Public health, or community health, nursing may be either governmental or private. Visiting-nurse services care for the sick at home and carry on individual, family, and community programs of prevention and health teaching. These services are usually supported by civic and other forms of private philanthropy and by fees from patients who can afford to pay. Some of the visiting-nurse services have contracts with local government health and welfare agencies and with industrial health units to provide services. Many have deep, traditional roots in their communities.

Support

In 1978 WHO declared its goal of "health for all" by AD 2000, stressing the efficient and effective use of medical personnel and urging the greater use of nurses in primary care. WHO called for the reorientation of the basic nursing curriculum from an institutional to a community basis, concentrating on education and prevention. Many member states have made progress in that direction. Nursing schools in such countries as Chile, Colombia, Costa Rica, Cuba, Ecuador, Honduras, Indonesia, Nicaragua, Panama, Peru, the Philippines, and Thailand have increased their emphasis on the theory and practice of community health. Other member states, including several in Africa, have begun training for an expansion of nursing functions to strengthen primary health care services, producing nurse practitioners licensed for diagnostic and therapeutic activities, previously considered the domain of the physician, and the prescription of drugs. These functions are usually performed in health centres and in health posts within an organized administrative framework. In some countries nurses are often the only health professionals in isolated rural areas, and even in urban areas specially trained nurses can save the physicians' time and allow for the provision of better service.

Nurses as educators. Nursing education is a field that combines nursing with the teaching of students of nursing and, for some, with the administration of educational programs. A high proportion of the teachers in nursing-education programs teach in clinical situations, in which students learn to care for patients and families in hospitals,

at home, and in other situations. Teaching by nurses in staff-development programs of hospitals and other health agencies and in continuing-education programs is usually offered by universities or associations.

Private, office, and industrial nursing. Private, office, and industrial (occupational health) nursing as fields of employment account for only a small percentage of registered nurses. Licensed practical nurses are also employed in these fields.

Military nursing. Military nursing provides an essential part of the health care given to men and women of the armed services in most countries. During World War I many of the trained nurses who were part of the military forces of the countries involved were drawn from a reserve maintained by the Red Cross. By the close of World War II much progress had been made in assigning nurses rank and responsibilities commensurate with their training and abilities. Medical corpsmen today save lives on battlefields and hospitals and are highly skilled. Some remain in health work after they leave military service.

Government nursing. This field includes military nursing and public health nursing as carried on by national and local health departments, functions discussed above.

Education. The basic educational program for nurses in many countries is scientific and humanistic in content. In the United States most programs lead to the bachelor's degree. Nursing specialists, teachers, and other leaders in the field may need advanced training at the master's or doctoral level. All educational programs include experience with patients in hospitals, homes, or other settings.

In almost all countries with nursing education there are at least two kinds of programs—those leading to diplomas and those that train auxiliaries, though a large portion of auxiliaries in some countries are untrained. A growing number of countries have one or more bachelor's degree programs; some have several. Among the latter are Australia, Canada, Colombia, Peru, Egypt, India, the Philippines, Taiwan, Thailand, the United Kingdom, and the United States. Master's degree programs have also been developed in several countries.

Countries with bachelor's degree programs

Post-basic programs for nurses with diplomas have been established in many countries. Some programs offer courses in general education, as well as nursing courses, and some, in universities, may become programs leading to a bachelor's degree. The purposes of such programs vary and include the preparation of teachers, supervisors, or administrators and of nurse specialists in various fields, including midwifery, public health, and teaching of auxiliaries. Some augment the education received in other programs. Enrollment is generally small in relation to the need for their graduates.

The development of nursing education in any country is affected by the developments in general education. In the United States and some other countries, for example, high school graduation or its equivalent has for many years been a requirement for admission to schools preparing registered nurses. In the United States this is also a requirement for admission to practical-nurse programs. In some countries fewer years of previous education are required.

Licensing and registration. After the number of nurses has become substantial and the essential nature of nursing has become established in a country, the need to regulate the practice of nursing under law becomes evident. These laws are aimed at the protection of the public.

Laws define nursing and establish titles under which nurses practice. Most laws establish boards empowered to give examinations and maintain registers of qualified nurses. Boards also are usually empowered to maintain lists of schools the graduates of which are eligible to take examinations. Further, the laws that regulate nursing usually provide ways to recognize licenses acquired by nurses from other states and countries.

The Nurses Act of 1919 established the General Nursing Council for England and Wales to maintain a register of nurses; similar acts were passed in 1919 for Scotland and in 1922 for Northern Ireland. The advent of the National Health Service (1946) called for enlargement of the council and its functions, and in 1949 an act was passed that consolidated all previous acts and established regional committees to work with schools of nursing for their improvement and for ways to meet new challenges.

In 1979 the Nurses, Midwifery and Health Visitors Act set up the Central Council for Nursing, Midwifery and Health Visiting and national boards for the four parts of the United Kingdom. The Central Council became responsible for the basic course, the registration of nurses, disciplinary machinery, and all post-basic education, including health visiting, district nursing, and school nursing.

In the United States, national and state nursing associations are prime movers in urging legislation for licensure and registration and in securing amendments. The American Nurses' Association gives leadership to state associations and state boards regarding changes in definitions, the desirable provisions of the law, and the legislative process.

The International Council of Nurses exerts leadership among national nursing associations in their efforts to pass and amend laws. The staff of the World Health Organization also has assisted new countries in the licensing process.

Organizations. Near the beginning of the 20th century, nurses began to organize national associations. The purposes of such organizations usually include the promotion of nursing care of high quality, promotion of desirable legislation in nursing and health, formulation of nursing and educational standards, professional development and welfare of nurses, and representation of nursing with other professional associations and government agencies. Many nursing associations publish professional journals. In large countries the national associations have state or provincial and district constituent associations.

An example of a thriving and vigorous national organization is the American Nurses' Association, founded in 1896, which operates aggressive programs. It affects legislation related to health and nursing in the U.S. Congress through its educational and lobbying activities and exerts leadership in the revision of state laws governing registration and the practice of nursing. A strong program for economic security has elevated nurses' salaries and the conditions of patient care, partly through collective bargaining. The official statements of the association set the course of action. Its official publication is the *American Journal of Nursing.* In England a large national nursing organization is the Royal College of Nursing.

In 1899 the International Council of Nurses was founded as a federation of autonomous national nursing associations. Today there are about 100 member associations from most major countries; and the council is working in other countries to assist in the development of national associations and ultimate membership in the council and to encourage national legislation on registration and nursing practice. An international congress is held every four years. The council's journal, *International Nursing Review,* makes a substantial contribution to worldwide literature of nursing. The Florence Nightingale International Foundation, which is related to the council, has conducted an international conference on research in nursing, as well as an international exchange of nurse scholars.

The International Council of Nurses

Several national colleges of nurse-midwifery and an international college work for the improvement of maternal and infant health. Nurses participate in international and national associations for public health, mental health, industrial health, and school health and in such areas as heart-disease, cancer, and respiratory-disease control. The Nursing Section of the American Public Health Association is one of its largest sections. Practical nurses and auxiliaries are usually organized, if at all, in separate organizations. Various groups form special organizations in several countries—deans of schools of nursing and surgical nurses are examples. Nursing councils are parts of regional commissions on higher education in the United States.

A unique organization—the National League for Nursing—combines nurses, related professionals, and other public-spirited citizens in the United States for meeting the nursing needs of people throughout the country. This organization also carries on a program of national accreditation of the various kinds of educational programs. Its journal is *Nursing Outlook.* With the American Nurses' Association it sponsors *Nursing Research.*

ROLES OF INTERNATIONAL AND MULTINATIONAL ORGANIZATIONS

The International Red Cross. The International Red Cross plays two major roles in nursing: it affords educational opportunities for nurses, and it affords nurses opportunity to serve in programs embodying the Red Cross principles of humanity, impartiality, and neutrality. Through participation in Red Cross activities, public-spirited citizens acquire valuable knowledge of the health needs of people and how nursing helps to meet these needs. Also, nurses broaden their concepts of community and worldwide service.

Some national Red Cross societies operate schools of nursing, post-basic educational programs, and training programs for auxiliaries. In addition, many societies carry on programs preparing nurses to teach health practices in the home; this instruction reaches many thousands of people, who thus learn individual responsibility for health and the care of mothers and children and of patients suffering from simple illnesses. Many societies also train nurses to care for people in disasters and emergencies and deploy nurses to sites of emergencies when needed; some nurses volunteer to serve in emergencies that occur in other countries.

National Red Cross societies emphasize community development and youth programs. Societies are urged to be constantly alert to social changes, such as those resulting from urbanization, migration, or drug abuse, and to coordinate their efforts with other voluntary and official agencies. Nurses with knowledge of community conditions can and do contribute to planning and action through board and community memberships, as well as in their participation in the carrying out of Red Cross programs.

The World Health Organization. The World Health Organization has included nursing in its activities from its beginning in 1948. Nurses are included on teams such as those concerned with maternal and child health, malaria, and tuberculosis. Member nations request assistance in developing educational programs for nurses, auxiliaries, and midwives and to organize public health programs and hospitals.

WHO assistance

Countries are assisted in establishing nursing as a part of their national health departments to assure the planning and implementing of the nursing portions of programs. Governments are aided in the establishment of nursing and nursing-education systems, including those in midwifery. Other programs in which assistance is given include upgrading diploma schools; development of basic and post-basic programs in universities; revision of entry requirements; coordination of classroom teaching with the clinical practice of students; adaptation of hospitals and health agencies for students' experience; preparation of public health nurses, administrators, midwives, and teachers, including teachers of auxiliaries and of indigenous midwives. (In many countries more than half of all the births are attended by untrained midwives.)

Fellowships are granted to nurses for overseas study, mostly in the areas of teaching, administration, public health nursing, midwifery, maternal and child health, and for learning to plan health services. When consultants from the World Health Organization work in countries, they strive to leave national counterpart personnel to continue the work. Study outside the country may be needed to develop such personnel.

European Economic Community. In countries of the European Economic Community, nursing is bound by the directives and regulations published by the Council of Ministers in 1977. The Permanent Committee of Nurses lays down broad guidelines for nurses' training and education and each country has a central authority, such as the Central Council of the United Kingdom. In 1979 the nursing directives required universal recognition of certificates and diplomas, and the European Economic Community countries now accept the competence of nurses trained in member states. Also established were a standard of general education to be attained before entry to a school of nursing, the number of hours the training must occupy, and the areas in which students must have theoretical and clinical knowledge. (L.P.Le./Ed.)

Dentistry

Dentistry is that profession concerned with the prevention and treatment of oral disease, particularly disease of the teeth and supporting tissues. In addition to general practice, dentistry includes many specialties and subspecialties, including oral surgery, prosthodontics, periodontics, orthodontics, pedodontics, and public health.

HISTORY OF DENTISTRY

Accounts of toothache and other oral complaints together with suggested forms of treatment have been found in some of the earliest texts of the ancient civilizations of the Middle and Far East, including China, Egypt, India, and Mesopotamia. Sumerian clay tablets probably of 5000–3000 BC record the belief, which persisted for several millennia, that dental decay and toothache were caused by the gnawing away of the tooth by a minute worm. In India a professional class of physicians who included dentistry in their activities described teething as a cause of serious infantile illnesses such as cough, diarrhea, vomiting, and convulsions, a myth that has persisted among some people to the present day.

Ancient remedies

In these early civilizations, a great variety of dental medications and operations, such as filling or extracting decayed teeth and the splinting of loose teeth or fractures of the jaws, were in use. In about 400 BC Hippocrates described many oral diseases, and he is credited by many with having introduced the term aphthae, which is still in use for the painful, but otherwise harmless, oral ulcers that occur particularly in adolescence or early adult life. Another observation made in ancient Greece, that sweet foods caused teeth to decay, was only experimentally confirmed in the 20th century.

Over the centuries many kinds of dental instruments and empirical forms of treatment developed, but dentistry remained essentially a craft included in the activities of surgeons or practiced by itinerant "healers." In the early 16th century in France, dentistry was practiced by barbers, and in England the Barber-Surgeons' Company (later to become the Royal College of Surgeons) was granted a charter by Henry VIII. Surgery and dentistry eventually passed out of the hands of barbers, and the 16th and 17th centuries saw the emergence of dentistry as a specialty with its own literature.

The first textbook in dentistry was published in Leipzig in 1530, and 50 years later students of dentistry were admitted to the University of France. By 1622 a number of men had been granted the title of surgeon-dentist, although the title was not fully established for a number of years after that. During the reign of Louis XIV, the surgeon-dentists formed a separate subdivision of the surgeons' guild and, the year after the subdivision was formed, it became law that those who wished to practice in the field of mouth surgery and artificial restoration had to pass prescribed examinations. At this time some women were permitted to practice dentistry in France, although the privilege was revoked during the mid-1700s.

In the 17th century in England dentistry was referred to as an independent specialty. The first English text on the subject of dentistry was published in 1685.

Textbooks

Publication of three highly significant textbooks occurred during the 18th century. The French dentist Pierre Fauchard's *Le Chirurgien Dentiste* ("The Surgeon Dentist") in 1728 put dental treatment on a more scientific plane than ever before and advocated a broader education for dentists. In 1756 the German dentist Philipp Pfaff's *Abhandlung von den Zähnen* ("Treatise on the Teeth") appeared, and in 1771 an anatomist and surgeon of England, John Hunter, who was giving lectures on dentistry, published *The Natural History of the Human Teeth.* Joseph Fox was appointed dental surgeon at Guy's Hospital in 1799, and at the same time lectures on dentistry were set up in Guy's and in a number of other areas.

The next major advance in dentistry was the introduction of general anesthesia to medicine by two dentists, Horace Wells and William Morton, between 1844 and 1846. In 1884 Sigmund Freud demonstrated the value of cocaine as a local anesthetic. William Stewart Halsted, a surgeon,

carried out the first regional local anesthetic of the inferior dental nerve, using a technique that has remained essentially unchanged.

In 1890 W.D. Miller published his finding that teeth could be caused to decay, in vitro, by the action of mouth bacteria on carbohydrates. More than half a century was to elapse, however, before the bacterial basis for dental caries could be definitively established. Since the 1940s, dental research has increased in range and scale in all of its specialities. The result has been an increasing number of publications and journals and the formation of national and international research associations.

No formal dental schools were established in England until 1858, although dental hospitals had been established earlier. These dental hospitals, which were created as centres to serve the poor, were founded and supported by dentists. For the most part they were independent of medical schools and general hospitals. They did accept some students, however, who partly provided cheap labour to operate the hospital and whose fees were used to support the charitable work undertaken. The honorary dental surgeons were responsible for any teaching that was done. In 1858 the first dental school in the United Kingdom was established by the Odontological Society of London and the second school the following year by the College of Dentists of England. Both were private schools.

Early dental schools

At about the same time the Royal College of Surgeons arranged to hold examinations for licensure for dental surgery. It remained the examining body for dentistry in Great Britain for about 20 years. In 1878 the first Dentists' Act in the United Kingdom was passed and the General Medical Council established a register of those qualified to hold the title of dental surgeon. The act did not, however, prohibit individuals without these qualifications from practicing. The General Medical Council also prescribed a curriculum for the training of dentists. It required two years of preceptorship to learn dental mechanics and three years in a medical school and dental hospital. An inspector of the examination procedures at the colleges in 1897 recommended that the one examination be replaced by three—in preliminary science and dental mechanics and a final examination. These recommendations were put into effect and remained without much alteration until 1922.

While schools were being established and standards for licensure created in England, development had also been going on in North America. Dentistry was being practiced in the United States by the late 1790s, and the first textbooks appeared in the early 1800s. By 1834, when the first meeting of a dental society in the United States was held, services were being provided by three distinct classes of people. First, there were those who had qualified by a course of study in the principles of medicine and surgery. Second, there were those who had obtained a preparatory course of medical study and then begun practicing dentistry without having studied the mechanics of dental operations. These people were held in comparative respect since in due course they did obtain some degree of skill in their operations, although they were criticized for not having obtained this skill before actually inflicting their services on the public. The third group included a great number of charlatans—anyone who decided he would like to practice dentistry. It is this group that brought dentistry into disrepute.

Early dentistry in the United States

During the 1830s two unsuccessful attempts to establish dental schools were made, the first in Kentucky and the second in New York City. The latter failed to get under way because of staff problems. No money was available at that time for salaries, and local dentists evidently felt that they could not spare time from their practice for teaching.

The first dental school in North America was established in Baltimore in 1840. Subsequently many other schools were started, many of which were operated privately or commercially. All of the early schools were separate from universities. The first, and for many years the only, school of dentistry in the United States associated with a university was opened at Harvard in 1867.

This was also the era during which dentistry became organized in Canada. The first dental act in Canada was passed by the legislature of the Province of Ontario in 1868. This act incorporated the Royal College of Dental Surgeons of Ontario and gave it the dual function of teaching and licensing. License requirements stipulated five years of practice in a dental office for registration. Except for a simple act that was passed in Alabama in 1841, the Ontario Act was the first law respecting dental practice in North America. The regulations for registration were fairly simple by today's standards, but with the large number of itinerant dentists, who had no formal training whatever, moving back and forth across the Canadian-U.S. border at that time, even this five-year requirement proved useful.

THE PRACTICE OF DENTISTRY

Licensure requirements. The practice of dentistry is now well controlled, and in all countries of the world in which dentistry is practiced there is a licensing requirement. The licensing authority may be the government or national dental organizations.

Canada. In Canada each province has its own licensing authority. This can be a college, such as the Royal College of Dental Surgeons of Ontario, or an association, such as the Manitoba Dental Association. There is also a national authority, the Dental Examining Board.

Requirements in Canada

The university degree (doctor of dental surgery or doctor of dental medicine) does not in itself entitle the holder to practice but is an academic qualification for presentation to the licensing board under whose jurisdiction the holder wants to obtain a license to practice. The regulations of the provincial licensing boards vary but usually require an examination for licensing.

United States. Licensing authority in the United States is vested in state boards of dental examiners, most of which require an examination. Most states require U.S. citizenship as a prerequisite. Some states require the declaration of intent to become a citizen or the first papers as a requirement for admission to the dental licensure examination; a few do not require citizenship for such admission.

Requirements in the United States

Nationals with foreign diplomas may be admitted to practice if their diplomas were issued by a school approved by the American Dental Association and if they pass the state licensure examination.

Soviet Union. There are three classifications of dentists in the Soviet Union, known as two-, three- and five-year dentists. The two-year dentist is a dental technician who studies for two years beyond secondary school, after which he is eligible to work in a dental laboratory. He cannot treat patients. The three-year dentist can treat patients, but his work is limited to restorative, prosthetics, and prophylactic dentistry. The five-year dentist is a stomatologist, who is on a par with that of the physician in the U.S.S.R. He receives training equal to that of the physician and, in addition, is trained in operative dentistry, crown and bridge dentistry, prosthetics, exodontia, and anesthesia, both general and local. His degree entitles him to operate on the hard as well as all soft tissues in the mouth and throat.

European Economic Community. With the advent of the European Economic Community and the Council of Europe, it has become accepted policy that doctors and dentists should be able to move freely and practice within any of the member countries. For this to be acceptable there has had to be mutual recognition of dental degrees and comparable forms of qualification. The Council of the European Communities has, therefore, agreed on a set of directives that set out the training requirements for dental education in the member states. This has created no difficulties for most European countries, where dentistry has long been recognized as a specialty in its own right. In Italy and Spain, however, where dentistry is a subspecialty of medicine, transitional provisions have had to be made until dental training can be harmonized with that of other member states.

Permission to practice in the United Kingdom is granted by the General Dental Council to those holding (1) a degree or diploma in dentistry or dental surgery conferred in Great Britain or Northern Ireland, (2) a degree or diploma in dentistry or dental surgery granted elsewhere

that has been recognized by the General Dental Council, or (3) a degree or diploma approved by the General Dental Council provided that these graduates have passed the statutory examination written under arrangements made by the General Dental Council.

Requirements in the United Kingdom, West Germany, and Italy

Dentists in West Germany must hold a dental surgeon's diploma, which authorizes private practice without further examination. They must be registered by local dental boards and by health authorities.

In Italy a diploma in dentistry, which allows the use of the title of Specialist in Diseases of the Mouth, Teeth, and Jaws, constitutes a license to practice. Holders of the diploma of Doctor of Medicine have passed examinations in dentistry and for this reason may also practice dentistry but do not have the specialist title.

Japan. Since about 1903 Japanese dentistry has been mainly patterned after that practiced throughout the United States. Those wanting to practice dentistry or dental surgery must be recognized by the national government. Applicants for registration must pass the national examination for dentists and obtain license to practice. These requirements must also be fulfilled by registered medical practitioners wanting to practice dentistry, by Japanese citizens, and by foreigners who have qualified in Japan.

Types of practice. *Private practice.* In such countries as Canada, the United States, the United Kingdom, France, West Germany, and Australia, dentists in private practice constitute the vast majority of all licensed dentists. In such countries as the Soviet Union dentists are employed by the government, and there is little private practice.

Shift from individual to group practice

Dental practice has changed significantly since 1920, without a concurrent change in the basic dental curriculum. Dental procedures have shifted from the repair and extraction of teeth for the relief of pain in 1920 to prevention of disease. Dental practice has also changed in larger urban centres from the isolated private practice common in 1920 to a complex system of groups of professions in a central location. Extensive use is now made of dental hygienists, who often receive the patient from the examining dentist. Dental hygienists provide services such as preventive procedures, prophylaxis, scaling, X rays, and dental-health education. Most practices also use dental assistants.

Another development that has occurred in dental health-care services is the extension of the duties currently carried out by dental auxiliaries. New Zealand has pioneered in the field with the creation of the dental nurse, an auxiliary who is trained to provide dental care for children without the supervision of a dentist. The United Kingdom has also developed the "dental auxiliary," who performs somewhat similar duties but under closer supervision. In Canada and the United States pilot projects have been conducted to test the feasibility of using dental auxiliaries for certain operative procedures in order to increase productivity, quality, and general service to the public.

France may be taken as an example of the development of the practice of dentistry in continental Europe. There are two types of dentists practicing in France, the *chirurgien dentiste* and the stomatologist. The practice of dentistry in France by a *chirurgien dentiste* has, since 1892, been restricted to persons of French nationality, who hold a state diploma and who are registered with the Order of Dentists. The Order of Dentists is responsible for registration and discipline but is not concerned with dental education, which is controlled by the state through the common state diploma.

Stomatologists are practitioners who have a diploma in medicine together with either a diploma in dental surgery or a certificate of special studies (two years) in stomatology. Specialization within the field of dentistry is not encouraged. There are no rules laid down for it nor are there any special courses or diplomas or titles.

Types of dental care

Hospital dental practice. Three types of dental care are normally carried out in the hospital environment: (1) clinical procedures normally provided in a dental office, for ambulatory inpatients and outpatients; (2) bedside care for persons admitted for other medical reasons; and (3) inpatient care for patients admitted to a hospital for purely dental conditions.

Dentists may treat patients in hospitals either privately, on a fee-for-service basis, or under some form of government program, such as the National Health Service in the United Kingdom or the Provincial Medicare Plan (surgery only) in Canada. Hospital dental services have for years been an integral part of dental health care and dental education in the United Kingdom, and such services by hospital dental departments have expanded steadily in the United States and Canada.

Hospital dental departments are normally established in the same manner as any other hospital department and are headed by a chief of service, who has the same status as other chiefs of service within the hospital. In some instances, the chief of the dental department may be responsible to the chief of surgery. There are two types of hospital dental departments—one that is established in a teaching hospital and the other in a general hospital with no teaching component. In the teaching hospital the dental department is associated with a faculty of dentistry and forms an integral part of the undergraduate curriculum and, if they exist, of the graduate and postgraduate programs. One of the chief purposes of hospital dental departments is to make available the service of consultants to other hospital departments and general practitioners. This service is most highly developed in teaching hospitals. Usually certain general dental treatment is provided for inpatients and outpatients.

Hospital dental services or departments are most prevalent in western Europe and the Soviet Union.

Public health practice. Generally typical of dental public health practice in Canada and in many areas of the United States is the program carried on in Ontario. There dentists trained in public health, hygienists, and dental assistants carry out a preventive and educational program basically concerned with the examination of children, the recording of basic dental conditions, and the provision of dental health education.

Military practice. Most countries of the world provide dental-care service for their armed forces. The organization of such a service varies extensively. In Canada the Royal Canadian Dental Corps has the same status as the Royal Canadian Medical Corps, with a brigadier-general as the director. Military service for dentists in the United States is under the U.S. Public Health Service, the chief of service being an assistant surgeon general. In the United Kingdom dental care is provided by three separate dental branches—Navy, Army, and Air Force.

Governmental practice. In certain countries, such as the Soviet Union, all dentists are employed by the government. In certain other countries dentists are required to work a number of years for the government before being considered private practitioners of the type known in Canada and the United States. This requirement may be based on the fulfillment of an obligation for government financial support during undergraduate training, or there may be a government regulation that all dental graduates must work for the state a prescribed number of years. Another example of government practice is in the United Kingdom, where dentists are employed by local authorities to provide dental care under the Maternal and Child Welfare Services and the School Dental Service.

The employment of dentists on a salary basis for the general practice of dentistry is not extensive in the United States or Canada. At the national level it may be the provision of dental care for eligible Indians and Eskimos, war veterans, or inmates of penitentiaries. At a municipal level, dentists may be employed in a school dental service. Dentists in both Canada and the United States commonly agree to provide service for families who qualify for social assistance. They are paid on a fee-for-service basis; the fee schedule is usually set, normally after consultation with the profession, by the agency responsible for the social service plan.

Government medical care was introduced in Japan in the late 1930s. This system was expanded until by 1962 almost the entire population was covered. There are limitations to the services offered by government medical care, as in orthodontics or in preventive dentistry.

Dental specialties and subspecialties. In most countries

The status of the specialist

that recognize specialties in dentistry, the specialist is limited to practice in the specialty and cannot carry out the practice of general dentistry. Where the specialty is thus limited the general dentist may refer patients, and a specialist's practice is mainly on a referral basis. In Britain and in certain provinces in Canada specialists may conduct a general practice.

Orthodontics. Orthodontics aims to prevent and correct malocclusion of the teeth and associated dentofacial incongruities. It has been practiced since the days of the ancient Egyptians, but methods of treatment involving the use of bands and removable appliances have become prominent only since the beginning of the 20th century. The United States gave impetus to the development of orthodontics, which was recognized as a specialty with the formation of the American Society of Orthodontists in 1900.

The demand for this service extends from the child to the mature adult, although human bone responds to tooth movement best in a person under 18, and it is generally agreed that children benefit from treatment more than do adults. In general, oral health and physical appearance are the two most important reasons for undertaking a course of orthodontic care.

Dental care for children

Pedodontics. Pedodontics, analogous to pediatrics in medicine, is concerned with the dental care of children and adolescents.

Much of the routine of practice is centred around the control of caries (tooth decay) and involves the use of fluoride and dietary and hygienic instruction. The need to influence tooth positions presents the next most frequently encountered problem. The correction of incipient abnormalities in tooth alignment may obviate the necessity for lengthy treatment. Many pedodontists use growth-influencing techniques to correct jaw alignments. A working knowledge of children's behaviour patterns, patience, and a knowledge of childhood physical and mental diseases and their ramifications are important qualifications of the pedodontist.

Periodontics. Periodontics is concerned with the prevention, diagnosis, and treatment of diseases of the periodontal tissues—the tissues that surround and support the teeth. These tissues consist mainly of the gums and the jaw bones and their related contiguous structures.

The most prevalent periodontal disease is periodontitis, commonly called pyorrhea, an inflammatory condition usually produced by local irritants. Periodontitis, if untreated, destroys the periodontal tissues and is a major cause of the loss of teeth in adults.

The advances of periodontics have been mostly in techniques of treatment. It is believed that bacterial plaque, a soft layer of substances rich in bacteria that adheres to the teeth, is the factor responsible for most destruction of the gums and the tissues surrounding the teeth. Periodontists advocate removal of such plaque by a specific regimen of controlled hygiene.

Prosthodontics. Prosthodontics is concerned with the restoration and maintenance of oral function, comfort, appearance, and health by the replacement of missing teeth and contiguous tissues with artificial substitutes, or prostheses.

Branches of prosthodontics

There are three main branches of the specialty, concerned, respectively, with removable prostheses, fixed prostheses, and maxillofacial prostheses. Maxillofacial prostheses are supplied to persons who have suffered congenital, traumatic, or surgical defects of the mouth, jaws, or associated facial structures.

The proper fitting of oral prostheses requires a detailed knowledge of the anatomy of the head and neck, of the physiology of the neuromuscular system, and of the science of occlusion and jaw movements. It also requires skill in planning, mouth preparation, impression making, registration of jaw relations, try-in procedures, placement of the prostheses, and follow-up care.

Oral medicine. Oral medicine, or stomatology, treats, the variety of diseases that affect both the skin and the oral mucous membranes. Some of these diseases, such as pemphigus vulgaris, can have their first manifestations in the mouth and can be life-threatening. Cancer of the mouth also has a high mortality rate, partly because it grows in such close proximity to so many vital structures and readily involves them. With all such diseases of the oral cavity, removal of a portion of the lesion for examination under the microscope (biopsy) by an oral pathologist is essential, and many other laboratory procedures are often also required for the diagnosis of oral mucosal diseases.

Oral pathology. Oral pathology is the study of the causes, processes, and effects of oral disease, together with the resultant alterations of oral structure and functions. The oral pathologist provides diagnoses on which treatment by other specialists will depend.

Oral surgery. Oral surgery deals with the diagnosis of, and the surgery required by, the diseases, injuries, and defects of the human jaws and associated structures.

Both dentists and physicians refer a wide variety of special dental problems to the oral surgeon. These may include the removal of impacted and infected teeth and the treatment of cysts, tumours, lesions, and infections of the mouth and jaws. In addition there are more complex problems, such as jaw and facial injuries, cleft palate, and cleft lip.

Public health dentistry. Dental public health is recognized as a specialty in Canada and the United States. The American Dental Association recognizes dental public health as a specialty if the holder of the master's degree proceeds to a further year of study in training and passes the examination of the American Board of Dental Public Health. Training in dental public health is available in the United Kingdom, but the specialty is not emphasized to the same degree in the rest of the world.

Forensic dentistry. Forensic dentistry is the study and practice of aspects of dentistry that are relevant to legal problems. It is a specialty practiced by few and is not usually part of dental education. Forensic dentistry is, however, of considerable legal importance for several reasons, one of the most important of which is the fact that the teeth are the structures of the body most resistant to fire or putrefaction. Moreover, the arrangement of the teeth or any restoration in them is virtually or completely unique to any given individual and, if dental records can be found, may enable identification with certainty similar to that provided by fingerprinting. The identification of human remains after aircraft accidents, for example, can often be made only by this means. Minor irregularities of the teeth can also be reproduced in bite marks, enabling a suspect to be identified if he has bitten another person.

Dental education. *Predental programs.* In a majority of countries in the world undergraduate training in dentistry is available. Many require predental training prior to acceptance into a school of dentistry. The predental training is in addition to primary and secondary education, which usually takes from 10 to 12 years. The required number of years in predental education varies from one to seven (a number of European countries require from five to seven years of medical education before entering dentistry). Predental course training usually includes such studies as biology, chemistry, physics, and mathematics. Certain faculties of dentistry in Canada and the United States require a bachelor's degree in arts or science as a prerequisite for admission into a dental faculty.

Dental school and training. After predental courses, training consists of four years in a faculty of dentistry to qualify as a doctor of dental surgery (D.D.S.) or doctor of dental medicine (D.M.D.). The program of studies during the four-year course includes the following biological sciences: human anatomy, biochemistry, bacteriology, histology, pathology, pharmacology, microbiology, and physiology, upon which the succeeding studies of the theory and techniques of dental practice are based. Studies required with respect to dental practice include restorative dentistry, prosthetics, orthodontics, surgery, preventive dentistry, medicine, dental public health, pedodontics, periodontics, radiology, clinical practice, and anesthesia.

ANCILLARY DENTAL FIELDS

Dental hygienists. The hygienist is a figure in the campaign to reduce periodontal disease and to improve physical well-being by promoting better care of the mouth.

The prevention of oral disease through education and

Chief function of hygienists

treatment is the chief function of hygienists. The specific duties and services that they are allowed to perform depend on the bylaws of the licensing bodies, the requirements of the dental offices in which they are employed, or the aims and objectives of the public health programs in which they are engaged. At all times hygienists work under the effective supervision of a qualified dentist, and they are not permitted to establish their own practice.

Hygienists employed in dental offices remove deposits and stains from the patient's teeth and apply fluorides. In addition, they observe and record conditions of decay and disease for the dentist's information. Further duties may include the taking of X-ray photographs of parts of the mouth, which the hygienist develops and mounts. Another function of the hygienist is to promote dental health by advising on diet and nutrition and encouraging oral hygiene.

Hygienists employed by educational authorities assist school dentists by examining children's teeth and performing other duties as required. They may also visit classrooms to explain the importance of oral hygiene and to give instruction in the proper care of the teeth and gums. In hospitals they perform mainly the same duties as for private practitioners.

Dental nurses and dental auxiliaries. In New Zealand an auxiliary known as the dental nurse has been carrying out a dental-care program for children for a number of years. The dental nurse receives minimal supervision but is equipped to provide a dental-care program for children up to 13 years of age. The dental nurse is trained for two years in a special course for dental nurses with entrance requirements below the university level .

Training and functions of dental nurses

Dental assistants. The majority of dentists in private practice employ one or more dental assistants to provide such services as the reception of patients, the keeping of records and accounts, assistance for the dentist while he is treating patients, general upkeep of the office, developing of dental X rays, and sterilization of instruments.

Dental technicians and dental mechanics. Dental technicians, also called dental mechanics, make artificial crowns, bridges, dentures, and other dental appliances according to dentists' specifications. Work orders, accompanied by models or impressions of patients' mouths, state the exact requirements for each particular job. In large laboratories the various stages of manufacture are often divided and the technicians employed may specialize. Sometimes partially skilled persons are hired to work in limited aspects of production on an assembly-line basis.

ORGANIZATIONS

Associations of dentists, dental journals, and dental schools now exist in almost every country of the world. The Fédération Dentaire Internationale (International Dental Federation) was founded in 1900 and has met annually except in times of war. It has sponsored international dental congresses that are planned to meet every five years. Other international organizations include the International Association for Dental Research (Association Internationale pour la Recherche Dentaire) and the Association pour les Recherches sur les Paradentopathies (Association for Research into Periodontal Diseases), which was organized in 1932. *The International Dental Journal,* published by the Fédération Dentaire Internationale, was founded in 1950.

Within the general framework of the World Health Organization, the dental health program has progressed steadily from the beginning. A proposal for a joint review of stomatology and dental hygiene in collaboration with the International Dental Federation was made at the first World Health Assembly in 1948.

Certain organizations, including the World Health Organization and Fédération Dentaire Internationale, and countries such as New Zealand and the United States offer direct and financial assistance to many developing countries in the development of health educational and dental care services. For example, New Zealand has long provided developing countries with the benefit of its experience in the use of dental auxiliaries or what is commonly known as school dental nurses. Direct assistance is provided in the development of other public health dental services to countries such as Sri Lanka, Malaysia, Singapore, Brunei, Thailand, Indonesia, Hong Kong, and Papua New Guinea. Dentists from these countries have had the opportunity of studying the New Zealand system and a number of school dental nurses have received their training there, enabling them to assist in the establishment of their own training facilities. (R.A.Co.)

Pharmacy

Pharmacy is the science and art concerned with the preparation and standardization of drugs. Its scope includes the cultivation of plants that are used as drugs, the synthesis of chemical compounds of medicinal value, and the analysis of medicinal agents. Pharmacists are responsible for the preparation of the dosage forms of drugs, such as tablets, capsules, and sterile solutions for injection. They compound physicians', dentists', and veterinarians' prescriptions for drugs. The science that embraces knowledge of drugs with special reference to the mechanism of their action in the treatment of disease is pharmacology.

HISTORY OF PHARMACY

The beginnings of pharmacy are ancient. When the first person expressed juice from a succulent leaf to apply to a wound, this art was being practiced. In the Greek legend, Asclepius, the god of the healing art, delegated to Hygieia the duty of compounding his remedies. She was his apothecary or pharmacist. The physician-priests of Egypt were divided into two classes: those who visited the sick and those who remained in the temple and prepared remedies for the patients.

In ancient Greece and Rome and during the Middle Ages in Europe, the art of healing recognized a separation between the duties of the physician and those of the herbalist, who supplied the physician with the raw materials from which to make medicines. The Arabian influence in Europe during the 8th century AD, however, brought about the practice of separate duties for the pharmacist and physician. The trend toward specialization was later reinforced by a law enacted by the city council of Bruges in 1683, forbidding physicians to prepare medications for their patients. In America, Benjamin Franklin took a pivotal step in keeping the two professions separate when he appointed an apothecary to the Pennsylvania Hospital.

Ancient and medieval pharmacy

The development of the pharmaceutical industry since World War II led to the discovery and use of new and effective drug substances. It also changed the role of the pharmacist. The scope for extemporaneous compounding of medicines was much diminished and with it the need for the manipulative skills that were previously applied by the pharmacist to the preparation of bougies, cachets, pills, plasters, and potions. The pharmacist continues, however, to fulfill the prescriber's intentions by providing advice and information; by formulating, storing, and providing correct dosage forms; and by assuring the efficacy and quality of the dispensed or supplied medicinal product.

THE PRACTICE OF PHARMACY

Education. The history of pharmaceutical education has closely followed that of medical education. As the training of the physician underwent changes from the apprenticeship system to formal educational courses, so did the training of the pharmacist. The first college of pharmacy was founded in the United States in 1821 and is now known as the Philadelphia College of Pharmacy and Science. Other institutes and colleges were established soon after in the United States, Great Britain, and continental Europe. Colleges of pharmacy as independent organizations or as schools of universities now operate in most developed countries of the world.

The course of instruction leading to a bachelor of science in pharmacy extends at least five years. The first and frequently the second year of training, embracing general education subjects, are often provided by a school of arts and sciences. Many institutions also offer graduate courses in pharmacy and cognate sciences leading to the degrees of master of science and doctor of philosophy in pharmacy, pharmacology, or related disciplines. These

advanced courses are intended especially for those who are preparing for careers in research, manufacturing, or teaching in the field of pharmacy.

The basic five-year curriculum

Since the treatment of the sick with drugs encompasses a wide field of knowledge in the biological and physical sciences, an understanding of these sciences is necessary for adequate pharmaceutical training. The basic five-year curriculum in the colleges of pharmacy of the United States, for example, embraces physics, chemistry, biology, bacteriology, physiology, pharmacology, and many other specialized courses. As the pharmacist is engaged in a business as well as a profession, special training is provided in merchandising, accounting, computer techniques, and pharmaceutical jurisprudence.

Licensing and regulation. To practice pharmacy in those countries in which a license is required, an applicant must be qualified by graduation from a recognized college of pharmacy, meet specific requirements for experience, and pass an examination conducted by a board of pharmacy appointed by the government.

Pharmacy laws generally include the regulations for the practice of pharmacy, the sale of poisons, the dispensing of narcotics, and the labeling and sale of dangerous drugs. The pharmacist sells and dispenses drugs within the provisions of the food and drug laws of the country in which he practices. These laws recognize the national pharmacopoeia (which defines products used in medicine, their purity, dosages, and other pertinent data) as the standard for drugs. The World Health Organization of the United Nations began publishing the *Pharmacopoeia Internationalis* in the early 1950s. Its purpose is to standardize drugs internationally and to supply standards, strengths, and nomenclature for those countries that have no national pharmacopoeia.

Research. Pharmaceutical research, in schools of pharmacy and in the laboratories of the pharmaceutical manufacturing houses, embraces the organic chemical synthesis of new chemical agents for use as drugs and is also concerned with the isolation and purification of plant constituents that might be useful as drugs. Research in pharmacy also includes formulation of dosage forms of medicaments and study of their stability, methods of assay, and standardization.

Bioavailability

Another facet of pharmaceutical research that has attracted wide medical attention is the problem of the "availability" to the body (bioavailability) of various dosage forms of drugs. Exact methods of determining levels of drugs in blood and organs have revealed that slight changes in the mode of manufacture or the incorporation of a small amount of inert ingredient in a tablet may diminish or completely prevent its absorption from the gastrointestinal tract, thus nullifying the action of the drug. Ingenious methods have been devised to test the bioavailability of pharmaceutical dosage forms. Although such in vitro, or test-tube, methods are useful and indicative, the ultimate test of bioavailability is the response of the patient to dosage form of the drug.

Licensing systems for new medicinal products in Europe and North America demand extensive and increasingly costly investigation and testing in the laboratory and in clinical trials to establish the efficacy and safety of new products in relation to the claims to be made for their use. Proprietary rights for innovation by the grant of patents and by the registration of trademarks have become increasingly important in the growth of the pharmaceutical industry and its development internationally.

The results of research in pharmacy are usually published in journals, such as the *Journal of Pharmacy and Pharmacology* (London), the *Journal of the American Pharmaceutical Association* and the *Journal of Pharmaceutical Sciences* (Washington, D.C.), the *American Journal of Pharmacy* and the *American Journal of Hospital Pharmacy* (Philadelphia), and the *Pharmaceutica Acta Helvetiae* (Zürich).

ORGANIZATIONS

There are numerous national and international organizations of pharmacists. The Pharmaceutical Society of Great Britain, established in 1841, is typical of pharmaceutical organizations. In the United States the American Pharmaceutical Association, established in 1852, is a society that embraces all pharmaceutical interests. Among the international societies is the Fédération Internationale Pharmaceutique, founded in 1910 and supported by some 50 national societies, for the advancement of the professional and scientific interests of pharmacy on a worldwide basis. The Pan American Pharmaceutical and Biochemical Federation includes the pharmaceutical societies in the various countries in the Western Hemisphere.

International organizations

There are also other international societies in which history, teaching, and the military aspects of pharmacy are given special emphasis. (J.C.K./Fr.H.)

LEGAL ASPECTS OF MEDICINE

Maintenance of professional standards

HISTORY

Physicians historically have set their own standards of care and their conduct has usually been judged by comparing it to that of other physicians. "Ethical" canons or codes generally focused on professional etiquette and courtesy toward fellow physicians rather than on relationships with patients. The Hippocratic oath was a notable exception, but its provisions were only ascribed to by a minority of Greek physicians.

The law has become intimately involved in medical practice only in the 20th century. Until recently legal medicine, or forensic medicine, was a field devoted exclusively to the uses of medicine in the courtroom, primarily in two settings: forensic pathology and forensic psychiatry. The pathologist has traditionally been asked to determine and testify to the cause of death in cases of suspected homicide and to aspects of various injuries involving crimes such as assault and rape. Pathological testimony may also be required in civil cases involving, for example, occupational injury, negligent injury, automobile accidents, and paternity suits. Similarly, when a defendant pleads insanity as a defense, a psychiatrist is asked to examine the defendant and to testify as to his mental state at the time of the crime. The relevant question is usually whether his criminal behaviour was the product of a mental illness or whether the defendant was able to distinguish right from wrong. In civil cases, psychiatrists frequently appear as witnesses in cases of child custody and involuntary commitment.

Changes since 1960

Since 1960 the legal climate has changed drastically. Civil lawsuits alleging medical malpractice have become a fact of professional life for many Western physicians. Issues formerly relegated to ethics, such as abortion and termination of treatment, also have become important civil rights issues in courtrooms across the world, as have issues of informed consent and patients' rights. Wide-ranging campaigns aimed at arresting the spread of infectious diseases, such as acquired immune deficiency syndrome (AIDS), have involved the legal system in issues of privacy, confidentiality, quarantine, and research using human subjects.

So great has the change been that forensic medicine has now become a subspecialty of a separate field, usually called health law to emphasize its application not only to medicine but to health care in general. This new field of health law is not limited to the courtroom but is active as well in legislatures, regulatory agencies, hospitals, and physicians' offices.

RELATIONSHIP OF LAW AND ETHICS

The legal philosopher Lon Fuller has distinguished between "the morality of aspiration" and "the morality of duty." The former may be denoted ethics, the latter law. Ethics tells people what they should do and embodies the ideals they should strive to attain. Unethical behaviour leads to punishments that are related to how an individ-

ual is perceived, both by himself and by his fellow man. Law, on the other hand, provides boundaries of actions, set by society, beyond which a person may go only by risking external sanctions, such as incarceration or loss of a medical license.

The Hippocratic oath

This may explain why ethical codes usually involve generalities, while laws tend to be more specific. For example, the Hippocratic oath, formulated in the 5th century BC, is concerned with the physician's doing no harm, refraining from performing abortions and giving deadly drugs, and maintaining strict confidentiality. Law, on the other hand, may permit abortions under certain circumstances, permit the giving of potentially lethal drugs in extreme situations, and sanction the violation of confidentiality when the interests of society demand it.

For example, in treating a patient with AIDS, a physician might administer a potentially fatal experimental drug in a desperate attempt to destroy the retrovirus that causes AIDS. The physician might also warn the patient's spouse that there is a danger that the virus will be transmitted, even if the patient does not want the spouse to be so informed. These scenarios present ethical dilemmas, and it is unlikely that any of these actions would have legal consequences today.

Although the Hippocratic oath has largely been superseded by such modern oaths as the Declaration of Geneva, the International Code of Medical Ethics, and the Canons of the American Medical Association, these codes of conduct retain the brevity and generality of the Hippocratic oath. For example, the International Code of Medical Ethics, developed and promulgated by the World Medical Association shortly after World War II, provides in part for the following:

> A doctor must always maintain the highest standards of professional conduct.
>
> A doctor must practice his profession uninfluenced by motives of profit. . . .
>
> A doctor must always bear in mind the obligation of preserving human life. . . .
>
> A doctor shall preserve absolute secrecy on all he knows about his patient because of the confidence intrusted in him. . . .

Effects of modern technology on ethics

Modern advancements in the field of medicine, such as cardiopulmonary resuscitation (which restores regular rhythm to an arrhythmic or failed heart) and mechanical respirators (which breathe for patients unable to expand their lungs) sometimes have been able to postpone a death that previously had been imminent. Under these circumstances, it may be difficult to relate the rules of ethics to the realities of the situation. For example, the meaning of "obligation of preserving human life" becomes unclear in the context of a young woman in a permanent coma who will probably die if the mechanical respirator is removed, but who may live for decades (in a coma) if the machine remains in place. It is not clear that the Hippocratic ideal of doing "no harm" requires that the machine remain in place or that it be removed.

In 1976 these same questions were confronted by the New Jersey Supreme Court in the landmark case of Karen Ann Quinlan. Her parents requested her physicians to remove the mechanical respirator in order to let their daughter die a natural death. The doctors refused, relying primarily on medical ethics, which they believed prohibited taking an action that might lead to the death of the patient.

In court, however, the lawyers for the Quinlan family argued that what was at stake was not medical ethics but the legal rights of the individual patient to refuse medical treatment that was highly invasive and offered no chance for a cure. The court agreed that patients have the legal right to refuse medical treatment, determined that honouring such a refusal was consistent with medical ethics, and decided that the parents of Karen Ann Quinlan could exercise her right to refuse treatment on her behalf. In order to reassure the physicians involved, however, the court also decreed that if a hospital's ethics committee agreed with the prognosis of permanent coma, the removal of the respirator could take place and all parties involved would have legal immunity from civil or criminal prosecution. Karen Ann Quinlan's respirator was removed, although she continued to breathe on her own. She survived in a coma until she died of pneumonia almost 10 years later.

The case of Karen Ann Quinlan has become a parable of modern medicine and of the relationship between medical ethics and the law. Although an issue was made of medical ethics by both the physicians and the court, the case primarily involved medical practice and the fear of potential legal liability. Modern physicians now also worry about the law as well as ethics, and they fear criminal lawsuits that allege homicide or assisting suicide, and civil lawsuits that allege malpractice. To address these concerns, the New Jersey court created an ethics committee with the power to grant legal immunity for actions and to diffuse the responsibility for them.

Ethics committees

This model has not been followed by other courts, however, although ethics committees have been established in North America, Europe, and Australia to help educate hospital staffs on such matters as the withholding and withdrawing of treatment and on general ethical conduct with patients. In fact, physicians are rarely taken to court on criminal charges for decisions about patient care that are made in good faith. The Massachusetts Supreme Court, for example, has summarized the criminal law in this regard: "Little need be said about criminal liability: there is precious little precedent, and what there is suggests that the doctor will be protected if he acts on a good faith judgment that is not grievously unreasonable by medical standards."

In the United States, the Quinlan rationale has been expanded to include the right of all mentally competent patients, whether terminally ill or comatose, to refuse any and all medical treatments (including artificial feedings). Some people state their wishes about treatment in documents called "living wills," to which physicians can refer after the patient is no longer able to speak for himself. People also may provide someone with power of attorney to make decisions about medical treatment once they become incompetent to act on their own behalf. The American Medical Association has stated that honouring patients' refusals of treatment is consistent with both medical practice and medical ethics. Other countries, like The Netherlands, have gone further and have held that it is legally and ethically acceptable for physicians to assist a patient with a terminal illness in his decision to die by providing lethal injections.

While ethics and law are concerned with different concepts of right and wrong, in medicine they find common ground in their fundamental principles. Both law and ethics in medicine rest on the principle of self-determination by competent individuals, beneficence (or at least nonmaleficence) on the part of medical practitioners, and a concept of justice as fairness to be afforded to all patients by both medical practitioners and society.

Law and medical practice

The law indirectly influences medical practice by structuring the delivery and financing of medical services and it does so directly in three major ways: licensure requirements; restrictions on practice; and redress for wronged patients.

GOVERNMENT FINANCING

Societies differ as to how much medical care is considered adequate. In the United States, the President's Commission for the Study of Ethical Problems in Medicine concluded in 1983 that even though access to health care might not be properly considered a legal right, "society has a moral obligation to ensure that everyone has access to adequate care without being subject to excessive burdens." Programs for governmental payment of medical bills are an example of law used as an instrument to effect social policy. Changing the laws regarding financing rules may, in fact, have more of an impact on medical practice than any other legal interventions.

LICENSE REQUIREMENTS FOR PRACTICE

Legislative approaches to health-care licensing have been divided into three general categories by the legal scholar

Jan Stepan: (1) exclusive or monopolistic systems, in which only the practice of modern, scientific medicine by professionals is lawful; (2) tolerant systems, in which only scientific medicine is officially sanctioned, although traditional medicine is tolerated; and (3) inclusive or integrated systems, in which traditional medicine is either recognized as a special part of the health system, or the integration of two or more systems of health care is officially promoted.

Requirements

The exclusive systems. The exclusive system of licensing, which first emerged in Europe, involves granting licenses only to those who meet certain formal educational requirements; they are often required to pass a standardized examination and to demonstrate that they are of good moral character. The educational requirements usually apply to physicians, nurses, pharmacists, dentists, and other allied health professionals. Most medical schools in the world have similar courses of study, and a person who is graduated from almost any of them receives the M.D. degree (doctor of medicine) and is entitled to sit for a licensure examination. In Britain and many Commonwealth countries the graduate receives a bachelor of medicine and a bachelor of surgery or bachelor of chirurgery (M.B., B.S.; or M.B., Ch.B.); the M.D. is a higher degree that requires further study. The license to practice, called registration, is granted by the General Medical Council, which was established by the Medical Act of 1858.

Because of the similarities between educational and practice patterns physicians often may be allowed to practice in countries other than those in which they are registered. In 1975 the Council of the European Economic Communities, for example, issued directives for the national recognition of primary and specialist medical qualifications so as to facilitate the free movement of doctors among the member states of the European Economic Community (EEC). Steps have been taken in most EEC countries to implement these directives. An Advisory Committee in Medical Training has been established by the EEC to assist in the formation of a common standard.

Legislatures originally adopted exclusive licensing in an effort to protect the public from untrained persons on the belief that the public was not qualified to judge medical skills and could easily be taken advantage of by unqualified practitioners. Licenses are generally granted for life and can be lost only by criminal conduct, gross incompetence, or mental disability. Virtually every country in the world, with the exceptions of China and Japan, initially followed the European model of exclusive licensing, and the type of health care this model imposed is based on Western science. What may have been reasonable for Europe and North America, however, makes little sense in rural Africa or South America, and it has become clear that more flexibility is needed to ensure better access to health care in many of the developing countries.

Approaches

Exclusive approaches to medical care exist in almost every state in the United States and in France, Belgium, and most other European countries. In the United States, for example, it is a criminal offense for anyone to practice medicine without a license. The practice of medicine is usually defined to include diagnosis, treatment, prescription, and surgery. Physicians may lawfully practice in any branch of medicine and any of its subspecialties. Further education, training, and certification as a specialist is available through private, voluntary organizations. All other licensed health professionals, however, are strictly limited by statute to the areas in which they may practice. For example, dentists must only do work relating to dentistry.

Licensing

Because the U.S. states have retained direct powers in health care, all licenses for health professionals are granted by the individual states. Licensing is regulated by administrative agencies, usually called state licensing boards, and licenses are valid only in the state issuing the license. Most states, however, grant a license by reciprocity upon request from a licensed physician, nurse, pharmacist, or dentist who is in good standing in another state. Licensing rules governing dentistry and pharmacy are parallel to medical licensure; the candidate must have been graduated from a professional school, have passed an examination, and be licensed by a state board set up for that purpose.

Registered nurses (R.N.'s) are required to have been graduated from a preparatory nursing program (with a bachelor's degree, associate degree, or diploma from a hospital-based program) and to have passed an examination. Licensed practical nurses (L.P.N.'s rather than R.N.'s), on the other hand, generally have to complete a one-year vocational program after graduating from high school. In almost all countries, national and local nursing associations are instrumental in proposing legislation for licensing or registration and in securing amendments to broaden the scope of nursing practices.

The tolerant systems. Tolerant systems operate in countries such as Britain, West Germany, the Scandinavian countries, and some states in the United States, such as California. In Britain, for example, there is no formal legal monopoly of medicine. The protected status is that of registered medical practitioner. Under the Medical Act of 1978 persons who have fulfilled statutory education and examination requirements are entitled to be registered.

Physicians' rights

Registered physicians have certain exclusive rights, such as employment by the National Health Service, prescribing, issuing medical certificates, and holding appointments in public hospitals. It is an offense to imply falsely that one is registered, but it is not an offense to otherwise engage in healing. Osteopaths, chiropractors, and acupuncturists may practice in the private sector. The Dental Act of 1984, however, gives dentists a monopoly, by making it an offense for anyone to practice dentistry who is not registered as a dentist. Sweden is similar in that anyone may be a healer, but it is a crime for a person to claim to be a physician when he is not so recognized by the state.

The inclusive and integrated systems. The inclusive and integrated systems are represented by countries such as India and China. In India the individual states regulate the practice of the health professions. A Medical Council maintains the Indian Medical Register, and any person on this register may practice anywhere in India. There are also state medical registers for the individual states, and people in these registers are also in the Indian Medical Register. Those who have obtained a degree from a university, usually M.B., B.S., or the qualification of licensed medical practitioner (L.M.P.) are entered in these registers. Qualifications from many foreign countries are also recognized for entry to the registers. Although it is now a crime to practice medicine without a license in India, this law was only enacted in 1970, and it does not apply to anyone who had practiced medicine without a license for at least five years before 1970.

China provides an example of an integrated system in which traditional Chinese doctors and Western-trained physicians practice side-by-side. The country has also developed a system in which certain persons receive a short, intensive training course and then provide primary health care in areas otherwise not served by health-care workers.

LEGAL RESTRICTIONS ON PRACTICE

Determination of death. The law generally supports customary medical practice and provides the medical profession with a great deal of autonomy. A dramatic example is the determination of death and the issuance of a death certificate. In almost every country of the world a physician declares a person dead and issues a death certificate after a determination of death is made in accordance with accepted medical standards. A question that recently appeared was whether physicians should continue to be given the authority to declare a person dead if the medical profession were to adopt whole brain death as an acceptable definition of death (instead of the past definition of irreversible cessation of respiration and heartbeat). A mechanical respirator can artificially maintain the respiration and circulation of a person whose functions would cease without such mechanical support. In the late 1960s the potentials of organ transplantation from such persons were becoming realized, and the seeming futility of devoting limited medical resources to maintaining circulation under such circumstances was of growing concern. Physicians began proposing that irreversible cessation of brain activity be used as an alternative definition of death.

Evolution of a standard

Since that time most Western countries have adopted this definition, either by continuing to permit physicians

to declare death, passing a specific statute endorsing this definition, or issuing court opinions giving approval to physicians' declarations of death in such circumstances. The law, in short, has continued to defer to medical practice in the definition of death itself.

The countries that have not adopted brain death criteria have not done so primarily for cultural and religious reasons. For example, Japan has refused to adopt a brain-based definition of death in part because it would conflict with religious tenets that require the death of all major organs prior to a pronouncement of death. Accordingly, such medical techniques as heart transplantation cannot be performed in Japan (or in any country that does not accept a brain-death definition). Harvesting a beating heart from a person is considered in such countries to be homicide, even though brain activity has ceased and respiration is being maintained artificially.

Termination of pregnancy. Physicians have broad legal authority and discretion in other controversial areas as well. In countries like the United States, for example, where the law permits termination of pregnancy prior to fetal viability, it is for the physician, and not the state, to determine whether or not an individual fetus is viable (*i.e.,* capable of living independently of its mother); the determination must be made consistent with accepted medical criteria. When the Supreme Court declared that the decision to terminate a pregnancy was protected by the U.S. Constitution, it emphasized that the decision should be made by a woman and her physician. The law may restrict procedures like termination of pregnancy, sterilization, and even birth control to mature minors and adults; but it is generally left to the physician to determine if a patient is mature or competent to consent.

Sudden death. Unexpected deaths must be reported to public authorities so that a determination can be made as to whether the death was a homicide, a suicide, or an accident. It is up to an investigator, whether he is called a medical examiner or coroner, or has some other title, to make a preliminary finding and then to refer the case to the police or prosecuting authorities if criminal activity is suspected. In most Western countries this person has either legal training, medical training, or both. In the United States, some jurisdictions use an elected coroner (who may or may not have medical training); however, the trend is toward a medical examiner system in which a physician is appointed to the post. Coroners in London are qualified both in medicine and in law. The principal evidence received by the coroner is the report of the autopsy carried out on the body of the deceased.

Public reporting. Physicians may also be expected to report certain patients to public authorities. For example, some communicable diseases are required to be reported to public health officials. Suspected child abuse and gunshot wounds may have to be reported to an authority (such as the child welfare authority or the police). Public reporting tends to put the physician in the position of an agent of the state rather than of his patient, making mandatory reporting uncomfortable. It is not surprising, therefore, that the incidence of public reporting is often thought to be lower than may be warranted.

Effect of public reporting

In extreme cases, physicians may also have a duty to warn specific individuals who may be at serious or mortal risk from their patients. For example, the California Supreme Court decreed that a psychologist had a "duty to warn" a person whom his patient had threatened to kill, if the psychologist believed "or should have believed" the threat to be real. In that particular case the patient, a graduate student, left the psychologist's care and murdered his former girlfriend, an action that the psychologist believed was so likely that he had sought unsuccessfully to have his patient involuntarily committed.

LEGAL REDRESS

When patients are injured by medical negligence the remedies they can pursue depend upon the country's legal system. In the United States, for example, lawsuits against physicians for negligent injury are not considered unusual.

Malpractice, or professional negligence, is the failure of a health-care provider (for example, a physician, dentist, nurse, or pharmacist) to exercise the ordinary care and skill a reasonably prudent, qualified person would exercise under the same or similar circumstances. The practitioner does not guarantee the outcome but must use diligence and ordinary skill in the treatment of a patient.

A valid malpractice claim must have four elements: duty, breach, damages, and causation. The practitioner must be shown to have a relationship to the patient (which establishes a duty to exercise ordinary care); must have breached that duty (as measured by the applicable standard of care); and through the breach must have caused the patient physical and monetary damages.

The central concern for physicians is usually to establish the standard of care through expert testimony, which may simply be the testimony of another qualified physician. Such testimony is necessary because the standard of medical practice is not something a lay jury is familiar with. Expert witnesses may themselves rely on the standards that have been set down by one or more of the medical speciality organizations, such as the American College of Obstetricians and Gynecologists. These medical speciality organizations provide certification to physicians who have fulfilled postgraduate training and practice requirements in the speciality. They maintain the standards necessary to practice in the specialities, and they provide reasonable assurance to patients that these standards will be upheld. Nonconformance with such standards by a specialist is evidence of negligence, although it is not conclusively negligence (the practitioner may have a valid excuse for not following custom, such as an emergency situation or lack of equipment). Conformance with the standards is evidence of due care, but it is not conclusive because other factors may have caused the physician's action to be imprudent under the circumstances.

If a practitioner consistently performs below the profession's standard of care (*i.e.,* the practitioner is a negligent physician who does not actually harm anyone) the remedy is not a malpractice action but a complaint to the licensing or registration authority to have the individual disciplined. Disciplinary action by public licensing authorities, however, is unusual.

Functions of malpractice

Medical malpractice actions have three basic functions: quality control, compensation for harm, and emotional vindication, all of which are achieved to varying degrees. Quality control is probably best achieved, since the standard of care is set by physicians themselves and enforced by patients and juries. Compensation for harm, on the other hand, is skewed toward major injuries. Attorneys in the United States, for example, represent malpractice cases on a contingency fee basis—*i.e.,* they are paid a proportion (usually 20–40 percent) of the total amount awarded to the plaintiff. Patients who suffer less severe injuries may have little redress for compensation. In countries that have a system of national health insurance, compensation for harm may not be a major issue (since all medical bills are paid regardless of cause). And countries with comprehensive social services for all citizens, like Sweden and New Zealand, have effectively developed "no fault" compensation systems. But in the United States, where more than 35,000,000 people do not have any form of health-care insurance, lack of coverage can transform a medically induced injury into a financial catastrophe.

Emotional vindication is a measure of the consumer's ability to make a complaint as well as to get a satisfactory response. A comparison in consumer complaints between the United States and Britain shows significant differences. In 1981 an estimated 700 writs for medical malpractice were issued in Britain. In 1983 about 42,000 claims were filed in the United States. These figures yield a per capita difference (even when the difference in years and between claims and writs is considered) that shows that Americans file claims against physicians more than 10 times as often as their British counterparts. The law professor Frances Miller has noted that many cultural and practical reasons serve to explain this difference, including different legal systems and rules, access to attorneys and courts, the method of paying for medical expenses, the special status of the National Health Service in Britain, and the existence of alternative complaint procedures.

Patients' rights

In addition to granting patients means for the effective redress for negligent injury (which increases the cost of malpractice insurance for physicians—and thus the cost of medical care), malpractice litigation has also promoted what have come to be called patients' rights.

Fundamental premises

Patients' rights are based upon two fundamental premises: (1) the patient has certain interests, many of which may be properly described as rights, that are not automatically forfeited by entering into a relationship with a doctor or health-care facility; and (2) doctors and health-care facilities may fail to recognize the existence of these interests and rights, fail to provide for their protection or assertion, and frequently limit their exercise without recourse.

Perhaps the most important development in patient rights has been that in the United States regarding the doctrine of informed consent. This doctrine requires physicians to share certain information with patients before asking for their consent to treatment. The doctrine is particularly applicable to the use of surgery, drugs, and invasive diagnostic procedures that carry risks. It requires the physician to describe the procedure or treatment recommended and to list its major risks, benefits, alternatives, and likely prospect for recuperation. The purpose is to promote self-determination by patients on the theory that it is the patient who has the most at stake in treatment and who relies largely on the physician for such information. British courts have rejected this formulation on the basis that the average British citizen does not want such information, and British physicians do not generally provide it unless requested.

In 1972 the American Hospital Association adopted a patient bill of rights based on the premise that "[the] traditional physician-patient relationship takes on a new dimension when care is rendered within an organizational structure . . . the institution itself also has a responsibility to the patient." The text of the American Hospital Association bill of patient rights calls for the rights of the patient to respectful care, complete medical information, information necessary for informed consent, refusal of treatment, privacy, confidentiality, response to requests for service, information on other institutions involved in the patient's care, refusal of participation in research projects, continuity of care, examination and explanation of financial charges, and knowledge of hospital regulations.

In the mid-1970s the Parliamentary Assembly of the Council of Europe submitted a draft recommendation to its member nations suggesting that all necessary action be taken to ensure that the sick can receive relief from their suffering and that patients be adequately prepared for death. They further recommended that commissions be established to study the issue of euthanasia and that physicians acknowledge that the sick have a right to full information, if requested, on their illness and the proposed treatment. The council listed the following basic rights for the humane and dignified treatment of patients: the right to freedom, to personal dignity and integrity, to information, to proper care, and to not suffer.

No one bill of rights is suitable for every health care institution, and specialty hospitals, such as maternity and pediatric hospitals, may require different approaches. Among

Proposed basic rights

the basic rights of a patient should be the right to clear communication: accurate information concerning possible medical care and procedures; informed participation in all decisions involving the patient's health-care program; and a clear, concise explanation of all proposed procedures, including possible risks, side effects, and problems related to recuperation.

Patients also should have rights regarding quality of care: a right to an accurate evaluation of their condition and prognosis without treatment; knowledge of the identity and professional status of those providing services; information contained in their medical record; access to consultant specialists; and refusal of treatment.

The patient should have basic human rights: the right to privacy of both person and information; of access to people outside the health-care facility; and to leave the health-care facility regardless of his condition.

The goals of such a system are to protect patients, especially those at a disadvantage within the health-care system (*e.g.*, the young, the illiterate, the uncommunicative, and those without relatives); to make available to those who seek it the opportunity to participate actively with the physician as a partner in a personal health-care program; and to put into proper perspective medical technology and pharmaceutical advances.

Until the 1960s law and medicine met only in the courtroom, and then usually only in cases involving pathology or psychiatry. Since then, however, civil litigation, public financing, and ethical issues have grown, at least partially as a result of the incredible successes of medicine. These successes have increased public expectations and increased the cost of medicine; they also have made decisions about terminating care more ambiguous. Enhancing patients' rights is one modern concern on which both medical and legal practitioners agree. (G.J.A.)

BIBLIOGRAPHY

General works: *Webster's Medical Desk Dictionary* (1986), is a reference source for the layman. *The Oxford Companion to Medicine,* 2 vol., edited by JOHN WALTON, PAUL B. BEESON, and RONALD BODLEY SCOTT (1986), is a comprehensive text of 20th-century developments and persons.

History: The literature on the history of medicine covers all topics and periods and includes biographies as well as descriptions of the development of hospitals, research institutes, health care, and medical education in different countries. Introductory studies include GEORGE T. BETTANY, *Eminent Doctors: Their Lives and Their Work,* 2 vol. (1885, reprinted 1972); ARTURO CASTIGLIONI, *A History of Medicine,* 2nd rev. ed. (1947; originally published in Italian, 1927), a classic work; FIELDING H. GARRISON, *An Introduction to the History of Medicine,* 4th rev. ed. (1929, reprinted 1967), a scholarly history; DOUGLAS GUTHRIE, *A History of Medicine,* rev. ed. (1958); HOWARD W. HAGGARD, *Devils, Drugs, and Doctors: The Story of the Science of Healing from Medicine-Man to Doctor* (1929, reprinted 1980); RICHARD H. MEADE, *An Introduction to the History of General Surgery* (1968), a well-documented work on developments in surgery on separate organs; CHARLES SINGER and E. ASHWORTH UNDERWOOD, *A Short History of Medicine,* 2nd ed. (1962); and PHILIP RHODES, *An Outline History of Medicine* (1985).

Ancient traditions of non-Western medicine are presented in P. KUTUMBIAH, *Ancient Indian Medicine* (1962); HEINRICH R. ZIMMER, *Hindu Medicine* (1948, reprinted 1979); EDWARD H. HUME, *The Chinese Way in Medicine* (1940, reprinted 1975); PAUL U. UNSCHULD, *Medicine in China: A History of Ideas* (1985; originally published in German, 1980); and EDWARD G. BROWNE, *Arabian Medicine* (1921, reprinted 1983).

For developments from the origin of Western medicine to the end of the 18th century, see WILLIAM G. BLACK, *Folk-Medicine: A Chapter in the History of Culture* (1883, reprinted 1970); W.H.R. RIVERS, *Medicine, Magic, and Religion* (1924, reprinted 1979), a comprehensive treatment of primitive medicine; JOHN SCARBOROUGH, *Roman Medicine* (1969, reprinted 1976); ROBERT S. GOTTFRIED, *Doctors and Medicine in Medieval England, 1340–1530* (1986); A. WEAR, R.K. FRENCH, and I.M. LONIE (eds.), *The Medical Renaissance of the Sixteenth Century* (1985); KATHARINE PARK, *Doctors and Medicine in Early Renaissance Florence* (1985); and GUY WILLIAMS, *The Age of Agony: The Art of Healing, c. 1700–1800* (1975, reprinted 1986).

Medicine and surgery during the 19th and 20th centuries are the subject of CARL J. PFEIFFER, *The Art and Practice of Western Medicine in the Early Nineteenth Century* (1985); THOMAS E. KEYS, *The History of Surgical Anesthesia,* rev. ed. (1963, reprinted 1978); M.H. ARMSTRONG DAVISON, *The Evolution of Anesthesia* (1965); ROBERT G. RICHARDSON, *The Scalpel and the Heart* (1970; U.K. title, *The Surgeon's Heart: A History of Cardiac Surgery,* 1969); JOHN S. HALLER, JR., *American Medicine in Transition, 1840–1910* (1981); RUTH J. ABRAM (ed.), *Send Us a Lady Physician: Women Doctors in America, 1835–1920* (1985); A. MCGEHEE HARVEY, *Science at the Bedside: Clinical Research in American Medicine, 1905–1945* (1981), a discussion of the institutionalization of clinical research; and LAWRENCE GALTON, *Med Tech: The Layperson's Guide to Today's Medical Miracles* (1985), a historical dictionary.

Medical practice: GEORGE ROSEN, *The Structure of American Medical Practice, 1875–1941* (1983), is a historical study. Particular kinds of medical practice are explored in WESLEY FABB and JOHN FRY (eds.), *Principles of Practice Management in Primary Care* (1984); SIR DOUGLAS BLACK *et al., Inequalities in Health: The Black Report,* edited by PETER TOWNSEND and NICK DAVIDSON (1982); DAVID SANDERS and RICHARD CARVER, *The Struggle for Health: Medicine and the Politics of Underdevelopment* (1985); and V. DJUKANOVIC and E.P. MACH (eds.),

Alternative Approaches to Meeting Basic Health Needs in Developing Countries: A Joint UNICEF/WHO Study (1975). Also see the articles of such journals as *Private Practice* (monthly) and *Modern Healthcare* (semimonthly). For a view of alternative medicine, see DOUGLAS STALKER and CLARK GLYMOUR (eds.), *Examining Holistic Medicine* (1985); and RICHARD GROSSMAN, *The Other Medicines* (1985). Health-care professions are the subject of LOUISE SIMMERS, *Diversified Health Occupations* (1983); C. WESLEY EISELE, WILLIAM R. FIFER, and TOMA C. WILSON, *The Medical Staff and the Modern Hospital* (1985); and ELI GINZBERG (ed.), *From Physician Shortage to Patient Shortage: The Uncertain Future of Medical Practice* (1986).

Among the many books devoted to the subject of medical and health-care education are the following historical discussions: ABRAHAM FLEXNER, *Medical Education in the United States and Canada: A Report to the Carnegie Foundation for the Advancement of Teaching* (1910, reprinted 1973); and KENNETH M. LUDMERER, *Learning to Heal: The Development of American Medical Education* (1985). For special information, see the following official publications: ASSOCIATION OF AMERICAN MEDICAL COLLEGES, *AAMC Directory of American Medical Education, 1986–87*, 33rd ed. (1986), *Medical School Admission Requirements, 1988–89*, 38th ed. (1987), and *Physicians for the Twenty-first Century: Report to the Project Panel on the General Professional Education of the Physician and College Preparation for Medicine* (1984). Studies include MOHAN L. GARG and WARREN M. KLEINBERG, *Clinical Training and Health Care Costs: A Basic Curriculum for Medical Education* (1985); and MARJORIE PRICE WILSON and CURTIS P. MCLAUGHLIN, *Leadership and Management in Academic Medicine* (1984).

For new developments in medical education, see the periodicals *The Journal of Medical Education* (monthly), *Medical Education* (bimonthly), and *WHO Chronicle* (bimonthly). Opportunities for continuing medical education appear semiannually in *JAMA: The Journal of the American Medical Association* (weekly). Education for health-care professionals other than physicians is the subject of COMMITTEE FOR THE STUDY OF NURSING EDUCATION (U.S.), *Nursing and Nursing Education in the United States* (1923, reprinted 1984), a classic yet still timely report on nursing training; BRYN DAVIS (ed.), *Nursing Education: Research and Developments* (1987); SHIRLEY A. BADASCH and DOREEN S. CHESEBRO, *The Health Care Worker: An Introduction to Health Occupations* (1985); JOSEPH S. GREEN *et al.* (eds.), *Continuing Education for the Health Professions* (1984); and T. DONALD RUCKER (ed.), *Pharmacy: Career Planning and Professional Opportunity* (1981).

Medical institutions: For historical discussions of hospitals, see TIMOTHY S. MILLER, *The Birth of the Hospital in the Byzantine Empire* (1985); GUENTER B. RISSE, *Hospital Life in Enlightenment Scotland: Care and Teaching at the Royal Infirmary of Edinburgh* (1986); and RICHARD H. THURM, *For the Relief of the Sick and Disabled: The U.S. Public Health Service Hospital in Boston, 1799–1969* (1972). Organization and administration of modern hospitals are discussed in I. DONALD SNOOK, JR., *Hospitals: What They Are and How They Work* (1981); JONATHON S. RAKICH and KURT DARR (eds.), *Hospital Organization and Management*, 3rd ed. (1983); THOMAS CHOI, ROBERT F. ALLISON, and FRED MUNSON, *Governing University Hospitals in a Changing Environment* (1986); HOWARD J. BERMAN, LEWIS E. WEEKS, and STEVEN F. KUKLA, *The Financial Management of Hospitals*, 6th ed. (1986); BRADFORD H. GRAY (ed.), *The New Health Care for Profit: Doctors and Hospitals in a Competitive Environment* (1983); and EVERETT A. JOHNSON and RICHARD L. JOHNSON, *Hospitals Under Fire: Strategies for Survival* (1986).

For developments in the field of hospitals, see the periodical *Hospitals* (semimonthly), published by the American Hospital Association. Social and psychological aspects of hospital life are the subject of GEOFFREY C. ROBINSON and HEATHER F. CLARKE, *The Hospital Care of Children: A Review of Contemporary Issues* (1980); and JUDITH WILSON ROSS, *Handbook for Hospital Ethics Committees* (1986). Hospital building is analyzed in W. PAUL JAMES and WILLIAM TATTON-BROWN, *Hospitals: Design and Development* (1986); OWEN B. HARDY and LAWRENCE P. LAMMERS, *Hospitals, the Planning and Design Process*, 2nd ed. (1986); and I. DONALD SNOOK, JR., and KATHRYN M. RUCK (eds.), *Using Hospital Space Profitably* (1987).

For organization and administration of other health facilities, see MARTIN P. CHARNS and MARGUERITE J. SCHAEFER, *Health Care Organizations, a Model for Management* (1983); and BRADFORD H. GRAY (ed.), *For-Profit Enterprise in Health Care* (1986). The following works discuss aspects of the activities of outpatient facilities: DREXEL TOLAND and SUSAN STRONG, *Hospital-Based Medical Office Buildings*, 2nd ed. (1986); PETER M. FRIEND and JOHN M. SHIVER (eds.), *Freestanding Emergency Centers: A Guide to Planning, Organization, and Management* (1985); and ROBERT KOHN, *The Health Centre Concept in Primary Health Care* (1983).

Facilities for geriatric care are discussed in WILLIAM B. BREGER and WILLIAM R. POMERANZ, *Nursing Home Development: A Guide to Planning, Financing, and Constructing Long-Term Care Facilities* (1985); and COLLEEN L. JOHNSON and LESLIE A. GRANT, *The Nursing Home in American Society* (1985). Medical institutions for terminal care are the subject of CHARLES A. CORR and DONNA M. CORR (eds.), *Hospice Care: Principles and Practice* (1983); GLEN W. DAVIDSON (ed.), *The Hospice: Development and Administration*, 2nd ed. (1985); JACK MCKAY ZIMMERMAN, *Complete Care for the Terminally Ill*, 2nd ed. (1986); and CHARLES A. CORR and DONNA M. CORR (eds.), *Hospice Approaches to Pediatric Care* (1985); as well as of articles in *The American Journal of Hospice Care* (bimonthly).

Public health: FRASER BROCKINGTON, *World Health*, 3rd ed. (1975), is a comprehensive discussion of public health concepts and the World Health Organization. JOHN BRYANT, *Health & the Developing World* (1969, reprinted 1972), studies in health-care in Africa, Asia, and Latin America. JOHN M. LAST (ed.), *Public Health and Preventive Medicine*, 11th ed. (1980), is a definitive text. Later surveys of the organized effort to protect and improve community health include DEREK FRASER, *The Evolution of the British Welfare State: A History of Social Policy Since the Industrial Revolution*, 2nd ed. (1984); ROBERT LANZA (ed.), *Medical Science and the Advancement of World Health* (1985); and GRACE BUDRYS, *Planning for the Nation's Health: A Study of Twentieth-Century Developments in the United States* (1987).

Nursing: The long historical tradition of the nursing occupation is studied in JOHN HOWARD, *An Account of the Principal Lazarettos in Europe* (1789); BRIAN ABEL-SMITH, *A History of the Nursing Profession* (1960, reprinted 1975); VERN L. BULLOUGH and BONNIE BULLOUGH, *The Emergence of Modern Nursing*, 2nd ed. (1969); CECIL BLANCHE WOODHAM SMITH, *Florence Nightingale, 1820–1910* (1950, reissued 1983); MONICA E. BALY, *Florence Nightingale and the Nursing Legacy* (1986); BARBARA MELOSH, *The Physician's Hand: Work Culture and Conflict in American Nursing* (1982); and M. PATRICIA DONAHUE, *Nursing, the Finest Art: An Illustrated History* (1985).

ESTHER LUCILE BROWN, *Nursing Reconsidered: A Study of Change*, 2 vol. (1970–71), reviews the developments of the 1960s and analyzes new functions performed by nurses. See also MONICA E. BALY, *Nursing and Social Change*, 2nd ed. (1980); BONNIE BULLOUGH, VERN L. BULLOUGH, and MARY CLAIRE SOUKUP, *Nursing Issues and Nursing Strategies for the Eighties* (1983); and NORMA L. CHASKA (ed.), *The Nursing Profession: A Time to Speak* (1983). Nursing practices are explored in SANDRA DEBELLA, LEONIDE MARTIN, and SANDRA SIDDALL, *Nurses' Role in Health Care Planning* (1986); MARY H. BROWNING (comp.), *The Nursing Process in Practice* (1974), and *Nursing and the Aging Patient* (1974); MARY H. BROWNING and EDITH P. LEWIS (comps.), *Nursing and the Cancer Patient* (1973), and *The Nurse in Community Mental Health* (1972); ANDREW JAMETON, *Nursing Practice: The Ethics Issues* (1984); and CAROL REN KNEISL and SUEANN WOOSTER AMES (eds.), *Adult Health Nursing: A Biopsychosocial Approach* (1986).

Dentistry: PIERRE FAUCHARD, *The Surgeon Dentist: or, Treatise on the Teeth*, 2 vol. in 1 (1946, reprinted 1969; originally published in French, 2nd ed., 1746) is a classic work. See also VINCENZO GUERINI, *A History of Dentistry: From the Most Ancient Times Until the End of the Eighteenth Century* (1909, reprinted 1967); and BERNHARD WOLF WEINBERGER, *An Introduction to the History of Dentistry, with Medical & Dental Chronology & Bibliographic Data*, 2 vol. (1948, reprinted 1981). MALVIN E. RING, *Dentistry: An Illustrated History* (1985), is a readable comprehensive survey. For a broad range of current information on dentistry, see *Dental Abstracts* (monthly), a compilation of abstracts of articles from periodicals; and *The Year Book of Dentistry*, an annual review of selected articles.

R.A. CAWSON and J.W. EVESON, *Oral Pathology and Diagnosis* (1987), is a text and colour atlas. Oral and dental surgery is the subject of W. HARRY ARCHER, *Oral and Maxillofacial Surgery*, 5th ed., 2 vol. (1975); GUSTAV O. KRUGER, *Textbook of Oral and Maxillofacial Surgery*, 6th ed. (1984); and DANIEL M. LASKIN (ed.), *Oral Surgery* (1985). Other dental specialties are covered in JOHN M. DAVIS, DAVID B. LAW, and THOMPSON M. LEWIS, *An Atlas of Pedodontics*, 2nd ed. (1981); RALPH E. MCDONALD and DAVID R. AVERY (eds.), *Dentistry for the Child and Adolescent*, 4th ed. (1983); THOMAS M. GRABER and BRAINERD F. SWAIN, *Orthodontics: Current Principles and Techniques* (1985); LEE W. GRABER (ed.), *Orthodontics: State of the Art, Essence of the Science* (1986); LOUIS I. GROSSMAN, *Endodontic Practice*, 10th ed. (1981); IRVING GLICKMAN, *Glickman's Clinical Periodontology*, 6th ed., edited by FERMIN A. CARRANZA (1984); HENRY M. GOLDMAN and D. WALTER COHEN, *Periodontal Therapy*, 6th ed. (1980); HENRY M. GOLDMAN *et al.* (eds.), *Current Therapy in Dentistry*, 7th ed. (1980); JOHN F. JOHNSTON, *Johnston's Modern Practice in Fixed Prosthodontics*, 4th ed., edited by ROLAND W. DYKEMA, CHARLES J. GOODACRE, and RALPH W. PHILLIPS (1986); EUGENE

WILLIAM SKINNER, *Skinner's Science of Dental Materials,* 8th ed., edited by RALPH W. PHILLIPS (1982); ROBERT G. CRAIG (ed.), *Restorative Dental Materials,* 7th ed. (1985); and FRANK M. MCCARTHY (ed.), *Emergencies in Dental Practice: Prevention and Treatment,* 3rd ed. (1979).

Prophylactic measures in dentistry are discussed in JOSEPH L. BERNIER and JOSEPH C. MUHLER (eds.), *Improving Dental Practice Through Preventive Measures,* 3rd ed. (1975); and ABRAHAM E. NIZEL, *Nutrition in Preventive Dentistry: Science and Practice,* 2nd ed. (1981). DAVID F. STRIFFLER, WESLEY O. YOUNG, and BRIAN A. BURT, *Dentistry, Dental Practice, & the Community,* 3rd rev. ed. (1983); and JAMES MORSE DUNNING, *Principles of Dental Public Health,* 4th ed. (1986), are studies of the relationship between the dentist and society.

Pharmacy: EDWARD KREMERS, *Kremers and Urdang's History of Pharmacy,* 4th ed., rev. by GLENN SONNEDECKER (1976), provides comprehensive coverage of the topic. Useful information can be found in UNITED STATES PHARMACOPEIAL CONVENTION, *Advice for the Patient: Drug Information in Lay Language,* 6th ed. (1986). Modern studies of practices in pharmacy include WILLIAM E. HASSAN, JR., *Hospital Pharmacy,* 5th ed. (1986); THOMAS R. BROWN and MICKEY C. SMITH (eds.), *Handbook of Institutional Pharmacy Practice,* 2nd ed. (1986); PAUL TURNER, ALAN RICHENS, and PHILIP ROUTLEDGE, *Clinical Pharmacology,* 5th ed. (1986); JOSEPH P. REMINGTON, *Remington's Pharmaceutical Sciences,* 17th ed., edited by ALFONSO R. GENNARO *et al.* (1985); D.R. LAURENCE and P.N. BENNETT, *Clinical Pharmacology,* 5th ed. (1980); and ALFRED GOODMAN GILMAN *et al.* (eds.), *Goodman and Gilman's The Pharmacological Basis of Therapeutics,* 7th ed. (1985).

Legal medicine: H. TRISTRAM ENGELHARDT, JR., *The Foundations of Bioethics* (1986), gives an overview of medical ethics. EMILY FRIEDMAN (ed.), *Making Choices: Ethics Issues for Health Care Professionals* (1986), is a selection of articles on decision-making in the rationing of health care and in the provision of critical care. UNITED STATES. PRESIDENT'S COMMISSION FOR THE STUDY OF ETHICAL PROBLEMS IN MEDICINE AND BIOMEDICAL AND BEHAVIORAL RESEARCH, *Securing Access to Health Care: A Report on the Ethical Implications of Differences in the Availability of Health Services,* 3 vol. (1983), is an authoritative document.

Issues accompanying the medical problems of special neonatal care are discussed in EARL E. SHELP, *Born to Die?: Deciding the Fate of Critically Ill Newborns* (1986); and SHERMAN ELIAS and GEORGE J. ANNAS, *Reproductive Genetics and the Law* (1987). The ethics of general critical care are the subject of STUART J. YOUNGNER (ed.), *Human Values in Critical Care Medicine* (1986); and ALONZO L. PLOUGH, *Borrowed Time: Artificial Organs and the Politics of Extending Lives* (1986). Patients' rights in view of the law of informed consent are explored in GEORGE J. ANNAS, *The Rights of Hospital Patients: The Basic ACLU Guide to a Hospital Patient's Rights* (1975); JOANNE LYNN (ed.), *By No Extraordinary Means: The Choice to Forgo Life-Sustaining Food and Water* (1986); WILLIAM J. WINSLADE and JUDITH WILSON ROSS, *Choosing Life or Death: A Guide for Patients, Families, and Professionals* (1986); and JAMES RACHELS, *The End of Life: Euthanasia and Morality* (1986).

Laws and legislation regulating health care are analyzed in GEORGE J. ANNAS, LEONARD H. GLANTZ, and BARBARA F. KATZ, *The Rights of Doctors, Nurses, and Allied Health Professionals: A Health Law Primer* (1981); SAMUEL C. URSU (ed.), *Symposium on Legal Considerations in Dentistry* (1982); FRANCES H. MILLER, "Medical Malpractice Litigation: Do the British Have a Better Remedy?", *American Journal of Law & Medicine,* 11:433–463 (1986), a comparison of the systems of medical malpractice litigation in the United States and Britain; PATRICIA M. DANZON, *Medical Malpractice: Theory, Evidence, and Public Policy* (1985); RAYMOND G. DEVRIES, *Regulating Birth: Midwives, Medicine, & the Law* (1985); and ROBERT D. MILLER, *Problems in Hospital Law,* 5th ed. (1986).

A summary of the major models of licensing of health-care professionals throughout the world is given in JAN STEPAN, "Traditional and Alternative Systems of Medicine: A Comparative Review of Legislation," *International Digest of Health Legislation,* 36:281–341 (1985). On professional accreditation and licensing, see DONALD G. LANGSLEY (ed.), *Legal Aspects of Certification and Accreditation* (1983); CLARE LABAR, *Statutory Requirements for Licensure of Nurses* (1985); and COUNCIL OF STATE GOVERNMENTS, *State Credentialing of the Health Occupations and Professions* (1986).

Forensic medicine is the subject of R.B.H. GRADWOHL, *Gradwohl's Legal Medicine,* 3rd ed., edited by FRANCIS E. CAMPS (1976), a comprehensive, fully illustrated text of traditional legal medicine. JOHN GLAISTER, *Glaister's Medical Jurisprudence and Toxicology,* 13th ed., edited by EDGAR RENTOUL and HAMILTON SMITH (1973), is a comprehensive text on forensic pathology and toxicology. See also WERNER U. SPITZ and RUSSELL S. FISHER (eds.), *Medicolegal Investigation of Death: Guidelines for the Application of Pathology to Crime Investigation,* 2nd ed. (1980); SIR JOHN JERVIS, *Jervis on the Office and Duties of Coroners,* 10th ed., edited by PAUL MATTHEWS and J.C. FOREMAN (1986); and WILLIAM J. CURRAN and E. DONALD SHAPIRO, *Law, Medicine, and Forensic Science,* 3rd ed. (1982).

(P.Rh./Fr.H./G.J.A./Ha.Sc./Ed.)

Mediterranean Sea

The Mediterranean Sea, the great expanse of waters that, stretching from the Atlantic Ocean on the west to Asia on the east, separates Europe from Africa, often has been called the incubator of Western civilization.

This ancient "sea between the lands" occupies a deep, elongated, and almost landlocked irregular depression lying between 30° and 46° N and between 5°50′ W and 36° E. Its east–west extent, from the Strait of Gibraltar to the shores of the Gulf of Iskenderum, is approximately 2,500 miles (4,000 kilometres), while its average north–south extent, between the shores of Yugoslavia and Libya, is about 500 miles. The Mediterranean Sea, including the Sea of Marmara and the Black Sea, occupies an area of 1,145,000 square miles (2,966,000 square kilometres); its area excluding the Black Sea is close to 970,000 square miles.

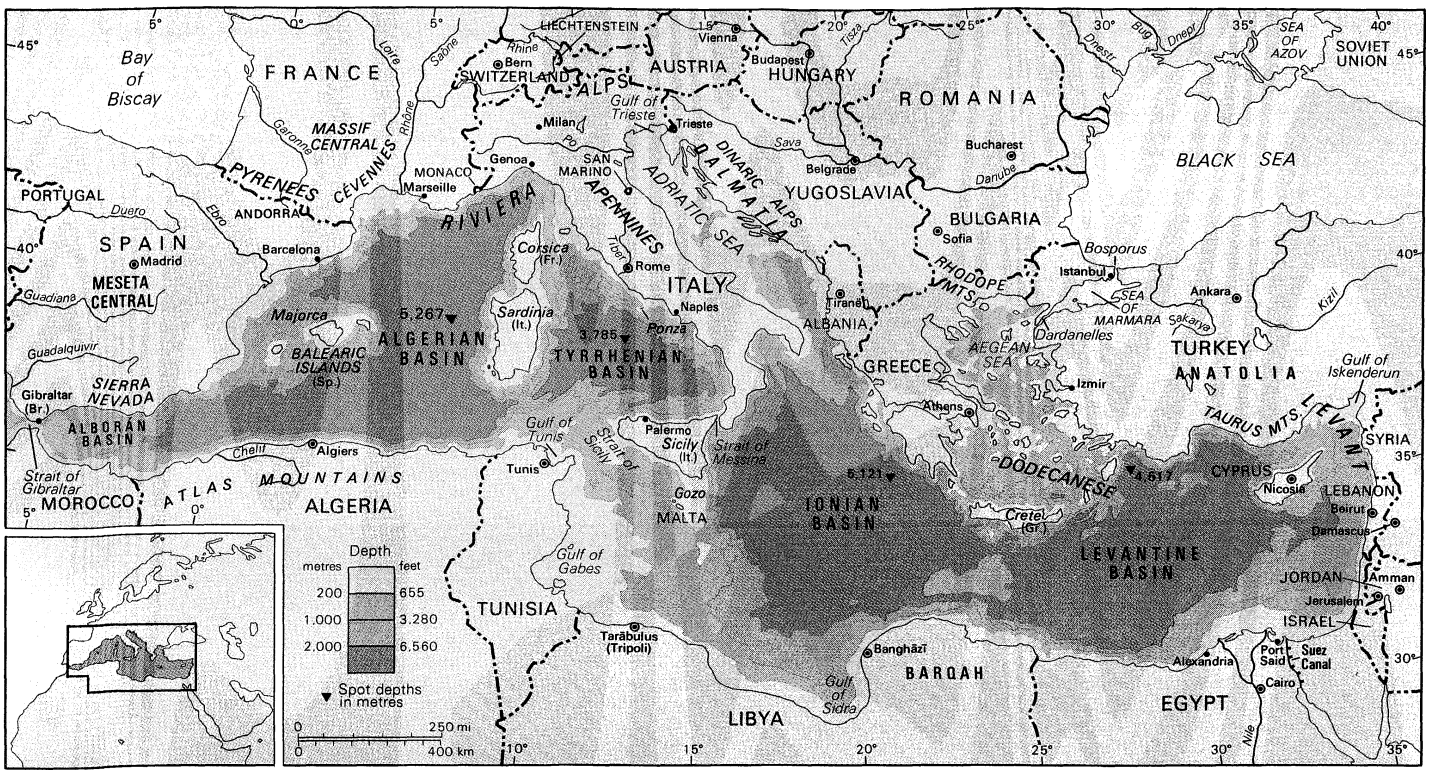
The western extremity of the Mediterranean connects with the Atlantic Ocean by the narrow and shallow channel of the Strait of Gibraltar, which is only eight miles wide at its narrowest point, and the depth of the sill, or submarine ridge separating the basins, is 1,050 feet (320 metres). To the northeast the Mediterranean Sea is connected with the Black Sea through the Dardanelles (with a sill depth of 230 feet), the Sea of Marmara, and the strait of the Bosporus. To the southeast it is connected with the Red Sea by the Suez Canal.

The article surveys the physical features of the Mediterranean and traces the history of the sea's influence on the various cultures of the region. The article is divided into the following sections:

Physical features 939
  Natural divisions 939
  Geology and physical geography 940
  Hydrography 941
  Climate 941
  Marine and plant life 942
History 942
  Early Mediterranean cultures 942
  19th and 20th centuries 943
  History of scientific study 943

Physical features

Natural divisions. A submarine ridge between the island of Sicily and the African coast, with a sill depth of about 1,200 feet, divides the Mediterranean Sea into western and eastern parts. The western part, in turn, is subdivided into three principal submarine basins. The Alborán Basin is east of Gibraltar, between the coasts of Spain and Morocco. The Algerian (sometimes known as the Algero-Provençal, or Balearic) Basin, east of the Alborán Basin, is west of Sardinia and Corsica, extending from off the coast of Algeria to off the coast of France. The Tyrrhenian Basin,



The Mediterranean Sea and its basin.

that part of the Mediterranean known as the Tyrrhenian Sea, lies between Italy and the islands of Sardinia and Corsica. The eastern Mediterranean is subdivided into two major basins. The Ionian Basin, in the area known as the Ionian Sea, lies to the south of Italy and Greece, where the deepest sounding in the Mediterranean, about 16,000 feet (4,900 metres), has been recorded. A submarine ridge between the western end of Crete and Barqah (Libya) separates the Ionian Basin from the Levantine Basin to the south of Asia Minor, and the island of Crete separates it from the Aegean Sea, which comprises that part of the Mediterranean Sea north of Crete and bounded on the west and north by the coast of Greece and on the east by the coast of Turkey. The Aegean Sea contains numerous islands known as the Grecian archipelago. The Adriatic Sea, northwest of the main body of the Mediterranean Sea, is bounded by Italy to the west and north and by Yugoslavia and Albania to the east.

Geology and physical geography. The Mediterranean is the main existing remnant of the Tethys Sea, which formerly girdled the Eastern Hemisphere. The structure and present form of the basin and its bordering mountain system have been determined by the convergence and recession of the relatively stable continental blocks of Eurasia and Africa. The interpretation of geological data suggests that there are, at present, areas of constriction between Africa and Eurasia the amplitude of which appears to be several centimetres a year. The use of laser tellurometry has demonstrated the intermittent nature of this phenomenon.

The western Mediterranean. The Tyrrhenian Basin of the western Mediterranean has two exits into the eastern Mediterranean, the Strait of Sicily and the Strait of Messina, both of which were of great strategic importance throughout Mediterranean history. The submarine relief of the Sicilian channel is rather complicated; the group of islands composed of Malta, Gozo, and Comino, all of which are of Tertiary limestone, stands on a submarine shelf that extends southward from Sicily.

Continental shelves and coasts

The widest of the continental shelves is off Spain's Ebro Delta, where it extends to 60 miles. Similarly, west of Marseille the shelf widens off the Rhône Delta area to 40 miles. Off the Italian peninsula the shelf is narrow, and the slopes are mostly quite low. The narrow shelves continue along the French Riviera, but the slopes show a marked increase where they are cut by many canyons and troughs.

Along the Atlas Mountain coast of North Africa, a narrow shelf bordered by slopes is found from the Strait of Gibraltar to the Gulf of Tunis. The slopes are cut by many canyons or troughlike indentations.

The coasts of the western Mediterranean, just as those of the eastern basin, have been subjected in recent geological times to the uneven action of deposition and erosion. This action, together with the movements of the sea and the emergence and submergence of the land, resulted in a rich variety of types of coasts. The Italian peninsula underwent considerable uplift in post-Pliocene times (the last 2,500,000 or so years), as a result of which a strip of Late Tertiary rocks has been exposed on the Adriatic flank of the Apennines. The Italian Adriatic coast exemplifies a coast owing much to emergence. Examples of coasts the outline of which is due to submergence are the granite coast in northeast Sardinia and the Dalmatian coast where the eroded land surface has been sunk, producing elongated islands parallel to the coast. The deltas of the Rhône, Po, Ebro, and Nile rivers are good examples of coasts resulting from the deposition of silt.

The eastern Mediterranean. The Sicilian straits scarcely exceed 1,500 feet in depth, so that there is essentially a shelf from Tunisia to Sicily separating the Mediterranean into two parts. South of the straits the shelf widens to as much as 170 miles off the Gulf of Gabès on the eastern coast of Tunisia. The first mud appears on the approach to the Nile Delta, and the shelf widens again to 70 miles off Port Said at the entrance to the Suez Canal. The shelves continue to be narrow along most of the northern shore of the Mediterranean. An exception is the broad shelf extending for 300 miles along the inner portion of the Adriatic Sea. Relatively deep water is found along much of the coast of Yugoslavia and along the Italian coast, in contrast to the gentle slopes of the Po Valley.

The northern shores of the eastern Mediterranean are very complex and, unlike the southern shores, have variable fold mountains that offered favourable sites for the development of the Mediterranean civilizations.

The north coast of Africa bordering the eastern Mediterranean is low-lying and of monotonous uniformity except for the Barqah highlands in Libya, which lie to the east of the Gulf of Sidra.

The largest islands of the eastern Mediterranean are Crete and Cyprus, both of which are mountainous.

Hydrography. The hydrological conditions in the Mediterranean are dominated by the presence of three layers of water masses: the surface layer, the intermediate layer, and the deep layer that continues down to the bottom, a separate bottom layer being absent.

The surface layer has a thickness varying between 250 and 1,000 feet. This variable thickness is determined in the western basin by the presence of a minimum temperature at its lower limit. In the eastern basin the temperature minimum is generally absent, and a layer of low temperature decrease is found instead.

The intermediate layer is accompanied by warm and saline water coming from the eastern Mediterranean and characterized by temperature and salinity maxima at 1,300 feet. This layer is situated at depths between 1,000 and 2,000 feet.

The deep layer occupies the remaining zone between the intermediate layer and the bottom. The approximate temperatures in this zone are: 55.2° F (12.9° C) at 3,000 feet and 55.6° F (13.1° C) at 8,200 feet; its salinity is 38.40 parts per thousand or slightly less. In general, the water of this layer is very homogeneous.

Currents. The Mediterranean Sea receives from the rivers that flow into it only about one-third of the amount of water that it loses by evaporation. In consequence there is a continuous inflow of surface water from the Atlantic Ocean. After passing through the Strait of Gibraltar, the main body of the incoming surface water flows eastward along the north coast of Africa. This current is the most constant component of the circulation of the Mediterranean. It is most powerful in summer, when evaporation in the Mediterranean is at a maximum. This inflow of Atlantic water loses its strength as it proceeds eastward, but it is still recognizable as a surface movement in the Sicilian channel and even off the Levant coast.

A small amount of water also enters the Mediterranean from the Black Sea as a surface current through the Bosporus, the Sea of Marmara, and the Dardanelles.

By evaporation, the Mediterranean water becomes more saline, with a corresponding increase in density. It therefore sinks, and the excess of this denser bottom water emerges into the Atlantic Ocean over the sill forming the shallow Strait of Gibraltar, as a west-going subsurface current below the inward current. The latter extends from the surface down to 230 or 260 feet. The Mediterranean has been metaphorically described as breathing, inhaling surface water from the Atlantic and exhaling deep water in a countercurrent below.

The fundamental surface circulation of the Mediterranean consists of separate counterclockwise movement of the water in each of the two basins.

Because of the complexity of the northern coastline and of the numerous islands, many small eddies and other local currents form essential parts of the general circulation. Tides, although only significant in range in the Gulf of Gabès and in the northern Adriatic, add to the complications of the currents in narrow channels such as the Strait of Messina.

Historically, large seasonal variations in the Nile discharge influenced the hydrology, productivity, and fisheries of the southeastern part of the Mediterranean. It reduced the salinity of the coastal waters and increased both the stratification and productivity of these waters. The Aswān High Dam (1970) stopped the seasonal fluctuation of the discharge of the Nile water into the Mediterranean.

The Suez Canal between the Mediterranean and Red seas passes negligible amounts of Red Sea water into the Mediterranean.

The general water budget of the Mediterranean is given in Table 1.

Table 1: Water Budget in the Mediterranean Sea

gains	cu ft/sec	cu m/sec	losses	cu ft/sec	cu m/sec
Inflow from the Atlantic	61,800,000	1,750,000	Outflow to the Atlantic	59,330,000	1,680,000
Inflow from the Black Sea	445,000	12,600	Outflow to the Black Sea	215,000	6,100
Precipitation	1,115,000	31,600	Evaporation	4,075,000	115,400
Runoff	260,000	7,300			
Total	63,620,000	1,801,500	Total	63,620,000	1,801,500

Seawater temperature. The 40° N parallel runs through the middle of the western basin, whereas the corresponding latitude of the eastern basin is 6° farther south, a fact that explains the higher surface temperature of the latter.

The highest temperature of the Mediterranean is met within the Gulf of Sidra, off the coast of Libya, where the mean temperature in August is around 88° F (31° C). This is followed by the Gulf of Iskenderun, Turkey (mean temperature about 86° F [30° C]). The lowest sea temperatures are found in the extreme north of the Adriatic, where the mean temperature in February falls down to 41.4° F (5.2° C) in the Gulf of Trieste. In this region ice occasionally forms in the depth of winter.

Salinity. Just as in all other seas and oceans, chlorides constitute more than half of the total ions present in the Mediterranean water, and the proportions of all the principal salts in the water are constant.

Assuming that in Mediterranean waters the average salinity (weight of dissolved ions per 1,000 parts; *i.e.,* grams per kilogram) is 38, usually expressed as 38 parts per thousand (3.8 percent), as compared with 35 parts per thousand for the oceans, then the composition of the major ions in the Mediterranean would be as given in Table 2.

Table 2: Major Ions Dissolved in One Kilogram of Mediterranean Seawater
(salinity = 38 parts per thousand)

	gm/kg		gm/kg
Chlorides	20.815	Magnesium	1.402
Sulfates	2.918	Calcium	0.437
Bicarbonates	0.156	Potassium	0.418
Sodium	11.636	Others*	0.218
Total			38.000

*Present in minute quantities.

Dissolved oxygen. The distribution of oxygen is in agreement with the origin of the different water masses. The surface layer down to 700 feet shows a high oxygen content all over the Mediterranean. The intermediate layer that is formed by the sinking of the surface layer in the eastern basin has a high oxygen content where it is freshly formed in this basin; but, as it moves westward, it loses some of its oxygen content, the lowest values occurring in the Algerian Basin. The transition layer between the intermediate and the deep water is characterized by an oxygen minimum.

Plant nutrients. In the Mediterranean Sea, plant nutrients such as phosphates, nitrates, and nitrites are scarce. Just as in all other seas, these nutrients show seasonal fluctuations, generally with a rise in the spring, the blooming season. The scarcity of nutrients in Mediterranean waters is attributed to several factors, the most important being that the Mediterranean receives its main bulk of water from the surface water of the Atlantic Ocean, which is itself not rich in these nutrients.

Scarcity of nutrients

In the western basin below 3,000 feet the contents of phosphates and nitrates are less than one-half of the amounts in the open ocean at the same depths.

The shortage of phosphates and nitrates in Mediterranean Sea water, which is so vital for the creation and the maintenance of the marine life cycle, results in the limited productivity of the region in terms of the life-forms it is able to support.

Climate. The flow of the air into the Mediterranean Sea takes place through gaps in the mountain ranges, except over the southern shores east of Tunisia. Strong winds funnelled through the gaps are the most important. The mistral, a cold, dry, northwesterly wind, passes through the Alps–Pyrenees gap, the strong northeasterly bora passes through the Trieste gap, and the cold easterly levanter and the westerly vendaval pass through the Strait of Gibraltar. The warm, dusty southeasterly that blows from Saharan Africa is known as the gibleh or khamsin. In the eastern Mediterranean the estian winds are steady, dry northwesterlies and northerlies; the sirocco, from Africa, is dry and hot.

The Mediterranean weather is characterized by windy, mild, wet winters and relatively calm, hot, dry summers. Spring, however, is a transitional season and is changeable. Autumn is relatively short.

The amount and distribution of rainfall in the Mediterranean cannot be generalized. Along the North African coast from Qābis (Gabès) in Tunisia to the Arab Republic of Egypt, more than 10 inches (250 millimetres) of rainfall per year is rare, whereas, on the Dalmatian coast of Yugoslavia, there are places with 100 inches.

Marine and plant life. Because the primary productivity of life-forms depends to a large extent on the concentration of the plant nutrient salts, it is, as would be expected, generally low in the Mediterranean Sea. The effective potential productivity in various regions of the Mediterranean as measured by radioactive methods using carbon-14 to determine the amount of carbon produced in milligrams of carbon per cubic metre of water (expressed as mg C/m^3) per 24 hours, varies from five to 150 mg C/m^3 per 24 hours. The lowest values are observed in the Levant, ranging from five to 30 mg C/m^3 per 24 hours, and also in the Ionian Basin, where values ranging between seven and 21 mg C/m^3 per 24 hours are observed.

The highest primary-production values in the Mediterranean Sea have been observed in springtime (March–May) off the Egyptian coast in areas under the influence of the Nile. There primary-production values as high as 700 mg C/m^3 per 24 hours have been found.

Fish stocks and fisheries. The environmental conditions of the Mediterranean already described, namely, the low concentration of phosphates and nitrates so necessary for the maintenance of marine pastures, together with the poor standing crop of bottom-dwelling organisms, provide no opportunity for the evolution of a big fisheries industry.

Although the general lack of large concentrations of fish has prevented the development of large-scale operations, the high prices of fresh fish in most Mediterranean countries have favoured the development of a large number of small-scale fisheries. Though the boats used are rather small, rarely exceeding 70 feet long, their numbers are sufficient to deplete the local stocks through overfishing. The tendency to overexploitation is strengthened by the use of trawl nets with very small mesh size that retain the smallest individuals.

The fish fauna of the Mediterranean is basically related to the fauna of the subtropical Atlantic with a large variety of species. Of the demersal (bottom-living) fishes, flounders, sole, turbot, whiting, congers, croakers, red mullet, gobies, gurnards, lizard fish, redfishes, sea bass, groupers, combers, sea breams, pandora, and jacks and cartilaginous fishes such as sharks, rays, and skates are all caught by the trawlers. Among the demersal fish, hake is one of the more important in all countries bordering the Mediterranean Sea. It is caught in deeper waters.

Rocky coasts provide a valuable harvest of lobsters, crabs, shrimp, and prawns. Oysters are more common on shallower alluvial coasts. The brackish lagoons and deltas provide eels, anchovies, and gray mullets.

Pelagic species

About half of the Mediterranean and Black Sea catches are of pelagic species (those caught in the upper layers of the sea). Sardines constitute the main catch in the western and northeastern parts of the Mediterranean. Occasionally sardines also appear in relatively small quantities in the southeastern part of the Mediterranean. Closely related fishes (*Sardinella aurita* and *S. maderensis*), however, occur in considerable quantities in the southern and southeastern region of the Mediterranean. The sprat is taken in some quantities in the most northern parts of the Mediterranean such as the northern Adriatic. It also occurs in commercial quantities in the Black Sea. Anchovy is important in most of the regions of the Mediterranean and the Black Sea.

Of the really large fish there is one of great commercial value, the bluefin tuna that moves into the Mediterranean from the Atlantic and disperses in several directions, toward the southern and eastern coasts of Spain, the coasts of the Balearic Islands, the northern coast of Morocco, and the coasts of Sardinia, Sicily, Algeria, Tunisia, and Libya. Related species of economic value are the bonitos and the mackerels. They are caught in various quantities in practically all the countries of the region.

Of the nonedible products of the Mediterranean, the most important are the corals of Naples and the sponges of the Dodecanese, the Gulf of Gabès, and the western coasts of Egypt.

Vegetation. The natural vegetations of the lands bordering the Mediterranean can be classified into six general types, one of which is the typical Mediterranean forest. The olive tree provides a satisfactory criterion for the distribution of Mediterranean climate. The holm (or holly) oak and the cork oak provide a better criterion. They tolerate low rainfall and a few months of drought. The conifers, especially the domestic pine, the maritime pine, and the Aleppo pine, are the remnants of the typical Mediterranean forests that suffered from extensive destruction by man. A characteristic feature of the landscape is provided by the macchia (or maquis, evergreen shrubs with hard, leathery leaves), which has widely replaced the conifers. Garigue in its different forms, including sage, prostrate prickly plants, and tuberous perennials, is commonly found in Greece, Spain, and Italy, especially where the soil is limey.

Another type of natural vegetation is the mixed green and deciduous forest, which occupies large areas in Greece, Albania, the north Aegean region, Spain, Portugal, the Alps, and the Apennines. It is made up of species adapted to short periods of drought and to cool winters, hence the presence of evergreens and also the deciduous oak.

A third type is the deciduous forest found in central Italy, where chestnuts occur between 2,000 and 3,000 feet. It extends to the Alps, Cévennes, and Pyrenees. The beech follows the chestnut in altitude.

Other vegetation types are the high coniferous zone, characterized by the white fir, widespread throughout the mountains of central Europe, the high pastures, and the arid zone.

Many plants have been introduced and cultivated in the Mediterranean lands by the Arabs and Spaniards. The more successful introductions include rice, cotton, oranges, sugarcane, corn (maize), tobacco, potatoes, and sisal.

History

Early Mediterranean cultures. The 2,000-mile access that the Mediterranean offers in the temperate zone to rain-bearing westerly winds; the ease of communications across its western, central, and eastern straits; and its prevailing freedom from storms in the summer months: all these made it the "inland sea" of early civilizations. The revolutionary development of a higher civilization based on agriculture evidently originated in the lands beyond the eastern Mediterranean, but that sheltered sea favoured its westward diffusion. The western Mediterranean lands followed the eastern into the full light of history in the 1st millennium BC with the expansion of Semitic Carthage in North Africa and into Spain and the establishment of Greek colonies as far west as the Gulf of Lions. By 300 BC the eastern Mediterranean basin belonged to the Greek kingdoms of Alexander the Great's successors; and in the following two centuries the expanding Roman Republic annexed the Carthaginian empire and began the absorption of the Greek world. When the Christian era opened, the entire Mediterranean basin as far south as the Sahara was for the first time politically and economically unified under the Roman Empire, a circumstance that greatly facilitated the expansion of Christianity.

By AD 400, however, the Germanic vanguard of migrations that had pushed forward from central Asia was breaking down the northern defenses of the Roman Empire. While its western half collapsed, the Byzantine Empire survived; but Levantine separatists found expression in the schism of the eastern churches and welcomed the 7th-century Muslim Arabs as liberators. The Arabs rapidly advanced through North Africa (whose economic ruin was effected by a later wave of Bedouin invaders), conquered Spain and Sicily, and harassed Italy and Provence. Although the Mediterranean basin was thus unequally partitioned between a Christian north and a Muslim

south, common economic interests reasserted themselves, notably through Jewish intermediaries. By the 11th century the Arab empire was lapsing into disorder, and the Christian conquest of Spain and Sicily began. The Crusades were ostensibly designed to win safe access to the Christian Holy Places but also offered estates to landless younger sons and commerce to the Mediterranean trading cities, notably Venice, Genoa, and Barcelona. The Crusades failed militarily after two centuries of struggle, completed the schism between Latin and Greek Christendom, announced the doom of the Byzantine Empire, and heightened the enmity between Christendom and Islām; but meanwhile western Christendom emerged from adolescence into maturity, and the widening of its cultural horizons heralded both the humanist Renaissance and the religious Reformation.

Until the 15th century, Constantinople, Barcelona, and the Italian commercial states assumed the role of intermediaries between the Orient and northwestern Europe. In the 15th century, however, the rise of the Ottoman Turks brought a second period of eclipse in the Mediterranen as well as an era of severe rivalries. Turkish conquest of the Levant, the Balkans, and North Africa was followed by a rule of oppression and exploitation, and piracy made traffic on the sea hazardous. Moreover, the discovery of the route to Asia around the Cape of Good Hope at the end of the 15th century created a safer and easier sea route connecting northwestern Europe directly with the Orient. Mediterranean lands lost their commercial function as intermediaries between Europe and Asia, and for more than two and a half centuries the Mediterranean Sea remained a backwater of world ocean trade and traffic.

The opening of the Suez Canal in 1869, together with the advent of the steamship, continued industrialization in northwestern and central Europe, and French colonization in North Africa made the Mediterranean again one of the busiest sea lanes of the world. Much of the traffic, however, passed through the sea en route between Asia and northwestern Europe. Countries in the Mediterranean basin had not been able to industrialize and had retained their agricultural and artisan economies, which limited their purchasing power and reduced their ability to trade. Nevertheless, a fierce competition ensued among ports in northwestern Europe and Mediterranean Europe for the capture of transit trade to and from the industrialized areas of western and central Europe, a rivalry that has continued to this day.

19th and 20th centuries. From the 18th century until World War II, the Mediterranean was dominated by Great Britain, which had gradually acquired control over several strategic points along the sea route so vital to its empire trade, among them Gibraltar (1704), Malta (1800), Aden (1839), Cyprus (1878), and finally Egypt (1882). Egypt, of course, became first a kingdom (1922), and then a republic (1953). After World War II Cyprus and Malta gained independence, Aden became the People's Democratic Republic of Yemen, and the status of Gibraltar was challenged seriously by Spain. The Suez Canal was closed to all traffic from 1967 to 1975 as a result of the continuing conflict between Israel and its Arab neighbours.

Britain lost its supremacy in the Mediterranean, and a new era began with U.S. entry into North Africa in 1942 during World War II. After the war it became apparent that only the United States was strong enough to oppose the Soviet Union, which made repeated attempts to expand into the Mediterranean sphere economically and politically. Thus, with the view of keeping the Mediterranean under non-Communist control and to protect Western interests in the Middle East and its oil fields, the United States took a leading role in the security of the Mediterranean and its borderlands by initiation of the Marshall Plan and the Truman Doctrine, by participation in the North Atlantic Treaty Organization, which it helped create, and by the enactment of the Eisenhower Doctrine for the Middle East. The sharpening of the Arab–Israeli conflict and the persistence of inter-Arab rivalries, most notable in the ravages of Lebanon, threatened to make the Mediterranean a backwater for world ocean shipping once again and to bring chaos and destruction to its eastern shores.

History of scientific study. Although well known to the geographers of antiquity and the Middle Ages, the Mediterranean Sea was scarcely subjected to any thorough modern study until the Danish expedition in the "Thor" undertook such a task in 1908–10. The investigations carried out covered as much of the Mediterranean Sea as possible with regard to pelagic (open-sea) animal life and its dependence on hydrographical (flow) conditions and also studied the seasonal circulation between the western and eastern basins and between the Balearic and Tyrrhenian seas.

The expedition of the "Thor"

Not until the end of World War II were major scientific surveys resumed when, in 1948, several surveys were made in the western basin. More systematic studies were carried out there (by the French) from 1957 to 1963.

The eastern basin has received less attention. Mention should be made, however, of investigations into the sources of the formation of deep water in the eastern basin (1957), studies of the vertical circulation of the Mediterranean Sea (1960, 1961), and studies of the hydrographic conditions of the bottom water during summer, using the observations of the French vessel "Calypso" (1955–60).

Cooperative investigations in the Mediterranean were initiated in 1969 by the Intergovernmental Oceanographic Commission (IOC/UNESCO), the International Commission for the Scientific Exploration of the Mediterranean Sea (ICSEM), and the General Fisheries Council for the Mediterranean (GFCM/FAO). Twenty-four nations agreed to cooperate in implementing these programs that cover physical, chemical, geophysical, biological, and fisheries research. Operations were started in 1969 by the German research vessel "Meteor." These operations covered seismic research, continuous seismic reflection measurements, geothermal measurements, radiological and petrophysical investigations, bathymetric measurements, volcanological studies, investigations of marine geology, biological investigations, studies regarding recent ore formation, and analysis for cosmic dust. (G.G.Wd./S.G./M.M.Sa/Ed.)

International cooperation

BIBLIOGRAPHY. General works include D.S. WALKER, *The Mediterranean Lands,* 3rd ed. (1965); V.M. ALAVEDRA, *Contribution à la connaissance des régimes fluviaux Méditerranéens* (1949); and PIERRE BIROT and JEAN DRESCH, *La Méditerranée et Le Moyen-Orient,* vol. 1, *La Méditerranée Occidentale,* vol. 2, *La Méditerranée Orientale et Le Moyen-Orient* (1953–56). See also M.I. NEWBIGIN, *The Mediterranean Lands* (1924), a general treatment of countries bordering the Mediterranean, with emphasis on human geography; and *Southern Europe,* 3rd ed. rev. (1949); E.C. SEMPLE, *The Geography of the Mediterranean Region* (1932); M. SORRE and J. SION, "Méditerranée," in *Géographie universelle,* vol. 7, pt. 1 (1934); and GEORG WUST, "On the Vertical Circulation of the Mediterranean Sea," *J. Geophys. Res.,* 66:3261–3271 (1961), which deals with Mediterranean currents studied from data already published.

Melbourne

Although the officially defined city is the home of fewer than 65,000 people, Melbourne is the core of an extensive metropolitan area and is the world's most southerly urban area with a population of more than 1,000,000. It is situated on the southeastern coast of Australia, at the head of Port Phillip Bay, and is the capital of the state of Victoria. In Australia its metropolitan area is second only to Sydney's in population, and there is a good-natured rivalry between the two cities, to which geography and history have bequeathed diverse characteristics. Though Melbourne's flat site has led to the regular development of a rectangular pattern of streets, the city has many beautiful parks, and the person with an eye for architectural detail and history can find much that is varied and attractive. Melbourne has a reputation for conservatism and financial soundness—attributes that have contributed to its growth and are revealed by the burgeoning skyline of the central city and the rapidly expanding eastern suburbs.

This article is divided into the following sections:

Physical and human geography 944
The landscape 944
The city site
Climate
The city layout
The people 944
Patterns of immigration
Move to the suburbs
The economy 945
Industry and trade
Transportation
Administration and social conditions 946
Government
Health and education
Cultural life 946
Arts
Recreation
History 946
Early settlement 946
Growth of the city 946
Bibliography 947

Physical and human geography

THE LANDSCAPE

The Melbourne Statistical Division, which includes all areas in close economic and social contact with the central city, covers 2,359 square miles (6,109 square kilometres). The continuous, densely settled metropolis, where a vast majority of the residents live, is less than half the size of the larger division.

The city site. Metropolitan Melbourne is situated at the northern end of Port Phillip Bay, 30 nautical miles (55 kilometres) from the bay's narrow entrance. Most of the flat terrain is less than 390 feet (120 metres) above sea level. The expansion of Melbourne from its origins at the mouth of the Yarra River to its present shape displays a strong correlation with the geology and drainage of the land. West of the original city site, basalt flows during the Tertiary Period filled the existing valleys and left flat, uniform plains. The eastern region, however, consists of undulating and dissected beds of sandstones, shales, and conglomerates laid down in the Silurian and Devonian periods. The thicker soils of the eastern region, together with its higher annual rainfall, supported a much denser cover of trees than on the basalt plains. Not surprisingly, the development of Melbourne has been mainly eastward into the broad reaches of land between Darebin Creek, the Plenty and Yarra rivers, and Koonung and Gardiners creeks. In a strikingly asymmetrical fashion, Melbourne's urban development presently lines the entire eastern shore of Port Phillip Bay, from the mouth of the Yarra River to Point Nepean, 60 miles (97 kilometres) distant; while corresponding development on the west coast of the bay extends for only 10 miles.

Asymmetrical urban growth

Climate. Melbourne's weather results from the eastward flow of high-pressure cells separated by low-pressure troughs. These patterns follow a course that passes south of the continent in summer and over northern Victoria in winter. The annual rainfall of 26 inches (657 millimetres) is fairly evenly distributed throughout the year, with October usually the wettest month and January the driest. Temperatures are moderate, only rarely falling below freezing; average daily maximum temperatures vary from 55° F (13° C) in July to 79° F (26° C) in January. Winds associated with the eastward passage of weather systems ensure that Melbourne is spared the serious air pollution of some other large cities.

The city layout. The area of original settlement in Melbourne, which today forms its financial, legal, administrative, and ecclesiastical heart, was laid out in a rectangular pattern that has not changed. The area has a frontage along the Yarra River of one mile and extends back from the river for half a mile. Within this core are the major suburban and interstate railway stations; Victoria's Houses of Parliament; the Anglican and Roman Catholic cathedrals; the Law Courts; the State Library; and many financial institutions, including the Melbourne Stock Exchange and the headquarters of major banks. Most of the city's buildings are modern, but the Town Hall, the Law Courts, and the Exhibition Building provide excellent examples of 19th-century official architecture.

The earliest suburbs—Carlton, Collingwood, Richmond, Prahran, St. Kilda, and Brighton—also have a generally rectangular pattern of streets. Row houses, often with verandas decorated with iron lacework, were common features in the suburbs close to the city's centre; and in parts of Carlton and South Melbourne some of these traditional terraces have been preserved. Today most houses are detached and are built of a single thickness of bricks rather than the weatherboards (clapboards) that were common until the 1940s.

THE PEOPLE

Patterns of immigration. The first official census of Melbourne, in 1836, numbered 177 persons, of whom 35 were females. In the 1850s the gold rush in nearby areas of Victoria resulted in the city's first major period of immigration. The immigrants came principally from other Australian colonies and Britain. By the 1920s Melbourne had become the home of more than half the residents of Victoria, and toward the end of World War II it reached a population of 1,000,000.

The second great wave of immigration came in the 1950s, when the Australian government pursued a deliberate policy of encouraging migration from Europe to provide workers for Australia's developing industries. The government provided assistance with travel costs and helped the immigrants to settle in Australia, learn English if necessary, and find employment. At first migrants were drawn mainly from the Baltic states and eastern Europe; many of these people were war refugees. Then larger numbers began coming from the United Kingdom and Ireland. Immigration agreements were signed with the Dutch, Maltese, West German, Italian, Greek, and Austrian governments. These programs sowed the seeds of Melbourne's present multicultural character. Italians, Greeks, and Yugoslavs formed the largest numbers of non-English-speaking migrants, and by 1976 a fifth of the city's population regularly spoke a language other than

Diverse ethnic population



The Exhibition Building in Melbourne.
By courtesy of Promotion Australia

English. Immigration from Southeast Asia, in particular refugees from Vietnam and Kampuchea (Cambodia), increased during the early 1980s.

Move to the suburbs. By 1966 the population of the Melbourne area had exceeded 2,000,000, and the city limits were rapidly being pushed eastward. The migrants from southern Europe were at first concentrated in the older industrial suburbs, where they found lower rents and nearby job opportunities. In those areas, too, support was available from earlier immigrants and the associations they had established. As these people prospered, however, they embraced the Australian dream of owning a home on a quarter-acre plot and began to move to the outer suburbs.

In the 1940s about 90 percent of Melbourne's residents lived within 10 miles of the central business district. By the 1980s this proportion had fallen to less than 50 percent, and the outer suburban areas had grown correspondingly. Although some western sections can be typified as working-class districts, on the whole Melbourne's suburbs represent a fairly even mixture of social classes and occupations.

THE ECONOMY

Industry and trade. That Melbourne dominates the economic life of Victoria is not surprising—the city contains the bulk of the state's population, is Victoria's financial centre and seat of government, and is at the hub of the communications network linking the state to the rest of Australia and the world.

The city's original core offers the most employment, but employment is growing at a faster rate in the outer suburbs. The central city mainly holds service industries such as banking, insurance, retailing, entertainment, public accommodation, and railway transportation. Surrounding this core is an incomplete ring of inner industrial suburbs, where the first clothing and metal factories were established in the 19th century. In the outer suburbs, particularly to the east, small manufacturing areas began to develop after World War II, when these suburbs could offer large areas of inexpensive land, few problems of traffic congestion, and an increasing population.

Melbourne's most important industries, in terms of numbers employed, are metal processing, including the manufacture of transportation equipment, and engineering. Other major industries include textile and clothing manufacture; food processing; papermaking and printing; and the manufacture of chemicals, furniture, and building materials. Melbourne is also one of Australia's leaders in the manufacture of computers.

The commercial port

The port of Melbourne occupies an area of level excavated land at the mouth of the Yarra River, southwest of the central business district. It is the nation's largest general-cargo port. Melbourne's chief imports are machinery, motor vehicles, textiles, newsprint and paper, iron and steel, and chemicals; its chief exports are wool, petroleum products, meat, fruit, wheat, and dairy products.

Since the 1960s large regional shopping centres have sprung up throughout Melbourne's outer suburbs, and the central city has lost its dominant retail function. Many residents do nearly all their shopping for clothes, appliances, furniture, and automobiles in the suburbs rather than coming into the city.

Transportation. Melbourne is well served by an integrated public system of electric trains, buses, and tramcars. In the 1970s and 1980s improvements to the road network reduced congestion to a level where commuting by private car has become a reasonable alternative to public transportation, which therefore must be financed by increasing annual deficits. Major improvements have included the opening of the Westgate Bridge over the lower Yarra River, linking the western and southeastern suburbs by a route that avoids the heart of the city, and the construction of an underground rail loop to serve the central business district. Melbourne's international and domestic airport is located on the flat basalt plain 11 miles northwest of the city's centre.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The Victorian state government has the ultimate responsibility for Melbourne's major planning decisions and for providing its principal health, educational, and transport services. Local government in the Melbourne Statistical Division is provided by more than 50 cities and shires. Councillors are elected annually in August under a system of compulsory, preferential voting; they serve in a voluntary capacity, although small sums are provided to reimburse out-of-pocket expenses. The councils pass local ordinances and control a number of services connected with building regulations, community welfare, garbage collection, and vehicle parking. Revenue for these purposes is raised by taxes levied on the valuation of property.

Health and education. Since the system of free public hospitals was started in 1846, it has grown to encompass numerous special facilities, which deal either with particular ailments or categories of patients, as well as general hospitals. There are in addition many private hospitals.

The University of Melbourne, one of the oldest in Australia, was established in 1853 (though the first students were admitted in 1855); the city's other two universities, Monash and La Trobe, date from the 1960s. All three universities are state-financed. Melbourne also has colleges of advanced education that offer degrees or diplomas in a variety of technical and academic subjects.

CULTURAL LIFE

Arts. Melbourne's already rich cultural life was greatly enhanced during the period between 1968 and 1984, when the Victorian Arts Centre was created on the south bank of the Yarra River close to the city's centre. It encompasses an art gallery, concert hall, and three theatres, among other facilities for the arts.

The arts centre

The National Gallery of Victoria was originally opened in 1861 and moved to its present site in the arts centre in 1968. It houses several outstanding collections including, most notably, Australian art ranging from the colonial period to modern times; European art, with 18th-century works particularly well represented; and decorative arts.

The Melbourne Concert Hall seats 2,600. Its patrons appreciate not only the technical brilliance of the acoustic engineering but also the hall's superb decoration in colours derived from the gemstone and mining industry, which make the hall appear to have been carved out of a hillside. Of the three theatre auditoriums, the State Theatre is used mainly for opera and ballet, the Playhouse for performances of drama, and the Studio, a more intimate theatre, for cabarets, television productions, and experimental plays. Yearly seasons of opera, ballet, and concert music include performances by international artists. The Melbourne Symphony Orchestra was formed in 1949 and has toured North America, Japan, and New Zealand as well as major Australian cities.

Recreation. Automobile license plates in Victoria carry the identification "Victoria—Garden State." Melbourne is worthy as the capital of a garden state, with 27 percent of its inner-city area consisting of public parks and reserves. These spaces were set aside in the mid-19th century, at a time when many civic leaders in other cities were concerned with commercial development rather than with the quality of life. Extensive tracts have also been allotted as parklands in the newer outer suburbs. The most famous park in Melbourne is the Royal Botanic Gardens. This area of 86 acres (35 hectares) was reserved in 1845 and today contains lakes, lawns, and thousands of named trees and shrubs. The associated Herbarium, which houses a collection of 1,500,000 pressed plant specimens, is internationally recognized and used by scholars.

Melbourne has hundreds of sports fields, tennis courts, swimming pools, and golf courses for active sports participants. Spectators find good accommodations at the Melbourne Cricket Ground, which holds 100,000 and is used for both cricket and Australian Rules football, and at the Flemington Racecourse, where the valuable Melbourne Cup race is held every November. Melbourne hosted the 1956 Summer Olympic Games. New sports facilities, including a large tennis stadium near the Melbourne Cricket Ground, have been added since the Olympics. Sailing and fishing on Port Phillip Bay and surfing on the ocean beaches are also popular, and the winter ski slopes around Mount Buller are within easy reach.

History

EARLY SETTLEMENT

Port Phillip Bay was discovered by Europeans in 1802, when captains John Murray and Matthew Flinders visited the bay within a few months of each other. This area was then part of the colony of New South Wales, and the colony's governor, Philip Gidley King, instructed the surveyor-general, Charles Grimes, to examine the shores of the bay with a view to identifying sites for future settlement. In 1803 Grimes and his party discovered the Yarra River and traveled along its lower course. Unlike some members of the party, Grimes was not enthusiastic about the Yarra River as a potential settlement. Later in the same year Captain David Collins arrived with a contingent of soldiers and convicts and settled near Sorrento, just inside the entrance to the bay on the east coast. Within a few months, however, he decided that the location was unsuitable and moved his group to Tasmania.

Exploration of Port Phillip Bay

Permanent settlement was delayed until 1835, when a pioneer settler and entrepreneur, John Batman, negotiated a treaty with the Aboriginal elders for the purchase of 500,000 acres at the head of Port Phillip Bay. The price he paid for the land was 40 blankets, 30 axes, 100 knives, 50 scissors, 30 mirrors, 200 handkerchiefs, 100 pounds of flour, and six shirts. Batman and his heirs were bound by the treaty to provide an annual "rent of tribute" of similar items.

A few days after the treaty was signed, Batman left, and two months later a party led by another pioneer, John Fawkner, settled on the banks of the Yarra River. There has been much debate about whether Batman or Fawkner should be regarded as the founder of Melbourne. Both seem to have an equal claim, but if the term is interpreted to include expansion and consolidation of the settlement, then the honour must go to Fawkner. Within four years of signing his treaty, which had been disallowed by the Governor of New South Wales, Batman died, at age 38; his financial affairs were in disarray, and prolonged litigation over his will destroyed the estate he had created. Fawkner lived to the age of 76. He died in Melbourne in 1869 after a rewarding career in which he established hotels, a newspaper, and a bookselling business, acquired large areas of land, and held a seat on the Legislative Council for 18 years.

Melbourne is distinguished from the other Australian state capitals in that it was founded unofficially, by individual enterprise. Once Batman, Fawkner, and others had established the settlement in 1835, the government in Sydney had to recognize the fact. In 1836 the first administrator of the Port Phillip District arrived, and in 1837 the new settlement was given its present name honouring the British prime minister, Lord Melbourne. Melbourne became a town in 1842 and a city in 1847, but its first main surge in growth came in the early 1850s following the discovery of gold near Bendigo and Ballarat less than 100 miles away. In three years the population of Melbourne increased fourfold to 80,000.

The gold rush

GROWTH OF THE CITY

Melbourne capitalized on its central position within Victoria and its port facilities to capture most of the region's trade. Between 1856 and 1873, railways were built to Geelong, Ballarat, Bendigo, Echuca, and Wodonga, and in 1883 a link with the New South Wales rail system was established at Albury. In 1877 the Melbourne Harbour Trust was created, and the Coode Canal was cut in the soft alluvial sediments of the lower Yarra River to provide a more direct course free from silting problems.

During the 1870s manufacturing flourished under the protection of a high tariff, and progress in most spheres continued until 1889, when a financial crisis and the collapse of many firms lowered public confidence. The following decade witnessed a severe economic depression

that began with a maritime strike and the failure of a number of banks and was sealed by seven years of drought from 1895 to 1902. In the decade before 1891 the population of Melbourne had increased by 200,000; in the following decade it rose by only 6,000.

In the early years of the 20th century, confidence gradually revived. Australia became a commonwealth, and Melbourne served as its federal capital until 1927, when Canberra was established. World Wars I and II encouraged the growth of manufacturing, and after 1945 European immigrants began to arrive in significant numbers.

After 1971 Melbourne's rate of growth slackened as immigration declined, and economic conditions worsened through the 1970s and early 1980s. Nevertheless, during this period of slower population growth a number of major changes took place. The appearance of the inner city was transformed by the replacement of old buildings with multistory office structures and hotels. The system of arterial roads was improved dramatically. Several important suburban economic areas emerged, reducing the retail and industrial importance of the city's centre. And the cultural life of the city was immeasurably enlivened by the completion of the Victorian Arts Centre.

BIBLIOGRAPHY. W.H. NEWNHAM, *Melbourne: Biography of a City* (1985), an excellent historical account; ANTHONY HARVEY (comp.), *The Melbourne Book* (1982), a good general description; J.S. DUNCAN (ed.), *Atlas of Victoria* (1982), which includes many well-illustrated references to Melbourne; and GRAEME DAVISON, *The Rise and Fall of Marvellous Melbourne* (1978), an economic history. See also AUSTRALIAN BUREAU OF STATISTICS, *Victorian Year Book* (annual), invaluable for statistics and developments in the area, with a good bibliography; and "Greater Melbourne," *International Demographics,* 5(3):1–9, 12 (March 1986), for recent demographic data on the city.

(J.R.V.P.)

Memory

That experiences influence subsequent behaviour is evidence of an obvious but nevertheless remarkable activity called remembering. Learning could not occur without the function popularly named memory. Practice results in a cumulative effect on memory leading to skillful performance on the tuba, to recitation of a poem, and even to reading and understanding these words. So-called intelligent behaviour demands memory, remembering being prerequisite to reasoning. The ability to solve any problem or even to recognize that a problem exists depends on memory. Typically, the decision to cross a street is based on remembering many earlier experiences.

Practice (or review) tends to build and maintain memory for a task or for any learned material. Over a period of no practice what has been learned tends to be forgotten; and the adaptive consequences may not seem obvious. Yet, dramatic instances of sudden forgetting (as in amnesia) can be seen to be adaptive. In this sense, the ability to forget can be interpreted to have survived through a process of natural selection in animals. Indeed, when one's memory of an emotionally painful experience leads to severe anxiety, forgetting may produce relief. Nevertheless, an evolutionary interpretation might make it difficult to understand how the commonly gradual process of forgetting survived natural selection.

In speculating about the evolution of memory, it is helpful to consider what would happen if memories failed to fade. Forgetting clearly aids orientation in time; since old memories weaken and the new tend to be vivid, clues are provided for inferring duration. Without forgetting, adaptive ability would suffer; for example, learned behaviour that might have been correct a decade ago may no longer be. Cases are recorded of people who (by ordinary standards) forgot so little that their everyday activities were full of confusion. Thus, forgetting seems to serve the survival of the individual and the species.

Another line of speculation posits a memory storage system of limited capacity that provides adaptive flexibility specifically through forgetting. In this view, continual adjustments are made between learning or memory storage (input) and forgetting (output). Indeed, there is evidence that the rate at which individuals forget is directly related to how much they have learned. Such data offer gross support of contemporary models of memory that assume an input-output balance.

Whatever its origins, forgetting has attracted considerable investigative attention. Much of this research has been aimed at discovering those factors that change the rate of forgetting. Efforts are made to study how information may be stored; that is, to discover the ways in which it may be encoded. Remembered experiences may be said to consist of encoded collections of interacting information; and interaction seems to be a prime factor in forgetting.

This article treats the phenomena and functions of normal memory; it also discusses the variety of disorders that affect the memory. The article is divided into the following sections:

The nature of memory: retention and forgetting 947
- Measuring retention 947
  - Recall
  - Recognition
  - Relearning
- Time-dependent aspects of retention; storage and retrieval 948
- Theories of forgetting 949
- Correlates of rate of forgetting 950
  - Degree of learning
  - Mnemonic systems
  - Individual differences

Abnormalities of memory 951
- Organic disorders 951
- Psychological studies of amnesia 953
- Psychogenic amnesia 954
- Paramnesia and confabulation 954
- Hypermnesia 955

The nature of memory: retention and forgetting

Psychologists of the modern era, from their earliest speculations about remembering to the formulation of most of their latest experimentally based views, commonly have assumed that the critical problems are concerned with the physiological mechanisms by which events and experiences can be retained so that they can be mentally reproduced, either in their original mode or with the assistance of signs and symbols that are regarded as equivalent to that mode. Memory is thus usually considered to function perfectly in proportion to its literal accuracy of reduplication. Investigators have generally supposed that anything that influences the behaviour of an organism endowed with a central nervous system leaves—somewhere in that system— a "trace" or group of traces. So long as these traces last they can, in theory, be restimulated and the event or experience that established them will be remembered. The experimental psychology of remembering—all modern experts claim to base their conclusions upon experimental evidence—endeavours to discover methods for identifying the necessary and sufficient conditions for the persistence and length of persistence of traces and for their restimulation.

MEASURING RETENTION

Standard sentences, prose passages, and poems have been used to control input in studies of retention; but discrete

verbal units (such as words or sets of letters) are most frequently employed. The letters usually comprise lists of consonant syllables (three consonants; *e.g.,* RQK) or so-called nonsense syllables (consonant-vowel-consonant; *e.g.,* ROK). The order in which verbal units are to be learned and to be recited may be left to the subject (free recall). A schoolchild who can recite the names of all African countries probably has learned such a free-recall task. Units also can be presented serially (in a constant order), the subject being asked to recite them in that order; reciting the alphabet in the usual way represents such serial learning.

Free recall, serial learning, paired-associate tasks

Pairs of words may be offered; in such paired-associate tasks the subject eventually is asked to produce the missing member of each pair when only one word is shown. This is akin to learning English equivalents for words from another language.

For these and similar tasks investigators commonly permit subjects enough practice trials to reach some preselected criterion or level of performance. This level effectively defines an immediate retention score against which later forgetting may be measured. Subsequent tests of retention are then made to investigate the rate at which forgetting proceeds. This rate tends to vary with the methods used, basically those of recall, recognition, or relearning.

Recall. Subjects may be asked to reproduce (recall) previously learned data in any order or in the original order in which they were learned.

In a free-recall test the instructions might be: "Yesterday you learned a list of words; please write as many of those words as you possibly can as they occur to you." For the paired-associate task the subject may be told: "Yesterday you learned some pairs of words; I will show you one word from each pair and you try to give the other." He may be paced, being limited to a few seconds to produce each word; or he may be unpaced, being given no rigidly specified limits.

If retention of any kind is to be measured over different periods (*e.g.,* an hour, a day, a week) a separate group of individuals should be used for each period. The reason is that the very act of remembering constitutes practice that keeps memory lively, tending to give misleading underestimates of the rate of forgetting if the same subjects are tested over successive intervals.

Recognition. The subject's task is simpler in tests of recognition, since reproduction or retrieval (as in recall) is not required. The subject simply is asked to remember previously presented information when it is offered to him again. For example, he may be given a list of words for study; on the subsequent test of retention these are mingled with additional words, the subject being asked to identify (recognize) the original words. Apparently the recognition test stresses ability to choose between "old" (studied) data and "new" words, although this need not mean that choices are based only on temporal discrimination (awareness of time distinctions).

In an alternative variety of recognition test, each word studied might be paired with a new one, the task being to choose the old member of each pair. Or, the test words might be presented one at a time for identification as old or new. Sometimes learning and testing are combined: a very long list of words may be presented one at a time, some being repeated; the task is to recognize the repeats.

Some recognition tests stress memory of the order of presentation. The subject learns a serial list (reciting in a prescribed order); the list then is scrambled, and he is tested on his ability to rearrange it appropriately. Order may be based on how units are arranged in space (*e.g.,* printed on a page) or on their numerical position in a series or on associative information. Thus, if a paired-associate list has been learned, the test may consist of the unmatched presentation of all units with a request to pair them properly. This sort of recognition seems to emphasize associative attributes. If some elements on the test were not presented originally, the temporal attribute also may be involved.

Relearning. The number of successive trials a subject takes to reach a specified level of proficiency may be compared with the number of trials he later needs to attain the same level. This yields a measure of retention by what is called the relearning method. The fewer trials needed to reach the original level of mastery, the better the subject seems to remember. The relearning measure sometimes is expressed as a so-called savings score. If 10 trials initially were required, and five relearning trials later produce the same level of proficiency, then five trials have been saved; the savings score is 50 percent (that is, 50 percent of the original 10 trials). The more forgetting, the lower the savings score.

Although it may seem paradoxical, relearning methods can yield both sensitive and insensitive measures of forgetting. Tasks have been devised that produce wide differences in recall but for which no differences in relearning are observed. (Some theorists attribute this to a form of heavy interference among learned data that has only momentary influence on retention.) Six months or a year after initial learning, some tests may give zero recall scores but can show savings in relearning.

Short-term and long-term memory

When relatively long retention intervals (usually hours or days) are used, the methods are said to involve long-term memory. In a sense, methods for studying short-term memory are miniaturized versions of these. A list may be as short as one item, level of proficiency is very low, and retention intervals are in seconds (or minutes at most).

For example, the subject may be shown a single nonsense syllable for a few seconds' study. Next he is given a simple task (such as counting backward) to occupy him for a half minute so that he cannot rehearse, and then is asked to recall the syllable. Forgetting is observed to occur over such short intervals, tending to be greater when length of interval increases, as in long-term memory. The same procedure can be used with a single paired-associate item or with a short list of four or five pairs. In a short-term counterpart of serial learning, a string of about eight single-digit numbers or letters is presented very rapidly (say, two per second), and the subjects are asked to recall them in the order in which they were presented. Recognition tests also can be adapted for measuring short-term retention. When only one presentation is used for learning, however, relearning measures are obviously unfeasible.

TIME-DEPENDENT ASPECTS OF RETENTION; STORAGE AND RETRIEVAL

Some workers theorize a distinct short-term memory system of sharply limited capacity that can retain information perhaps only a few seconds and a long-term system of relatively unlimited capacity and retention.

Among typical people, short-term function seems limited to about seven separate units (*e.g.,* seven random letters or unrelated common words). Thus, one may consult a telephone directory and forget the number before dialing is completed. Information seems to enter long-term storage by such processing as rehearsal and encoding, as if short-term retention is a way station between incoming information and more enduring memory.

Other theorists do not distinguish short- and long-term systems as inferred from observed differences in capacity and retention. Positing only one storage system, they attribute short-term phenomena to very low levels of learning. Those who postulate distinct systems point to the results of injury to a specific brain region (the hippocampus): (1) information stored prior to hippocampal damage seems to be retained; (2) sufferers seem incapable of new long-term storage; (3) the short-term functions appear to be unimpaired and subjects perform as well as ever in tests of immediate memory (*e.g.,* for a set of random numbers). It is as if new information no longer can be transferred from some sort of short-term system to relatively enduring storage.

Neurological evidence

Other data that bear on the controversy among theorists come from studies of people without known brain injury. When one has just seen a new list of words one at a time, the initial words in the list tend to be recalled best (primacy effects), those at the end next best (recency effects), while items from the middle are least likely to be recalled. This is quite consistently found as long as recall begins immediately following presentation of the last word. If, however, a short interval follows, during which

the subject is otherwise occupied to prevent rehearsal, the recency effect may completely disappear; words at the end are no better recalled than those in the middle. Primacy effects are essentially undisturbed, while a delay as short as perhaps 15 seconds is enough to abolish the recency phenomenon. Although some suggest that recency effects depend on a separate short-term memory system and that primacy effects are mediated by a long-term system, a single memory function also may be invoked to accommodate the findings. Nevertheless, interest is growing in multisystem theories on the grounds that they enhance appreciation of the processes involved in establishing relatively enduring memory.

Investigators concerned with physiological bases for memory seek a kind of neurochemical code with enough stability physically to produce a structural change or memory trace (engram) in the nervous system; mechanisms for decoding and retrieval also are sought. Efforts at the strict behavioral level similarly are directed toward describing encoding, decoding, and retrieval mechanisms as well as the content of the stored information.

One way to characterize a memory (or memory trace) is to identify the information it encodes. A learner may encode far more information than is apparent in the task as presented. For example, if a subject is shown three words for a few seconds and (after 30 seconds of diversion from rehearsal) is asked to recall and then another triad of words is given under the same procedure, then another, and so on, then if all triads share some common element (*e.g.*, all are animal names), poorer and poorer recall is observed on successive trials. Such findings may be explained by assuming that the learner encodes this animal category as part of his memory for each word. Initially, the common code might be expected to aid recall by sharply delimiting the word population. Successive triads, however, should tend to be encoded in increasingly similar ways, blurring their unique characteristics for the subject. An additional step provides critical supporting evidence for such an interpretation. If a final triad of vegetable names is unexpectedly presented, recall recovers dramatically. The subject tends to reproduce the vegetable names much better than he does those of the last animal triad; recall is about as efficient as it was for the first three animal names. This particular shift clearly seems to provide escape from earlier confusion or blurring, and it may be inferred that a common conceptual characteristic was encoded for each animal name.

Encoding mechanisms

Any characteristic or attribute of a word may be investigated in this way to infer whether it is incorporated in memory. When recall does not recover it would seem that the manipulated characteristic has little or no representation in memory. For example, grammatical class typically does not appear to be encoded; decrement in recall produced after a series of triads consisting of verbs tends to continue when a shift is made to three adjectives. Such an experiment does not indicate what common encoding characteristic might be responsible for the decrement, suggesting only that it is not grammatical class.

Encoding mechanisms also may be inferred from tests of recognition. For example, subjects study a long list of words, being informed of a multiple-choice memory test to follow. Each word studied is made part of a test item that includes other carefully chosen new words (distractors). Distractors are selected to represent different types of encoding the investigator suspects may have occurred in learning. If the word presented for study is chosen by the subject, little can be inferred about the nature of the encoding. Any errors, however, can be most suggestive. Thus, if the word presented for study was TABLE, the multiple-choice item might be TABLE, CHAIR, ABLE, FURNITURE, PENCIL. If CHAIR is incorrectly selected, it may be suspected that this associatively related word occurred to the subject implicitly during learning and became so well encoded that the subject later could not determine whether it or TABLE had been presented. If the wrong choice is ABLE, acoustical resemblance to TABLE may have contributed to the confusion. If FURNITURE is erroneously chosen, perhaps conceptual category was prominent in the encoding.

Since it is not related in any obvious way to TABLE, the word PENCIL may be intended as a control, unlikely to be a part of the memory for TABLE. If this is the case, subjects should be more likely to select distractor words other than PENCIL (if indeed they have been encoded along with TABLE).

Although a subject may have encoded in ways suggested by particular distractors, he still may be able to choose the correct word. Or he may have encoded in ways not represented by the distractors.

Evidence has been accumulating to suggest that a long-term memory is a collection of information or of attributes that can serve in discriminating it from other memories and can function as retrieval cues. In addition to verbal attributes, visual images may be a part of the memory; emotional responses produced at the time the memory is established may be incorporated.

"Tip-of-the-tongue" phenomenon

The common experience of having a name or word on the tip of the tongue seems related to specific perceptual attributes. In particular, people who report the tip-of-the-tongue" feeling tend to identify the word's first letter and number of syllables with an accuracy that far exceeds mere guessing. There is evidence that memories may encode information about when they were established and about how often they have been experienced. Some seem to embrace spatial information; *e.g.*, one remembers a particular news item to be on the lower right-hand side of the front page of a newspaper. Research indicates that the rate of forgetting varies for different attributes. For example, memories in which auditory attributes seem dominant tend to be more rapidly forgotten than those with minimal acoustic characteristics.

If a designated (target) memory consists of a collection of attributes, its recall or retrieval should be enhanced by any cue that indicates one of the attributes. For example, on failing to recall the term horse (included in a list he has just seen), one may be told that there was an animal name among the words. Or he may be asked if an associate term (say, barn or zebra) helps him think of a word he missed. While some additional recall has been observed with this kind of help, failures are common even with ostensibly relevant cues. Though it is possible that the cues frequently are inappropriate, nevertheless, if words were not learned (encoded or stored) with accompanying attributes, cuing of any kind should be ineffective.

THEORIES OF FORGETTING

When memory of past experience is not activated for days or months, forgetting tends to occur; and any theory of forgetting must cope with this primitive observation. Such auxiliary phenomena as differences in the rates of forgetting for different kinds of information also must be accommodated.

It has been theorized that as time passes the physiological bases of memory tend to change. With disuse, it is held that the neural engram (the memory trace in the brain) gradually decays or loses its clarity. While such a theory seems reasonable, it would, if left at this point, do little more than restate behavioral evidence of forgetting at the nervous-system level. Decay or deterioration does not seem attributable merely to the passage of time; some underlying physical process needs to be demonstrated. Until a neurochemical basis for memory can be more explicitly described, any decay theory of forgetting must await detailed development.

Interference theory; retroactive and proactive inhibition

A pre-eminent theory of forgetting at the behavioral level is anchored in the phenomena of interference; in what are called retroactive and proactive inhibition. In retroactive inhibition, new learning interferes with retention of the old; in proactive inhibition, old memories interfere with the retention of new ones. Both phenomena have great generality in studies of any kind of learning, although most research among humans has considered verbal learning.

People may, for example, learn two successive verbal lists; the next day some are asked to recall the first list and others to recall the second. Still a third (control) group learns only one list and is asked to recall it a day later. People who learn two lists almost unfailingly will recall much less than do the control group. The amount

by which controls exceed those who recall the first list is a measure of retroactive inhibition; the degree to which they are better than those who recall the second list is a measure of proactive inhibition. While retroactive inhibition usually will be observed in relearning, it is unusual to detect proactive deficit under such circumstances.

Theorists attribute the loss produced by these procedures to interference between list-learning tasks. When lists are constructed to exhibit varying differences, the degree of interference seems to be related to the amount of similarity. Thus loss in recall will be reduced when two successive lists have no identical terms. Maximum loss generally will occur when there appears to be heavy (but not complete) overlap in the memory attributes for the two lists. One may recall parts of the first list in trying to remember the second and vice versa. (This breakdown in discrimination may reflect the presence of dominant attributes that are appropriate for items in both lists.) Discrimination tends to deteriorate as the number of lists increases, retroactive and proactive inhibition increasing correspondingly, suggesting interference at the time of recall.

In retroactive inhibition, however, all of the loss need not be attributed to competition at the moment of recall. Some of the first list may be lost to memory in learning the second; this is called unlearning. If one is asked to recall from both lists combined, first-list items are less likely to be remembered than if the second list had not been learned. Learning the second list seems to act backward in time (retroactively) to destroy some memory for the first. So much effort has been devoted to studying conditions that affect unlearning that it has become a major topic in interference theory.

Retroactive and proactive effects can be quite gross quantitatively. If one learns a list one day and tries to recall it the next, learns a second list and attempts recall for it the following day, learns a third and so on, recall for each successive list tends to decline. Roughly 80 percent recall may be anticipated for the first list; this declines steeply to about 20 percent for the tenth list. Learning the earlier lists seems to act forward in time (proactively) to inhibit retention of later lists. These proactive phenomena indicate that the more one learns the more rapidly one will forget. Similar effects can be demonstrated for retroactive inhibition within just one laboratory session.

Such powerful effects have led some to theorize that all forgetting is produced by interference. Any given memory is said to be subject to interference from others established earlier or subsequently. Interference, theoretically, may occur when memories conflict through any attributes. With a limited group of attributes and an enormous number of memories, it might seem that everyday attempts to recall would be chaotic. Yet even if all of the memories shared some information, other attributes not held in common could still serve to distinguish them. For example, every memory theoretically is encoded at a different time and temporal attributes might serve to discriminate otherwise conflicting memories. Indeed, when two apparently conflicting lists are learned several days apart, proactive inhibition is markedly reduced. Assuming memories to be multiple-encoded, interference theory need not predict utter confusion in remembering.

Limits on interference

Sources of interference are most pervasive and should not be considered narrowly. For example, any memory seems to be established in specific surroundings or context, and subsequent efforts to remember tend to be less effective when the circumstances differ from the original. Alcoholics, when sober, tend to have trouble finding bottles they have hidden while intoxicated; when they drink again, the task is much easier. Some contexts also may be associated with other memories that interfere with whatever it is that one is trying to remember.

Each new memory tends to amalgamate information already in long-term storage. Encoding mechanisms invariably adapt or relate fresh data to information already present, to the point that what is coded may not be a direct representation of incoming stimuli. This is particularly apparent when input is relatively meaningless; the newly encoded memory comes to resemble those previously established (*i.e.,* it accrues meaning). For example, a nonsense word such as LAJOR might be encoded as MAJOR.

To recall any nonsense word correctly requires that an appropriate decoding rule be a part of the memory; but coding rules are subject to forgetting (interference) in the same way that any attribute is. Qualitative changes in memory may result when the information presented does not allow precise decoding; thus, when one sees a drawing of a jagged figure that resembles a star he might encode it as a star, knowing full well that it is not perfect. Subsequent decoding in recall (or recognition) thus produces only an approximation of the original jagged figure; it may well be influenced by other equally imprecise decoding rules already stored. In like fashion, somewhat incoherent sentences may become more reasonable during encoding; they tend to be reproduced in memory tests more coherently. When the learner has trouble making sense of any new stimulus (when he cannot specify encoding and decoding rules with precision) the decoded memory tends to resemble previously established memories.

Although interference has attracted wide support as an account of forgetting, it must be placed in perspective. Interpretations that emphasize distinctions between short- and long-term memory and that posit control processes for handling information are potentially more comprehensive than is interference theory. Behavioral evidence for interference eventually may be explained within such systems.

In addition, a number of deductions from interference theory have not been well supported by experiment. The focus of difficulty lies in the hypothesis that interference from established memories is a major source of proactive inhibition. The laboratory subject is asked to learn tasks with attributes that have varying degrees of conflict with memories established in daily life. Theoretically, the more conflict, the greater the proactive interference to produce forgetting. Yet a number of experiments have failed to provide much support for this prediction.

Interference theory also fails to account for some pathological forms of forgetting. Repression as observed in psychiatric practice, for example, represents almost complete, highly selective forgetting, far beyond that anticipated by interference theorists. Attempts to study repression through laboratory procedures have failed to yield systematic data that could be used to test theoretical conclusions.

CORRELATES OF RATE OF FORGETTING

Reminiscence

Although forgetting normally is expected to begin as soon as practice ceases, at times an exception (known as reminiscence) has been reported. In reminiscence, memory seems to improve without practice; retention is even better if tested after a rest period than if tested immediately after learning trials stop. Observed only over periods of a few minutes, this elusive phenomenon produces very small improvements, and forgetting follows. Scores of studies designed to elicit reminiscence have failed to do so, yielding only evidence of forgetting. Reasons for the conflicting findings have not been identified.

Degree of learning. The degree of learning is found to be directly associated with the amount of practice. In a metaphoric sense, specific memory may be said to grow stronger and stronger as practice proceeds. Even after a task can be performed or recited perfectly, continued practice (sometimes called overlearning) increases the "strength" of the memory. The rate of forgetting is slower when the degree of learning is greater. If there were one universal prescription for resisting forgetting, it would be to learn to a very high level initially; results seem even better when learning trials are not bunched together. Practice trials may be given en masse in a single session or the same number of trials may be distributed in sessions held on different days. The interrupted schedule is far superior to massed practice in that the rate of forgetting that follows distributed practice is much slower. The laboratory evidence also confirms the belief that cramming for an examination may produce acceptable performance shortly afterwards, but that such massed study results in poor long-term retention. Information learned in widely distributed practice appears less susceptible to interference; memories established under distributed schedules also are

less likely to produce proactive inhibition than are those learned in massed trials.

Mnemonic systems. The principle that new information is encoded to previously stored data has been used in an effort to aid memory function. When encoding techniques are formally applied, they are called mnemonic systems or devices. (The popular rhyme that begins "Thirty days hath September . . . " is an example.) Verbal learning can be enhanced by providing an appropriate mnemonic system (even to a bright college student who may have devised efficient systems of his own). Thus, paired associates (*e.g.,* DOG-CHAIR) will be learned more rapidly if they are included in a simple sentence (*e.g.,* The dog jumped over the chair). Imagery that can relate different words to be learned (even in a bizarre fashion) has been found beneficial. Some investigators hold that pure rote learning (in which no use is made of established memories except to directly perceive the stimuli) is rare or nonexistent. They suggest that all learning elaborates on memories already available. This could be taken to mean that the rate of forgetting would be the same whether or not a formal mnemonic system were used in learning.

Indeed, there seem to be no experimental results in which formal mnemonic instruction has resulted in forgetting more rapidly than when such special training is not given. Yet, while there often is no difference in the rate of forgetting, a number of studies indicate slower forgetting following instruction in a mnemonic system. These discrepancies may mean that some mnemonic systems are more subject to interference than are others. Perhaps the methods used fail to adequately distinguish between learning and forgetting.

Rates of learning and forgetting

Factors that influence rate of learning should be distinguished from those that affect rate of forgetting. For example, nonsense syllables are learned more slowly than are an equal number of common words; if both are studied for the same length of time, the better learned common words will be forgotten more slowly. But this does not mean that rate of forgetting *intrinsically* differs for the two tasks. Degree of learning must be held constant before it may be judged whether there are differences in rate of forgetting. Rates of forgetting can be compared only if tasks are learned to an equivalent degree. Indeed, when degree of learning is experimentally controlled, different kinds of information are forgotten at about the same rate. Nonsense syllables are *not* forgotten more rapidly than are ordinary words. In general, factors that seem to produce wide differences in rate of learning show little (if any) effect on rate of forgetting. (Despite discrepant evidence, mnemonic systems may prove an exception.)

Individual differences. Experimental findings seem to contradict the common intuition that people inherently differ in the rate at which they forget. This intuitive belief appears largely to derive from definite, wide individual differences in rate of learning; some people do learn faster than others. Thus, given the same number of trials or identical time in which to study, people will vary widely in the level of learning they achieve. Individual differences in forgetting then can be predicted efficiently, merely on the basis of how well each person has learned. This powerfully indicates that ordinary estimates of one's rate of forgetting are spurious, being obscured by uncontrolled differences in learning ability. One's talent for learning seems to swamp efforts to assess his inherent tendency to forget. Under less ordinary circumstances, however (*e.g.,* selective brain injury, stroke, neurotic amnesia), the degree of learning does seem to be almost completely irrelevant to the rate at which one forgets. An amnesia sufferer may forget his own name and still may be able to remember that Sofia is the capital of Bulgaria (see below). (B.J.U./Ed.)

Abnormalities of memory

Disorders of memory must have been known to the ancients and are mentioned in several early medical texts; but it was not until the closing decades of the 19th century that serious attempts were made to analyze them or to seek their explanation in terms of brain disturbances. Of the early attempts, the most influential was that of a French psychologist, Théodule-Armand Ribot, who, in his *Diseases of Memory* (1881, English translation 1882), endeavoured to account for memory loss as a symptom of progressive brain disease by embracing principles describing the evolution of memory function in the individual, as offered by an English neurologist, John Hughlings Jackson. Ribot wrote:

> The progressive destruction of memory follows a logical order—a law. It advances progressively from the unstable to the stable. It begins with the most recent recollections, which, being lightly impressed upon the nervous elements, rarely repeated and consequently having no permanent associations, represent organization in its feeblest form. It ends with the sensorial, instinctive memory, which, having become a permanent and integral part of the organism, represents organization in its most highly developed stage.

The statement, amounting to Ribot's "law" of regression (or progressive destruction) of memory, enjoyed a considerable vogue and is not without contemporary influence. The notion has been applied with some success to phenomena as diverse as the breakdown of memory for language in a disorder called aphasia and the gradual return of memory after brain concussion. It also helped to strengthen the belief that the neural basis of memory undergoes progressive strengthening or consolidation as a function of time. Yet students of retrograde amnesia (loss of memory for relatively old events) agree that Ribot's principle admits of many exceptions. In recovery from concussion of the brain, for example, the most recent memories are not always the first to return. It has proved difficult, moreover, to disentangle the effects of passage of time from those of rehearsal or repetition on memory.

Ribot's "law"

A Russian psychiatrist, Sergey Sergeyevich Korsakov (Korsakoff), may have been the first to recognize that amnesia need not necessarily be associated with dementia (or loss of the ability to reason), as Ribot and many others had supposed. Korsakov described severe but relatively specific amnesia for recent and current events among alcoholics who showed no obvious evidence of shortcomings in intelligence and judgment. This disturbance, now called the Korsakoff syndrome, has been reported for a variety of brain disorders aside from alcoholism and appears to result from damage in a relatively localized part of the brain.

The neurological approach may be combined with evidence of psychopathology to enrich understanding of memory function. Thus, a French neurologist, Pierre Janet, described amnesia sufferers who were apparently very similar to those observed by Korsakov but who gave no evidence of underlying brain disease. Janet also studied people who had lost memory of extensive periods in the past, also without evidence of organic disorder. He was led to regard these amnesias as hysterical, explaining them in terms of dissociation: a selective loss of access to specific memory data that seem to hold some degree of emotional significance. In his experience, reconnection of dissociated memories could as a rule be brought about by suggestion while the sufferer was under hypnosis. Freud regarded hysterical amnesia as arising from a protective activity or defense mechanism against unpleasant recollections; he came to call this sort of forgetting repression, and he later invoked it to account for the typical inability of adults to recollect their earliest years (infantile amnesia). He held that all forms of psychogenic (not demonstrably organic) amnesia eventually could resolve after prolonged sessions of talking (psychotherapy) and that hypnosis was neither essential nor necessarily in the amnesiac's best interest. Nevertheless, hypnosis (sometimes induced with the aid of drugs) has been widely used in the treatment of hysterical amnesia, particularly in time of war when only limited time is available.

Repression

ORGANIC DISORDERS

Defect of memory is one of the most frequently observed symptoms of impaired brain function. It may be transitory, as after an alcoholic bout or an epileptic seizure; or it may be enduring, as after severe head injury or in association with brain disease. When there is impaired ability to store

memories of new experiences (up to total loss of memory for recent events) the defect is termed anterograde amnesia. Retrograde loss may progressively abate or shrink if recovery begins, or it may gradually enlarge in scope, as in cases of progressive brain disease. Minor grades of memory defect are not uncommon aftereffects of severe head injury or infections such as encephalitis; typically they are shown in forgetfulness about recent events, in slow and insecure learning of new skills, and sometimes in a degree of persistent amnesia for events preceding the illness.

Transient global amnesia. Apparently first described in 1964, transient global amnesia consists of an abrupt loss of memory lasting from a few seconds to a few hours, without loss of consciousness or other evidence of impairment. The individual is virtually unable to store new experience, suffering permanent absence of memory for the period of the attack. There is also a retrograde loss that may initially extend up to years preceding the attack. This deficit shrinks rapidly in the course of recovery but leaves an enduring gap in memory that seldom exceeds the three-quarters of an hour before onset. Thus the person is left with a persisting memory gap only for what happened during the attack itself and in a short period immediately preceding. Such attacks may be recurrent, are thought to result from transient reduction in blood supply in specific brain regions, and sometimes presage a stroke.

Traumatic amnesia. On recovery of consciousness after trauma, a person who has been knocked out by a blow on the head at first typically is dazed, confused, and imperfectly aware of his whereabouts and circumstances. This so-called posttraumatic confusional state may last for an hour or so up to several days or even weeks. While in this condition, the individual appears unable to store new memories; on recovery he commonly reports total amnesia for the period of altered consciousness (posttraumatic amnesia). He also is apt to show retrograde amnesia that may extend over brief or quite long periods into the past, the duration seeming to depend on such factors as severity of injury and the sufferer's age. In the gradual course of recovery, memories are often reported to return in strict chronological sequence from the most remote to the most recent, as in Ribot's law. Yet this is by no means always the case; memories seem often to return haphazardly and to become gradually interrelated in the appropriate time sequence. The amnesia that remains seldom involves more than the events that occurred shortly before the accident though in severe cases careful inquiry may reveal some residual memory defect for experiences dating from as long as a year before the trauma. It is thought by some that, after recovery, the overall period of time for which there is no recollection may indicate the degree of severity of the head injury.

Haphazard recovery of memory

Traumatic automatism. Posttraumatic amnesia sometimes is observed after mild head injury without loss of consciousness and with no apparent change in ordinary behaviour. A football player who is dazed but not knocked out by a blow on the head, for example, may continue to play and even score a goal. But he may be going through these motions automatically and may later have no memory whatever of the part of the game that followed his injury. The phenomenon is known as traumatic automatism and seems similar to, if not identical with, transient global amnesia.

Memory defect after electroconvulsive therapy. Electroconvulsive treatments have been widely used in psychiatry, particularly for depressed people. A seizure or convulsion is induced by passing current through electrodes placed on the forehead. Each treatment is followed by a period of confusion for which the person is subsequently amnesic; at this time there is also a rapidly abating amnesia of some seconds for events that immediately preceded the shock. After a number of treatments, however, some individuals complain of more persistent memory defect, shown mainly in exaggerated forgetfulness for day-to-day events. These difficulties nearly always clear up within a few weeks after treatment ends. Experimental evidence tentatively suggests that electroshock administered to only one side of the head produces therapeutic results equal to those of the standard procedure but with significantly reduced impairment of memory.

Korsakoff's syndrome. First described in cases of chronic alcoholism, Korsakoff's psychosis, or syndrome, occurs in a wide variety of toxic and infectious brain illnesses, as well as in association with such nutritional disorders as deficiency of the B vitamins. The syndrome also has been observed among people with cerebral tumours, especially those involving the third ventricle (one of the fluid-filled cavities in the brain). The main psychological feature is gross defect in recent memory, sometimes so severe as to produce "moment-to-moment" consciousness; such people can store new information only for a few seconds and report no continuity between one experience and the next. They seem incapable of learning, even after many trials or repetitions. Although cases of such severity are relatively rare, the ability to store experience only briefly is quite characteristic of Korsakoff's syndrome.

In addition, sufferers almost always show evidence of retrograde amnesia that can span as little as a few weeks past to as much as 15 or 20 years before onset of the disorder. These extensive retrograde amnesias are seldom total or uniform, and "islands" of memory often can be found by persistent interrogation. The person's memory function depends heavily on circumstances; for example, a man with Korsakoff's syndrome who recognizes his wife instantly when she visits may in her absence vehemently deny that he is married. Commonly, there is disorientation in place and time; the individual often underestimates his own age, sometimes grossly. Some sufferers characteristically confabulate; *i.e.,* they remember experiences they never personally had or they falsely localize their memories in time. Not uncommonly, these people deny being ill or that their memory is in any way affected. Otherwise, they can exhibit good intelligence and, apart perhaps from some lack of spontaneity, may show little or no personality change.

Extensive retrograde amnesia

While Korsakoff's syndrome is commonly encountered as a transitory sign of brain disorder, it can be chronic, remaining effectively unimproved over many years. Even with improvement, however, an appreciable weakness in recent memory, particularly in regard to sequence in time, is quite apparent.

Persistent defect after encephalitis. Attention repeatedly has been drawn to severe and persistent memory defect following attacks of a form of brain inflammation called acute inclusion body encephalitis. The individual's behaviour closely resembles that of Korsakoff's syndrome except that his insight into the memory disorder is usually good and confabulation is infrequent or absent. Indeed, the memory disorder is sometimes so limited and specific as to raise the possibility of a psychogenic (*i.e.,* hysterical) amnesia. In cases of this kind there may be little or no impairment of intelligence or judgment.

Defect following brain surgery. Surgical operations on the sides of the brain (the temporal lobes) to remove tissues that produce symptoms of epilepsy are routine. While good results are often achieved, a degree of memory defect ensues. Operations on the dominant (usually left) temporal lobe tend to hamper one's ability to learn verbal information by hearing or reading. Usually observable even before surgery, the defect tends to be more marked after operation and has been reported to persist for up to three years before eventual recovery. Operations on one temporal lobe when there is unsuspected damage to its fellow on the other side of the brain (or on both lobes, in surgery very rarely undertaken) produce severe and persistent general memory defect, altogether comparable to postencephalitic amnesia. There is gross defect in recent memory and in learning (except perhaps in motor learning), with retrograde amnesia that initially may involve several years of the person's past. Intelligence otherwise appears to be well preserved; the individual shows insight into his memory difficulty, and seldom, if ever, confabulates.

Diffuse brain diseases. Some memory failure is almost universal during old age, particularly in forgetfulness for names and in the reduced ability to learn. Many people of advanced age, nevertheless, show adequate memory function if they suffer no brain disease. Impairment of memory is a characteristic early sign of senility, as well as of hardening of the brain arteries (cerebral arterioscle-

Senility

rosis) at any age, with exaggerated forgetfulness for recent events and progressive failure in memory for experiences that preceded the disorder. As arteriosclerotic brain disease progresses, amnesia tends to extend further into the past, embracing personal experience and general or common information. When the symptoms are almost those of Korsakoff's syndrome, the disturbance is called presbyophrenia. In most cases the amnesia is complicated by failure in judgment and changes in character. It has been suggested that severe memory defect in an elderly person carries a poor prognosis, being related to such factors as a shortened survival time and an increased death rate.

A Swiss psychiatrist, Eugen Bleuler, held that amnesia results only from a diffuse disorder of the outer layers (cortex) of the brain and suggested that memory depends on the integrity of the cortex as a whole. Indeed, the removal of brain tissue from rats and monkeys in experimental studies has indicated that retention of complex habits by the animals depends on the total amount of cortex that remains. It was claimed that the degree to which memory is lost depends not on where the brain is injured but on the extent of the damage. (This is the "law" of mass action, which asserts that the brain functions in a unitary manner; *i.e.*, as a whole.) While the extent of diffuse brain damage is roughly related to the severity of memory defect, the principle of mass action is manifestly inadequate. Whatever its physical basis, memory seems to depend on the integrity of relatively limited parts of the brain, rather than on that organ (or even the cortex) as a whole.

Severe and highly specific amnesic symptoms principally stem from damage to such brain structures as the mammillary bodies, circumscribed parts of the thalamus, and of the temporal lobe (*e.g.*, the hippocampus). While the ability to store new experience (and perhaps to retrieve well-established memories) appears to depend on a distinct neural system involving the temporal cortex and limited parts of the thalamus and hypothalamus, understanding of the neuroanatomy of memory remains sketchy enough to generate major differences of opinion. French and German workers tend to stress the role of the mammillary bodies, while U.S. investigators tend to implicate the thalamus. It has been pointed out that circumscribed damage to the mammillary bodies is not invariably associated with memory defect; cases of amnesia evidently occur in which these structures are spared. Nevertheless, implication of the mammillary bodies in a large number of verified cases of Korsakoff's syndrome seems incontrovertible. Injury to other neural tissues (*e.g.*, the so-called fornix bundle deep within the brain) that anatomically might be expected to produce severe memory disorder rarely does so. While evidence for amnesia as a sign of localized brain damage is impressive, much remains to be understood about the physical system that sustains memory.

Brain structures in amnesia

PSYCHOLOGICAL STUDIES OF AMNESIA

Short-term memory. The so-called short-term memory is typically intact among amnesia sufferers. Such victims usually can repeat a short phrase or a series of words or numbers from immediate memory as adequately as anyone of comparable age and intelligence. Such an amnesic person can retain the gist of a question or request long enough to respond appropriately, unless, of course, there is enough delay in performance or attention is diverted. Evidently the ability to register information is intact, if this means availability of data in short-term memory. Thus, experimental psychologists who favour a sharp distinction between short-term and long-term storage systems contend that the primary deficit in amnesia is an inability to transfer information from short-term to long-term storage.

Associative learning. It has been argued that the basic deficit in the amnesic state is a loss of learning ability. In a series of experiments with amnesic patients, using, for the most part, verbal material, the subjects evidenced failure to link new with old associations, rapid fading of new associations, and great difficulty in reproducing whatever associations might have been formed. These findings have been amply confirmed. In one view, however, the weakness resides less in the failure to establish new associations than in their rapid decay (*i.e.*, accelerated forgetting). On the other hand, it has often been noticed that if a Korsakoff patient can once succeed in learning an item, he may be able to reproduce it correctly after an appreciable interval of time. Further experiments, using a variety of techniques for assessing learning and retention, have suggested that retrieval rather than learning is at fault.

Motor skill. It has been noted that the fact that the acquisition of manual skill in Korsakoff patients is less impaired than either verbal learning or the solution of puzzles or mazes. This is confirmed in the observation that a severely amnesic patient who had undergone an extensive operation on the temporal lobes could perform rotary-pursuit and tracking tasks at a level not greatly inferior to that of healthy subjects. A second case of the same kind has been described, in which memory for motor tasks such as maze learning or the rendering of new compositions on the piano is said to have been completely preserved. These observations suggest that the acquisition of motor skill may remain relatively unaffected by lesions that give rise to a severe defect of general memory. What is commonly called global or generalized memory defect may, therefore, become increasingly subject to fractionation.

Generalized versus specific memory loss

Residual learning capacity. Korsakov himself pointed out that a patient who consistently denies having seen his doctor before does not necessarily react to him on each successive encounter as a total stranger. It thus appears that, despite gross amnesia, some learning, perhaps implicit, can still take place. This view has gained much support from clinical and experimental studies. About 1900 it was reported that even severely affected Korsakoff patients show appreciable savings in relearning verbal material after an interval of several hours or days, thus indicating minimal retention. Some Korsakoff patients, in spite of gross amnesia, eventually learn their way about the hospital. Again, some patients who disown any knowledge of their whereabouts may nevertheless give the correct name of the hospital, when asked to guess or to select it from a list containing the names of several hospitals. Thus, while learning capacity is seldom, if ever, wholly destroyed, there is failure to integrate new knowledge within the total personality. It is apparently a lack of mental cohesion that lies at the basis of Korsakoff's psychosis.

Forgetting. While some clinicians have attributed memory defect largely to defective registration of experience (*i.e.*, failure to form memory traces), the widely accepted view is that it results primarily from a greatly increased rapidity of forgetting (*i.e.*, rapid decay of memory traces). This view has also been held by the great majority of experimental psychologists who have worked with amnesic people. The consensus is that amnesia sufferers characteristically lose much of the memory they once had. This conclusion finds support in the very rapid extinction of conditioned eyeblink responses to a buzzer. It is notable that, in Korsakoff states, forgetting appears to be due to the passage of time (oblivescence) rather than to retroactive inhibition or some kindred interference effect.

Time disorders. Estimation of time is typically poor in amnesic states. The individual is prone to underestimate grossly the time in which he has been engaged on any particular activity. Conversely, he may equally grossly overestimate the time that has elapsed since a particular event (*e.g.*, the visit of a relative) of which he has preserved some recollection. Indeed, amnesic patients exhibit a remarkable want of coherence in their thought processes, suggesting that a lack of temporal synthesis underlies, and may indeed in large part explain, the defect of memory. Yet although difficulties in dating particular past events and in building a coherent time framework are characteristic of amnesic states and may persist after otherwise good recovery, an explanation couched wholly in terms of time disturbance is scarcely convincing.

Retrograde amnesia. Since retrograde amnesia relates to memory for events that took place when brain function was unimpaired, it clearly cannot be ascribed to failure of registration—with the exception, perhaps, of the very brief permanent amnesias following electroconvulsive shock or head injury. Retrograde amnesia otherwise would appear to be wholly due to a failure of retrieval, though this failure is evidently selective. That recent memories are

Consolidation of the memory trace

generally harder to evoke than those more remote is usually explained on the basis of consolidation; *i.e.,* progressive strengthening of memory traces with the passage of time. Yet, recency is not the only factor, and in some cases memory for a relatively recent event may still be preserved while that for one more remote is inaccessible. Much depends, too, on the method used to test retrieval; *e.g.,* recognition may succeed when voluntary recall entirely fails. By and large, the availability of information in memory would seem to depend to a considerable extent on its relation to the person's current interests and preoccupations. When these are severely curtailed by an amnesic state, the links connecting present and past are severed, with a consequent failure of reproduction.

PSYCHOGENIC AMNESIA

Some forms of amnesia appear to be quite different from those associated with detectable injury or disease of the brain. These comprise, first, amnesias that can be induced in apparently normal individuals by means of suggestion under hypnosis; and secondly, amnesias that arise spontaneously in reaction to acute conflict or stress, and which are commonly called hysterical. Such amnesias are reversible and have been explained wholly in psychological terms. Nevertheless, organic factors are not infrequently involved to some extent, and the distinction between organic and psychogenic amnesia may turn out to be far less absolute than has been supposed.

Hypnotic amnesia. Memory of a hypnotic trance is often vague and fragmentary, as in awakening from an ordinary dream. This may be due in part to defect of registration during the period of altered consciousness. At the same time, very much more complete posthypnotic amnesia can be induced if an individual is told that, when he awakens, he will remember nothing of what went on during the period of hypnosis. This is clearly a psychogenic phenomenon; memory is fully regained if the patient is rehypnotized and an appropriate counter-suggestion given. It may also be regained if the person is persistently interrogated in the waking state, again suggesting that the amnesia is apparent rather than real. This observation led Freud to seek access to ostensibly forgotten (repressed) memories in his patients without the use of hypnosis.

Hysterical amnesia. Hysterical amnesia is of two main types. One involves the failure to recall particular past events or those falling within a particular period of the patient's life. This is essentially retrograde amnesia but it does not appear to depend upon an actual brain disorder, past or present. In the second type there is failure to register—and, accordingly, later to recollect—current events in the patient's ongoing life. This is essentially anterograde amnesia and, as an ostensibly psychogenic phenomenon, would appear to be rather rare and almost always encountered in cases in which there has been a preexisting amnesia of organic origin. Rarely, amnesia appears to cover the patient's entire life, extending even to his own identity and all particulars of his whereabouts and circumstances. Although most dramatic, such cases are extremely rare and seldom wholly convincing. They usually clear up with relative rapidity, with or without psychotherapy.

Hysterical amnesia differs from organic amnesia in important respects. As a rule it is sharply bounded, relating only to particular memories, or groups of memories, often of direct or indirect emotional significance. It is also usually motivated in that it can be understood in terms of the patient's needs or conflicts; *e.g.,* the need to seek financial compensation after a road accident causing a mild head injury or to escape the memory of an exceptionally distressing or frightening event. Hysterical amnesia also may extend to basic school knowledge, such as spelling or arithmetic, which is never seen in organic amnesia unless there is concomitant aphasia or a very advanced state of dementia. A most distinctive feature of hysterical amnesia is that it can almost always be relieved by such procedures as hypnosis. Although distinguishing organic from psychogenic amnesia is not always easy, it can usually be achieved on the basis of such criteria, especially when there is no reason to suspect actual brain damage.

Relief of amnesia through hypnosis

Legal implications. The differentiation of organic from functional amnesia not uncommonly assumes legal importance, as in cases in which compensation is sought for disability held to be due to industrial or road accidents causing head injuries. If there is a complaint of defective memory, it is legally important to ascertain what part of it can be ascribed to the aftereffects of the head injury and what part of it to subsequent psychogenic elaboration. Similar issues may also arise on occasion in criminal cases, as in a trial in England (1959) in which it was contended that the accused man had a total amnesia for the circumstances of his alleged offense—the murder of a police officer—and should therefore be regarded as unfit to plead. After much discussion as to whether the amnesia was organic, hysterical, or feigned, the jury found it not to be genuine and the trial proceeded to conviction.

Mixed amnesic states. Students of amnesia have been increasingly impressed by the frequency with which psychogenic factors appear to reinforce, prolong, or otherwise complicate an organic memory defect. Hysterical reactions appear to be far from uncommon in brain-damaged patients: conversely, there is little or nothing in the pathology of hysterical amnesia that has not been observed in the organic syndrome. One case reported in the German literature in 1930 aroused great controversy. A young man developed severe and persistent amnesia following accidental carbon monoxide poisoning. His consciousness was virtually restricted to a second or two and no lasting memory traces could apparently be formed. While the original defect of memory may have been largely, if not wholly, organic, it was sustained thereafter on a hysterical basis. Conversely, a case has been reported in which the diagnosis, originally hysterical amnesia, had to be altered in light of the discovery that the patient had suffered from progressive brain disease. In such cases, organic and psychogenic factors appear to interact to produce complex and atypical symptoms.

Fugue states. The fugue is a condition in which the individual wanders away from his home or place of work for periods of hours, days, or even weeks. One celebrated case was that of the Rev. Ansell Bourne, described by the U.S. psychologist William James. This clergyman wandered away from home for two months and acquired a new identity. On his return, he was found to have no memory of the period of absence, though it was eventually restored under hypnosis. In not all cases, however, is the basis of the fugue so manifestly psychogenic. Indeed, close observation in some instances may reveal minor alterations in consciousness and behaviour that suggest an organic basis, probably epileptic. According to one view, pathological wandering with subsequent amnesia is due to a constellation of factors, among which are a tendency toward periodic depression, history of a broken home in childhood, and predisposition to states of altered consciousness, even in the absence of organic brain lesion. Psychoanalysts, on the other hand, see in the fugue a symbolic escape from severe emotional conflict.

Possible basis in epilepsy

PARAMNESIA AND CONFABULATION

The term paramnesia was introduced by a German psychiatrist, Emil Kraepelin, in 1886 to denote errors of memory. He distinguished three main varieties; one he called simple memory deceptions, as when one remembers as genuine those events imagined or hallucinated in fantasy or dream. This is not uncommon among confused and amnesic people and also occurs in paranoid states. Kraepelin also wrote of associative memory deceptions, as when a person meeting someone for the first time claims to have seen him on previous occasions. This has been renamed reduplicative paramnesia or simply reduplication. Lastly there was identifying paramnesia, in which a novel situation is experienced as duplicating an earlier situation in every detail; this is now known as déjà vu or paramnesia *tout court.* The term confabulation denotes the production of false recollections generally.

Déjà vu. The déjà vu experience has aroused considerable interest and is occasionally felt by most people, especially in youth or when they are fatigued. It has also found its way into literature, having been well described by, among other creative writers, Shelley, Dickens,

Hawthorne, Tolstoy, and Proust. The curious sense of extreme familiarity may be limited to a single sensory system, such as the sense of hearing, but as a rule it is generalized, affecting all aspects of experience including the subject's own actions. As a rule, it passes off within a few seconds or minutes, though its repercussions may persist for some time. For some epileptics, however, déjà vu may continue for hours or even days and can provide a fertile subsoil for delusional elaboration.

In view of its occurrence among organically healthy individuals, déjà vu commonly has been regarded as psychogenic and as having its origin in some partly forgotten memory, fantasy, or dream. This explanation has appealed strongly to psychoanalysts; it also gains support from the finding that an experience very similar to déjà vu can be induced in normal people by hypnosis. If a picture is presented to a hypnotized person with the instruction to forget it and then is shown with other pictures when he is awake, the subject may report an intense feeling of familiarity that he is at a loss to justify. The déjà vu phenomenon also is attributable to minor neurophysiological abnormality; it is frequent in epilepsy. Indeed, déjà vu is accepted as a definite sign of epileptic activity originating in the temporal lobe of the brain and may occur as part of the seizure activity or frequently between convulsions. It seems to be more frequent in cases in which the disorder is in the right temporal lobe and has on occasion been evoked by electrical stimulation of the exposed brain during surgery. Some have been tempted to ascribe it to a dysrhythmic electrical discharge in some region of the temporal lobe that is closely associated with memory function.

Hypnotic induction of déjà vu

Reduplicative paramnesia. Reduplication is observed mainly among acutely confused or severely amnesic people; for example, a patient may say that he has been in one or more hospitals that are very similar to his present location and that all bear the same name. The effect also can be induced by showing the person an object such as a picture and by testing him for recognition of the same picture a few minutes later. He is apt to say that he has seen a similar picture but definitely not the one now being shown. This effect appears to depend on loss of a sense of familiarity and on failure to treat a single object seen on a number of occasions as one and the same. It has been reported that reduplication of this kind is typically associated with confabulation, speech disorder (paraphasia), disorientation, and denial of illness.

Confabulation. Spurious memories or fabrications are very common in psychiatric disorders and may take on an expansive and grandiose character. They may also embody obvious elements from fantasy and dream. At a more realistic level, the production of false memories (confabulation) is best studied among sufferers of Korsakoff's syndrome, for whom consciousness and reasoning remain clear. When asked what he did on the previous day, such a person may give a detailed account of a typical day in his life several months or years earlier. Evidently his retrograde amnesia and his disorientation in time provide fertile soil for false reminiscence. When the confabulation embodies dramatic, fanciful elements, it is the exception rather than the rule.

Confabulation once was regarded as one's reaction to the social embarrassment produced by a memory defect; *i.e.,* as an attempt to fill memory gaps plausibly. Despite this possibility, many severely amnesic patients confabulate little, if at all; and there appears to be no relation between the severity of amnesia and frequency of confabulation. In consequence, individual differences in preamnesic personality have been stressed, particularly in regard to suggestibility. While many patients who confabulate are obviously highly suggestible, precise tests of suggestibility have not been used in most clinical evaluations. It also has been claimed that the superficially sociable, but basically secretive, individual is particularly prone to confabulate. The most critical factor appears to be the sufferer's degree of insight into his disorder; it has been observed that the amnesia sufferer who most strongly denies any lapse in memory is most prone to confabulate. By contrast, it also has been claimed that in chronic Korsakoff states the individual's insight into his condition is no guarantee of freedom from confabulation.

Inventive memory

While confabulation is pathological by definition, all people include an inventive, imaginative (and therefore spurious) element in their remembering. Indeed, it seems valid to say that all remembering depends heavily on reconstruction rather than on mere reproduction alone. Among amnesiacs, reconstruction is especially drastic, inventive, and error prone, particularly in regard to chronological sequence. The difference, therefore, between normal and grossly amnesic confabulation may well be one of degree rather than kind.

HYPERMNESIA

Enhancement of memory function (hypermnesia) under hypnosis and in some pathological states was frequently described by 19th-century medical writers; for example, cases were recorded of delirious people who would speak fluently in a language they had not had occasion to use for up to 50 or more years and apparently had forgotten. It was then categorically claimed that anyone under hypnosis would recollect events with invariably greater efficiency than in the waking state. It is true that experience inaccessible to ordinary recall sometimes can be recollected under hypnosis; some have attributed this effect to release from emotional inhibition. Nevertheless, evidence indicates that previously memorized material (*e.g.,* poetry) in many cases is reproduced no better under hypnosis than in the waking state.

Memory prodigies. Few individuals who exhibit exceptional memory have been studied extensively. The case of a Russian mnemonist (memory artist), "S," was studied over a period of 30 years, and his story has been delightfully written by a Soviet psychologist (see *Bibliography*). This man's exceptional mnemonic ability seemed largely to depend on an outstandingly vivid, detailed, and persistent visual memory, almost certainly eidetic ("photographic") in nature. "S" also reported an unusual degree of synesthesia, though whether this helped or hindered his feats of memory is not clear. (A person shows signs of synesthesia when he reports that stimulation through one sense leads to experiences in another sense; for example, such a person may say that he sees vivid flashes of colour when he hears music.) Although "S's" highly developed power of concrete visualization made possible feats of memory far beyond the ordinary, he exhibited weakness in abstract thinking.

Photographic memory

Exceptional memory capacity is occasionally observed among mathematicians and others with exceptional talent for lightning calculation. A mathematics professor at the University of Edinburgh, for example, was reported to be capable of remarkable feats of long-term memory for personal experiences, music, and verbal material in either English or Latin. This talented mathematician has been said to recall with complete accuracy a list of 25 unrelated words after only a brief effort to memorize, and to recite the value of pi (an endless number) to a thousand places or more. Likewise, some composers and musicians appear to possess exceptional auditory memory, though no systematic study of their attainments appears to have been made. The anatomical or physiological basis of hypermnesia remains most incompletely understood.

(O.L.Z.)

BIBLIOGRAPHY

General: J.A. ADAMS, *Human Memory* (1967), provides a well-considered account of memory as viewed from laboratory findings. A somewhat more abbreviated account is J. JUNG, *Verbal Learning* (1968). Two versions of formal models of memory are given in K.W. and J.T. SPENCE (eds.), *The Psychology of Learning and Motivation,* 3 vol. (1967–69), one by G. BOWER, "A Multicomponent Theory of the Memory Trace," 1:229–325 (1967); the other by R.C. ATKINSON and R.M. SHIFFRIN, "Human Memory: A Proposed System and its Control Processes," 2:89–195 (1968). The most comprehensive account of interference in long-term memory is provided by G. KEPPEL, "Retroactive and Proactive Inhibition," in *Verbal Behavior and General Behavior Theory,*ed. by T.R. DIXON and D.L. HORTON, pp. 172–183 (1968). A parallel analysis for short-term memory is given by L. POSTMAN, "Short-term Memory and Incidental Learning," in *Categories of Human Learning,*

ed. by A.W. MELTON, pp. 145–201 (1964). B.J. UNDERWOOD, "Attributes of Memory," *Psychol. Rev.*, 76:559–573 (1969), summarizes evidence that memory can be described as a collection of information.

Abnormalities of memory: A.R. LURIA, *The Mind of a Mnemonist* (1968), a fascinating account of a "memory prodigy" studied over many years by an outstanding Soviet psychologist; B. MILNER (ed.), "Disorders of Memory After Brain Lesions in Man," *Neuropsychologia*, 6:175–291 (1965), a symposium on memory disorders with major emphasis on psychological aspects; T. RIBOT, *Les maladies de la mémoire* (1881; Eng. trans., *Diseases of Memory*, 3rd ed., 1885); the classical text on disorders of memory. G.A. TALLAND, *Deranged Memory: A Psychonomic Study of the Amnesic Syndrome* (1965), a thorough historical, clinical, and experimental study of memory defect associated with chronic alcoholism, and *Disorders of Memory and Learning* (1968), a popular survey of memory and some of its disorders; G.A. TALLAND and N.C. WAUGH (eds.), *The Pathology of Memory* (1969), the proceedings of an international symposium on the topic; C.W.M. WHITTY and O.L. ZANGWILL (eds.), *Amnesia* (1966), a review of amnesia considered from the neurological point of view.

Mental Disorders and Their Treatment

Mental disorders, in particular their consequences and their treatment, are of more concern and receive more attention now than in the past. Mental disorders have become a more prominent subject of attention for several reasons. They have always been common, but with the eradication or successful treatment of many of the serious physical illnesses that formerly afflicted humans, mental illness has become a more noticeable cause of suffering and accounts for a higher proportion of those disabled by disease. Moreover, the public has come to expect the medical profession to help it obtain an improved quality of life in its mental as well as physical functioning. And indeed, there has been a proliferation of both pharmacological and psychotherapeutic treatments in psychiatry in this regard, many of which have proved effective. The transfer of many psychiatric patients, some still showing conspicuous symptoms, from mental hospitals into the community has also increased the public's awareness of the importance and prevalence of mental illness.

No simple definition of mental disorder is universally satisfactory. This is partly because mental states or behaviour that are viewed as abnormal or pathological in one culture may be regarded as normal or acceptable in another, and in any case there is no line clearly demarcating healthy from pathological mental functioning.

A narrow definition of mental illness would insist upon the presence of organic disease of the brain, either structural or biochemical; this condition does not pertain, as far as is known, to the majority of mental disorders. An overly broad definition would define mental illness as simply being the lack or absence of mental health; *i.e.*, that condition of mental well-being, balance, and resilience in which the individual can successfully work and function and in which he can both withstand and learn to cope with the conflicts and stresses encountered in life. A more generally useful definition than either of the above is that a mental disorder is an illness with significant psychological or behavioral manifestations that occurs in an individual and that is associated with either a painful or distressing symptom and/or with impairment in one or more important areas of functioning. The mental disorder may be due to either a psychological, social, biochemical, or genetic dysfunction or disturbance in the individual.

A mental illness can affect every aspect of a person's life, including thinking, feeling, mood, and outlook and such areas of external activity as family and marital life, sexual activity, work, recreation, and management of material affairs. Most mental disorders negatively affect how an individual feels about himself and impair his capacity for conducting mutually rewarding relationships.

Psychopathology is the systematic study of the significant causes, processes, and symptomatic manifestations of mental disorders. The meticulous study, observation, and enquiry that characterize the discipline of psychopathology are in turn the basis for the practice of psychiatry; *i.e.*, the science and practice of treating mental disorders, as well as dealing with their diagnosis and prevention. Psychiatry and its related disciplines in turn embrace a wide spectrum of techniques and approaches for treating mental illnesses. These include the use of drugs to correct biochemical imbalances in the brain or otherwise to relieve depression, anxiety, and other painful emotional states. Another important group of treatments are the psychotherapies, which seek to treat mental disorders by psychological means and which involve verbal communication between the patient and a trained person in the context of an interpersonal relationship between them. An important variant of this latter mode of treatment is behavioral therapy, which concentrates on changing or modifying observable pathological behaviours by the use of conditioning and other experimentally derived principles of learning.

This article treats the types, causes, and treatment of mental disorders. Neurological diseases with behavioral manifestations are treated in NERVES AND NERVOUS SYSTEMS. Alcoholism and other substance use disorders are discussed in ALCOHOL AND DRUG CONSUMPTION. Disorders of sexual functioning and behaviour are treated in SEX AND SEXUALITY. Mental retardation is treated in INTELLIGENCE, THEORIES AND DISTRIBUTION OF. Tests used to evaluate mental health and functioning are discussed in PSYCHOLOGICAL TESTS AND MEASUREMENT. The various theories of personality structure and dynamics are treated in PERSONALITY, while human emotion and motivation are discussed in EMOTION AND MOTIVATION, HUMAN.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, especially section 436.

The article is divided into the following sections:

Types and causes of mental disorders 957
  Classification and epidemiology 957
    Classification
    Epidemiology
  Theories of causation 958
    Organic and hereditary etiologies
    Psychodynamic etiologies
    Behavioral etiology
    Other etiologies
  Major diagnostic categories 960
    Organic mental disorders
    Substance use disorders
    Schizophrenia
    Paranoid disorders
    Affective disorders
    Anxiety disorders
    Obsessive-compulsive disorder or neurosis
    Posttraumatic stress disorder
    Somatoform disorders
    Dissociative disorders or hysterical neurosis, dissociative type
    Personality disorders
    Psychosexual disorders
    Disorders usually first evident in infancy, childhood, or adolescence
    Other mental disorders
Treatment of mental disorders 967
  Historical overview 967
    Early history
    The mental hospital era

The biological movement
Development of psychotherapy
Development of physical and pharmacological treatment
Deinstitutionalization
Development of behaviour therapy
The mental health profession in the late 20th century
Physiological treatments 969
Pharmacological treatments
Electroconvulsive treatment
Psychosurgery
The psychotherapies 971
Dynamic psychotherapies
Behavioral psychotherapy
Other therapies
Bibliography 974

Types and causes of mental disorders

CLASSIFICATION AND EPIDEMIOLOGY

Psychiatric classification attempts to bring order to the enormous diversity of mental symptoms, syndromes, and illnesses encountered in clinical practice. Epidemiology is the measurement of the prevalence, or occurrence, of these psychiatric disorders in different human populations.

Classification. Diagnosis is the process of identifying an illness by studying its signs and symptoms and by considering the patient's history. Much of this information is gathered by the psychiatrist during his initial interviews with the patient, who describes his main complaints and symptoms and any past ones and briefly gives his personal history and current situation. The psychiatrist may administer any of several psychological tests to the patient and may supplement these with a physical and a neurological examination. These data, along with the psychiatrist's own observations of the patient and of the patient's interaction with him, form the basis for a preliminary diagnostic assessment. For the psychiatrist, diagnosis involves finding the most prominent or significant symptoms, upon which the patient's disorder can be assigned to a category as a first stage toward rational treatment. This is as essential in psychiatry as in the rest of medicine.

Classificatory systems in psychiatry aim to distinguish groups of patients who share the same or related clinical symptoms in order to provide an appropriate therapy and accurately predict the prospects of recovery for any individual member of that group. Thus, the diagnosis of, for example, depressive illness having been made, it becomes logical to consider antidepressant drugs when preparing a course of treatment.

The diagnostic terms of psychiatry have been introduced at various stages of the discipline's development and from very different theoretical standpoints. Sometimes two words with quite different derivations have come to mean almost the same thing, for example, dementia praecox and schizophrenia. Sometimes a word, such as hysteria, carries many different meanings depending on the psychiatrist's theoretical orientation.

Psychiatry is hampered by the fact that the cause of many mental illnesses is unknown, and so convenient diagnostic distinctions cannot be made among such illnesses as they can, for instance, in infectious medicine, where infection with a specific type of bacterium is a reliable indicator for a diagnosis of tuberculosis. But the greatest difficulties presented by mental disorders as far as classification and diagnosis are concerned is that the same symptoms are often found in patients with different or unrelated disorders, or a patient may show a mix of symptoms properly belonging to several different disorders. Thus, although the categories of mental illness are defined according to symptom patterns, course, and outcome, the illnesses of many patients constitute intermediate cases between such categories, and the categories themselves may not necessarily represent distinct disease entities and are often poorly defined.

Mixed symptoms in many mental disorders

The two most frequently used systems of psychiatric classification are the *International Classification of Diseases* produced by the World Health Organization and the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III) produced by the American Psychiatric Association. The ninth revision of the former, published in 1977, is widely used in western Europe and other parts of the world for epidemiological and administrative purposes. Its nomenclature is deliberately conservative in conception so that it can be used by clinicians and mental health care systems in different countries.

This article, however, will follow the third edition of the DSM-III, which was published in 1980 and revised in 1987. The DSM-III differs from the International Classification in its introduction of precisely described criteria for each diagnostic category; its categorizations are usually based upon the detailed description of symptoms.

The DSM-III has been widely used, especially in the United States, and its detailed descriptions of diagnostic criteria have been useful in eradicating the inconsistencies of earlier classifications. However, there are still some major problems in its everyday clinical use. Chief among these problems is the DSM-III's innovative and controversial abandonment of the general categories of psychosis and neurosis in its classificatory scheme. These terms have been and still are widely used to distinguish between classes of mental disorders.

Psychoses are major mental illnesses that are characterized by severe symptoms such as delusions, hallucinations, disturbances of the thinking process, and defects of judgment and insight. Psychotic patients exhibit a disturbance or disorganization of thought, emotion, and behaviour so profound that they are often unable to function in everyday life and may be incapacitated or disabled. The psychotic patient is often unable to realize that his subjective perceptions and feelings do not correlate with objective reality, a phenomenon evinced by psychotics who do not know or will not believe that they are ill despite the distress they feel and their obvious confusion concerning the outside world. The psychoses are broadly divided into organic and functional psychoses. In organic psychoses the mental disturbance results from a physical defect of or damage to the brain. In functional psychoses, notably the schizophrenias and affective psychoses, no physical brain disease is evident upon clinical examination, but there is some research evidence pointing to an underlying organic abnormality.

Neuroses, or psychoneuroses, are less serious disorders in which a person may experience such negative feelings as anxiety or depression and his functioning may be significantly impaired, but his personality remains relatively intact, he maintains a capacity for recognizing and objectively evaluating reality, and he is basically able to function in everyday life. In contrast to the psychotic, the neurotic patient knows or can be made to realize that he is ill, and he usually wants to get well. His chances for recovery are better than those of the psychotic patient. The symptoms of neurosis sometimes resemble the coping mechanisms used in everyday life by most people, but in the neurotic person these defensive reactions are inappropriately severe or prolonged in response to an external stress. Anxiety disorders, phobic disorders, conversion disorders, obsessive-compulsive disorders, and depressive disorders have been traditionally classed as neuroses.

Responsiveness of neuroses to therapy

There are various other mental illnesses, such as personality disorders (or "character disorders"), that cannot be classed as either psychoses or neuroses.

Epidemiology. Epidemiology is the study of the distribution of disease in different populations. Prevalence denotes the number of cases of a condition present at a particular time or over a specified period, while incidence denotes the number of new cases occurring in a defined time period. Epidemiology is also concerned with the social, economic, or other contexts in which mental illnesses arise.

The understanding of mental disorders is aided by knowledge of the rate and frequency with which they occur in different societies and cultures. Looking at the worldwide prevalence of mental disorders reveals many surprising findings. It is remarkable, for instance, how constant the

rate for schizophrenia is; in widely different cultures there is generally a lifetime risk of developing the illness of just under 1 percent.

Gradual historical changes in the incidence and prevalence of particular disorders have often been described, but it is very difficult to obtain firm evidence that such changes have actually occurred. On the other hand, prevalence has been seen to increase for a few syndromes due to general changes in living conditions over time. For example, dementia inevitably develops in some 20 percent of those persons over age 80, so that with the increase in life expectancy common to developed countries the number of people with dementia is bound to increase. Other factors, such as the presence of small quantities of aluminum in drinking water, may also play a part in the increased prevalence of dementia. There also seems to be some evidence of an increased prevalence of affective disorders over the last century.

Several large-scale epidemiological studies have been conducted to determine the incidence and prevalence of mental disorders in the general population. Simple statistics based on those people actually under treatment for mental disorders cannot be relied upon in making such a determination, because the number of those who have sought treatment is substantially smaller than the actual number of people afflicted with mental disorders, many of whom do not seek professional treatment. Moreover, surveys to determine incidence and prevalence depend for their statistics on the clinical judgment of the survey takers, which can always be fallible because there are no objective tests for the assessment of mental illness. Given such objections, one ambitious study conducted by the National Institutes of Mental Health in the United States examined thousands of persons in several American localities and yielded the following results concerning the prevalence of mental disorders in the general population. About 0.6 percent of those surveyed were found to be schizophrenic, 0.5 percent had a manic episode, 5 percent suffered from depression, 5 percent suffered from phobias or other anxiety disorders, and about 1 percent had obsessive-compulsive disorders.

Schizophrenia and socioeconomic class

There is a relatively strong epidemiological association between socioeconomic class and the occurrence of certain types of mental disorders and of general patterns of mental health. One study found that the lower the socioeconomic class, the greater the prevalence of psychotic disorders; schizophrenia was found to be 11 times more frequent among the lowest of the five classes surveyed (unskilled manual workers) than among the highest class (professionals). (Anxiety disorders were found to be more common among the middle class, however.) Two possible explanations for the elevated frequency of schizophrenia among the poor would be that schizophrenics "drift downward" to the lowest socioeconomic class because they are impaired by their illness, or alternatively that unfavourable sociocultural conditions create circumstances that help induce the illness.

The manifestation of particular psychiatric symptoms is sometimes closely associated with particular epochs or periods in life. The symptoms of infantile autism are usually evident by early childhood, for example. Childhood and adolescence may produce a variety of psychiatric symptoms peculiar to those periods of life. Anorexia nervosa, several types of schizophrenia, drug abuse, and manic-depressive psychosis often first appear during adolescence or in young adult life. Alcohol dependence and its consequences, paranoid schizophrenia, and repeated attacks of depressive illness are more likely to occur in middle age. Involutional melancholia and presenile dementias typically occur in late middle age, while senile and arteriosclerotic dementias are characteristic of the elderly.

There are also marked sex differences in the incidence of certain types of mental illness. For instance, anorexia nervosa is 20 times more common in girls than boys; schizophrenia occurs more commonly in men than women, and they tend to develop the illness at a younger age; depressive illness is more common in women than men; and many sexual deviations occur almost exclusively in men.

THEORIES OF CAUSATION

Very often the etiology, or cause, of a particular type of mental disorder is unknown or is understood only to a very limited extent. The situation is complicated by the fact that a mental disorder such as schizophrenia may be caused by a combination and interaction of several factors, including a probable genetic predisposition to develop the disease, a postulated biochemical imbalance in the brain, and a cluster of stressful life events that help to precipitate the actual onset of the illness. The predominance of these and other factors probably varies from patient to patient in schizophrenia. A similarly complex interaction of constitutional, developmental, and social factors can influence the formation of neurotic disorders.

No single theory of causation can explain all mental disorders or even all those of a particular type, and, moreover, the same type of disorder may have different causes in different patients; *e.g.,* an obsessive-compulsive disorder may have its origins in a biochemical imbalance, in an unconscious emotional conflict, in faulty learning processes, or in a combination of these. The fact that quite different therapeutic approaches can produce equal improvements in different patients with the same type of disorder underscores the complex and ambiguous nature of the causes of mental illness. The major theoretical and research approaches to the causation of mental disorders are treated below.

Organic and hereditary etiologies. Organic explanations of mental illness have usually been genetic, biochemical, neuropathological, or a combination of these.

Genetics. The study of the genetic causes of mental disorders involves the statistical analysis of the frequency of a particular disorder's occurrence among individuals who share related genes; *i.e.,* family members and particularly twins. Family risk studies compare the observed frequency of occurrence of a mental illness in close relatives of the patient with its frequency in the general population. First degree relatives (parents, siblings, and children) share 50 percent of their genetic material with the patient, and higher rates of the illness in these relatives than expected indicate a possible genetic factor. In twin studies the frequency of occurrence of the illness in both members of pairs of identical (monozygous) twins is compared with its frequency in both members of a pair of fraternal (dizygous) twins. A higher concordance for disease among the identical than the fraternal twins suggests a genetic component. Further information on the relative importance of genetic and environmental factors accrues from comparing identical twins reared together with those reared apart. Adoption studies comparing adopted children whose biological parents had the illness with those whose parents did not can also be useful in separating biological from environmental influences.

Schizophrenia and genetic predisposition

Such studies have pointed up a clear role for genetic factors in the causation of schizophrenia. When one parent is found to have the disorder, the probability that his children will develop schizophrenia is at least 10 times higher (about a 12-percent risk probability) than it is for children in the general population (about a 1-percent risk probability). If both parents are schizophrenic, their children stand anywhere from a 35- to 65-percent probability of becoming schizophrenic. If one member of a pair of fraternal twins develops schizophrenia, there is about a 10-percent chance that the other twin will also develop the disorder. If one member of a pair of identical twins has schizophrenia, the other identical twin has at least a 40–50 percent chance of developing the disease.

Genetic factors seem to play a less significant role in the causation of other psychotic disorders and in personality disorders, and they seem to be even less of a factor in the causation of the neuroses.

Biochemistry. If a mental disease is caused by a biochemical abnormality, investigation of the brain at the site where the biochemical imbalance occurs should show neurochemical differences from normal. In practice such a simplistic approach is fraught with practical, methodological, and ethical difficulties. The living human brain is not readily accessible to direct investigation, and the dead one undergoes chemical change; moreover, findings

of abnormalities in cerebrospinal fluid, blood, or urine may have no relevance to the question of a presumed biochemical imbalance in the brain. Human mental illnesses cannot be adequately studied using animals as analogues, because most mental disorders either do not occur or are not recognizable in animals. Even when biochemical abnormalities have been found in mental patients, it is difficult to know whether such abnormalities are the cause or the result of the illness, or of its treatment, or of other consequences. Despite these problems, progress has been made in unraveling the biochemistry of affective disorders, schizophrenia, and some of the dementias.

Certain drugs have been demonstrated to have beneficial effects upon mental illnesses. Antidepressant, antipsychotic, and antianxiety drugs are thought to achieve their therapeutic results by the selective inhibition or enhancement of the quantities, action, or breakdown of neurotransmitters in the brain. Neurotransmitters are a group of chemical agents that are released by neurons (nerve cells) to stimulate neighbouring neurons, thus allowing impulses to be passed from one cell to the next throughout the nervous system. Neurotransmitters play a key role in transmitting nerve impulses across the microscopic gap (synaptic cleft) that exists between neurons. The release of such neurotransmitters is stimulated by the electrical activity of the cell. Among the principal neurotransmitters are norepinephrine, dopamine, and serotonin. Some neurotransmitters excite or activate neurons, while others act as inhibiting substances. Abnormally low or high concentrations of neurotransmitters at sites in the brain are thought to change the synaptic activities of neurons, thus ultimately leading to the disturbances of mood, emotion, or thought found in various mental disorders.

Importance of neurotransmitters

Neuropathology. In the past the pathological study of the brain at post mortem revealed information upon which great advances in understanding the etiology of neurological and some mental disorders were based, leading to the German psychiatrist Wilhelm Griesinger's postulate: "all mental illness is disease of the brain." The application of the principles of pathology to general paralysis of the insane, one of the most common conditions found in mental hospitals in the late 19th century, resulted in the discovery that this was a form of neurosyphilis and was caused by infection with the spirochete bacterium *Treponema pallidum*. The examination of the brains of patients with other forms of dementia has given useful information concerning other causes of this syndrome, for example, Alzheimer's disease and arteriosclerosis. The pinpointing of abnormalities of specific areas of the brain has aided understanding of some abnormal mental functions, such as disturbances of memory or speech disorders. However, no abnormal pathology has been demonstrated for most mental disorder.

Psychodynamic etiologies. Up to the 1970s theories of the etiology of mental disorders, especially of neuroses and personality disorders, were dominated in the United States by Freudian psychoanalysis and the derivative theories of the post-Freudians. In western Europe the influence of Freudian theory upon psychiatric theory diminished after World War II.

Theories of personality development. Freudian and other psychodynamic theories view neurotic symptoms as arising from intrapsychic conflict; that is, as being caused by conflicting motives, drives, impulses, and feelings held within various components of the mind. Central to psychoanalytic theory is the postulated existence of the unconscious, which is that part of the mind whose processes and functions are inaccessible to the individual's conscious awareness or scrutiny. One of the functions of the unconscious is thought to be that of a repository for traumatic memories, feelings, ideas, wishes, and drives that are threatening, abhorrent, anxiety-provoking, or socially or ethically unacceptable to the individual. These mental contents may at some time be pushed out of conscious awareness but remain actively held in the unconscious. This process is a defense mechanism for protecting the individual from the anxiety or other psychic pain associated with those contents and is known as repression. The repressed mental contents held in the unconscious retain much of the psychic energy or power that was originally attached to them, however, and they can continue to influence significantly the mental life of the individual even though (or because) he is no longer aware of them.

The natural tendency for repressed drives or feelings, according to this theory, is to break through into conscious awareness so that the individual can seek the gratification, fulfillment, or resolution of them. But this threatened release of forbidden impulses or memories provokes anxiety and is seen as threatening, and a variety of psychic processes known as defense mechanisms may then come into play to provide relief from the state of psychic conflict. Through reaction formation, projection, regression, sublimation, rationalization, and other defense mechanisms, some component of the unwelcome mental contents can emerge into consciousness in a disguised or attenuated form, thus providing partial relief to the individual. Later, perhaps in adult life, some event or situation in the person's life triggers the abnormal discharge of the dammed-up or strangulated emotional energy in the form of neurotic symptoms in a manner mediated by defense mechanisms. Such symptoms can form the basis of neurotic disorders such as conversion and somatoform disorders, anxiety disorders, obsessional disorders, and depressive disorders. Since the symptoms represent a compromise within the mind between letting the repressed mental contents out and continuing to deny all conscious knowledge of them, the particular character and aspects of an individual's symptoms and neurotic concerns bear an inner meaning that symbolically represents the underlying intrapsychic conflict. Psychoanalysis and other dynamic therapies associate the patient's controlled and therapeutic recovery to conscious awareness of repressed mental conflicts, and his understanding of their influence on both his past history and present difficulties, with the relief of symptoms and improved mental functioning.

Childhood experience and neurosis

Freudian theory views childhood as the primary breeding ground of neurotic conflicts. This is because children are relatively helpless and are dependent on their parents for love, care, security, and support and because their psychosexual, aggressive, and other impulses are not yet integrated into a stable personality framework. Children are thus liable to emotional traumas, deprivations, and frustrations which they lack the resources to cope with and which can become grounds for intrapsychic conflicts that are not resolved but rather are merely held in abeyance through repression, producing insecurity, unease, or guilt and subtly influencing the individual's developing personality, interests, attitudes, and ability to cope with later stresses.

Non-Freudian psychodynamics. Psychoanalytic theory's emphasis on the unconscious mind and its influence on human behaviour resulted in a proliferation of other, related theories of causation incorporating many basic psychoanalytic precepts. Most subsequent psychotherapies have stressed in their theories of causation aspects of earlier, maladaptive psychological development that had been missed or underemphasized by orthodox psychoanalysis, or they have incorporated insights taken from learning theory. The Swiss psychiatrist Carl Jung, for instance, concentrated on the individual's need for spiritual development and concluded that neurotic symptoms could arise from a lack of self-fulfillment in this regard. The Austrian psychiatrist Alfred Adler emphasized the importance of feelings of inferiority and the unsatisfactory attempts to compensate for it as important causes of neurosis. Neo-Freudian authorities such as Harry Stack Sullivan, Karen Horney, and Erich Fromm modified Freudian theory by emphasizing social relationships and cultural and environmental factors as being important in the formation of mental disorders. Many other highly specific theories of causation have been developed by particular psychotherapies, and in general, psychiatric scrutiny has come to extend far beyond the confines of early psychosexual development that were originally posited by Freud as the prime ground for the causation of neuroses. More modern psychodynamic theories have moved away from the idea of explaining and treating neurosis on the basis of a defect in a single psychological system and have instead adopted

a more complex notion of multiple causes, including emotional, psychosexual, social, cultural, and existential ones. A notable trend in the more recently developed psychotherapies has been the incorporation of approaches derived from theories of learning. Such psychotherapies pay special attention to acquired, faulty mental processes and maladaptive behavioral responses that act to sustain neurotic symptoms, and there has generally been increased interest in the patient's present circumstances and his learned responses to those conditions as a causative factor in mental illness. In this way, psychoanalytic theory and behavioral theory have tended somewhat to converge and intermingle in their views of disease causation.

Behavioral etiology. Behavioral theories for the causation of mental disorders, especially neurotic symptoms, are based upon learning theory, which was in turn largely derived from the study of the behaviour of animals in laboratory settings. Most important theories in this area arose out of the work of the Russian physiologist Ivan Pavlov and such American psychologists as Edward L. Thorndike, Clark L. Hull, John B. Watson, Edward C. Tolman, and B.F. Skinner. In the classical Pavlovian model of conditioning, an unconditioned stimulus is followed by an appropriate response; for example, food placed in a dog's mouth is followed by the dog salivating. If a bell is rung just before food is offered to a dog, eventually the dog will salivate at the sound of the bell only, even though no food is offered. Because the bell could not originally evoke salivation in the dog (and hence was a neutral stimulus) but came to evoke salivation because it was repeatedly paired with the offering of food, it is called a conditioned stimulus. The dog's salivation at the sound of the bell alone is called a conditioned response. If the conditioned stimulus (the bell) is no longer paired with the unconditioned stimulus (the food), extinction of the conditioned response gradually occurs (the dog ceases to salivate at the sound of the bell alone).

The role of conditioned responses

Behavioral theories for the causation of mental disorders rest largely upon the assumption that the symptoms or symptomatic behaviour found in persons with various neuroses (particularly phobias and other anxiety disorders) can be regarded as learned behaviours that have been built up into conditioned responses. In the case of phobias, for example, a person who has once been exposed to an inherently frightening situation afterward experiences anxiety even at neutral objects that were merely associated with that situation at the time but that should not reasonably produce anxiety; *e.g.,* a child who has had a painful session with a dentist may subsequently have an unreasonable dread of men in white coats or of any kind of drill. The neutral object alone is enough to arouse anxiety, and the person's subsequent effort to avoid that object is a learned behavioral response that is self-reinforcing, since the person does indeed procure a reduction of his anxiety by avoiding the feared object and is thus likely to continue to avoid it in the future. But his fear of the object persists, since it is only by confronting the object that he can eventually lose his irrational, association-based fear of it.

Other etiologies. Mental illness may be contagious in a psychological sense; that is, close contact with an individual who has symptoms may result in the transmission of those symptoms to one or many others who were previously unaffected. This may occur either through the powerful influence of long-term cohabitation of one person with one other—a phenomenon known as folie à deux—or through the volatile collective emotions of a group—mass hysteria. Epidemic, communicated, or mass neurosis is particularly likely to spread through a closed community such as a boarding school. The transmission of symptoms occurs first to those who are psychologically vulnerable, and stops once the closed population is scattered.

Social values can sometimes determine or encourage the formation of particular syndromes. Prime examples of this are anorexia nervosa and bulimia nervosa, which predominantly affect young females in affluent Western societies. The value and attractiveness of physical slimness are communicated via the media and respected adults, and an eating disorder resulting in emaciation subsequently occurs in some neurotically susceptible individuals.

Another approach to the causation of mental disorders focuses on the effects and consequences of stress, which is a state of bodily or mental tension resulting from external factors such as marital conflicts, excessive work demands, or serious financial problems. Stress is known to cause psychosomatic illnesses, and an accumulation of stressful life events can help cause depression in psychologically vulnerable individuals.

MAJOR DIAGNOSTIC CATEGORIES

Organic mental disorders. This category includes both those psychological or behavioral abnormalities that arise from structural disease of the brain and also those that arise from brain dysfunction caused by disease outside the brain. These conditions differ from those of other mental illnesses in that they have a definite and ascertainable cause—*i.e.,* brain disease. Treatment, when possible, is aimed at both the symptoms and the underlying physical dysfunction in the brain.

There are several types of psychiatric syndromes that arise from organic brain disease, chief among them being dementia and delirium. Dementia is a gradual and progressive loss of such intellectual abilities as thinking, remembering, paying attention, judging, and perceiving, without an accompanying disturbance of consciousness. The syndrome may also be marked by the onset of personality changes in the sufferer. Dementia is usually a chronic condition and frequently worsens over the long term. Delirium is a diffuse or generalized intellectual impairment marked by a clouded or confused state of consciousness, an inability to attend to one's surroundings, difficulty in thinking coherently, a tendency to perceptual disturbances such as hallucinations, and difficulty in sleeping. Delirium is generally an acute condition and is not long-lasting. Other specific psychological impairments associated with organic brain disease are amnesia (in which there is a gross loss or disorder of recent memory and time-sense without other intellectual impairment), recurring or persistent hallucinations or delusions, or marked personality changes.

Dementia and delirium

In the diagnosis of suspected organic disorders, a full history has first to be taken from the patient and his mental state must be examined in detail, with additional tests for particular functions added if necessary. A physical examination is also carried out with special attention to the central nervous system. In order to determine whether a metabolic or other biochemical imbalance is causing the condition, blood and urine tests, liver function tests, thyroid function tests, and other evaluations may be carried out. Chest and skull X rays are made, and computerized axial tomography (CAT scan) is used to reveal focal or generalized brain disease. Electroencephalography may show localized abnormalities in the electrical conduction of the brain caused by a lesion. Detailed psychological testing may reveal more specific perceptual, memory, or other disabilities.

Senile and presenile dementia. In these dementias there is a progressive intellectual impairment that proceeds to lethargy, inactivity, and gross physical deterioration and eventually to death within a few years. Presenile dementias are arbitrarily defined as those that begin in persons under the age of 65. In old age the most common causes of dementia are Alzheimer's disease and cerebral arteriosclerosis. Dementia from Alzheimer's disease usually begins in people over age 65 and is much more common in women than in men. It begins with incidences of forgetfulness, which become more frequent and serious, and the disturbances of memory, personality, and mood progress steadily toward physical deterioration and death within a few years. In dementia caused by cerebral arteriosclerosis there are multiple areas of destruction of the brain caused by pieces of the damaged lining of arteries outside the skull lodging in the small arteries of the brain. The course of the illness is stepwise, with rapid deterioration followed by periods of slight improvement. Death may be delayed slightly longer than with dementia from Alzheimer's disease and often occurs from ischemic heart disease (heart attack) or from massive cerebral infarction, causing a stroke.

Other causes of dementia include Pick's disease, a rare inherited condition that occurs in women twice as often as men, usually between the ages of 50 and 60; Huntington's chorea, an inherited disease that usually begins at about the age of 40 with involuntary movements and proceeds to dementia and death within 15 years; and Creutzfeldt-Jakob disease, a rare condition that is probably caused by a transmissible agent known as a slow virus. Head injury, for instance, resulting from a boxing career or from an accident, may produce dementia. Infection, for example, with neurosyphilis or encephalitis, various tumours, toxic conditions such as chronic alcoholism or heavy metal poisoning, metabolic illnesses such as liver failure, reduced oxygen to the brain due to anemia or carbon monoxide poisoning, and the inadequate intake or metabolism of certain vitamins may all result in dementia.

There is no specific treatment for the symptoms of dementia; the underlying physical cause needs to be identified and treated when possible. The aims in the care of the demented patient are to relieve distress, prevent behaviour that might result in accident, and optimize his remaining physical and psychological faculties.

Substance-induced organic disorders. A variety of psychiatric conditions can result from the use or abuse of alcohol or other drugs. Mental disorders resulting from the ingestion of alcohol include intoxication, withdrawal, hallucinations, and amnesia. Similar syndromes may occur following the use of other drugs (see ALCOHOL AND DRUG CONSUMPTION). Those most commonly used nonmedically to alter mood are barbiturates, opioids (such as heroin), cocaine, amphetamines, hallucinogens such as lysergic acid diethylamide (LSD), cannabis, tobacco, and caffeine. Treatment is directed at alleviating symptoms and preventing the patient's further consumption of the substance.

Other organic syndromes. Delirium occurs secondarily to many other physical conditions such as drug intoxication or withdrawal, metabolic disorders (for example, liver failure or low blood sugar), infections such as pneumonia or meningitis, head injuries, brain tumours, epilepsy, or nutritional or vitamin deficiency. There are a clouding or confusion of consciousness and disturbances of thinking, behaviour, perception, and mood, with disorientation being prominent. Treatment is aimed at the underlying physical condition.

Damage to different areas of the brain may cause particular psychological symptoms. Damage to the frontal lobe of the brain may manifest itself in such disturbances of behaviour as loss of inhibitions, tactlessness, and overtalkativeness. Lesions of the parietal lobe may result in difficulties of speech and language or of the perception of space. Lesions of the temporal lobe may lead to emotional instability, aggressive behaviour, or problems with learning new information.

Substance use disorders. This category refers to maladaptive behaviour associated with the regular nonmedical use of substances that affect the central nervous system. Substance abuse implies a sustained pattern of pathological use resulting in impairment of the drug abuser's social or occupational functioning. Substance dependence implies tolerance, in which markedly increased amounts of the drug must be administered to achieve the same effect, and withdrawal, in which symptoms follow the cessation of drug use or decreases in the dose of the substance.

Schizophrenia. The term schizophrenia was introduced by the Swiss psychiatrist Eugen Bleuler in 1911 to describe what he considered to be a group of severe mental illnesses with related characteristics; "schizophrenia" eventually replaced the earlier term dementia praecox, which the German psychiatrist Emil Kraepelin had first used in 1899 to distinguish the disease from manic-depressive psychosis. Schizophrenic patients have a wide variety of symptoms; thus, although different authorities may agree as to whether a particular patient suffers from the condition, they might disagree about which constellations of symptoms are essential in clinically defining schizophrenia.

In 12 very different countries, rates for schizophrenia have been found to be surprisingly similar, the annual prevalence, that is, the number of cases both old and new recorded in one year, being between two and four per 1,000 persons. The lifetime risk of developing the illness is between seven and nine per 1,000. Schizophrenia is the single largest cause of admissions to mental hospitals, and it accounts for an even larger proportion of the permanent populations of such institutions. It is a severe and frequently chronic illness that typically first manifests itself during the teen years or during early adult life. More severe levels of impairment and personality disorganization are reached in schizophrenia than in almost any other mental disorder.

Severity of schizophrenia

Clinical features. The principal clinical signs of schizophrenia are delusions, hallucinations, a loosening or incoherence of a persons's thought processes and train of associations, deficiencies in feeling appropriate or normal emotions, and a withdrawal from reality. A delusion is a false or irrational belief that is firmly held despite obvious or objective evidence to the contrary. The delusions of schizophrenics may be persecutory, grandiose, religious, sexual, or hypochondriacal in nature, or they may be concerned with other topics. Delusions of reference, in which the patient attributes a special, irrational, and usually negative significance to people, objects, or events in relation to himself, are common in the disease. Especially characteristic of schizophrenia are delusions in which the patient believes his thinking processes, parts of his body, or his actions or impulses are controlled or dictated by some external force. Schizophrenic delusions are frequently bizarre or absurd.

Hallucinations are false sensory perceptions that are experienced without an external stimulus but that nevertheless seem real to the subject. Auditory hallucinations, experienced as "voices" and characteristically heard commenting negatively about the patient in the third person, are prominent in schizophrenia. Hallucinations of touch, taste, smell, and bodily sensation also occur. Disorders of thinking vary in nature but are quite common in schizophrenia. The thought disorders may consist of a loosening of associations, so that the speaker jumps from one idea or topic to another unrelated one in an illogical, inappropriate, or disorganized way. At its most serious, this incoherence of thought extends into pronunciation itself, and the speaker's words become garbled or unrecognizable. Speech may also be overly concrete and inexpressive; it may be repetitive, or, though voluble, it may convey little or no real information. Usually a schizophrenic patient has little or no insight into his own condition and realizes neither that he is suffering from mental illness nor that his thinking is disordered.

Among the so-called negative symptoms of schizophrenia are a blunting or flattening of the person's ability to experience (or at least to express) emotion, indicated by speaking in a monotone and by a peculiar lack of facial expressions. The person's sense of self (*i.e.,* of who he is) may be disturbed. He may be apathetic and may lack the drive and ability to pursue a course of action to its logical conclusion, or he may withdraw from the world, become detached from others, and become preoccupied with silly, bizarre, or nonsensical fantasies. Such symptoms are more typical of chronic rather than of acute schizophrenics.

Different authorities have recognized many different types of schizophrenia, and there are intermediate stages between the disease and other conditions. Four major types of schizophrenia are still recognized by the DSM-III: the disorganized or hebephrenic type, the catatonic type, the paranoid type, and the simple or undifferentiated type. Hebephrenic schizophrenia is characterized by grossly inappropriate, shallow, or silly emotional responses and by incoherent thought and speech. Catatonic schizophrenia is marked by striking motor behaviour, such as remaining motionless in a rigid posture for hours or days, and by stupor or mutism. Paranoid schizophrenia is marked by the presence of prominent delusions of a persecutory and/or grandiose nature. Undifferentiated schizophrenia is marked by an insidious or gradual reduction in the person's interest in and relations with the external world and by a pervasive impoverishment of his personality and emotional responses.

Course and prognosis. The course of schizophrenic ill-

Variable course of schizophrenia

ness is extremely variable. It may be said that roughly one-third of schizophrenic patients make a complete recovery and have no further recurrence, one-third have recurrent episodes of the illness, and one-third deteriorate into chronic schizophrenia with severe disability. The prognosis for schizophrenics has improved during the 20th century due to the use of antipsychotic drugs and community supportive measures.

About 10 percent of schizophrenic patients die by suicide. The prognosis of schizophrenia is poor when it has a gradual rather than a sudden onset, when the patient is quite young at onset, when there is a long duration of illness, when the patient exhibits blunted feelings or has displayed an abnormal personality previous to the onset of the disease, and when such social factors as never having been married, poor sexual adjustment, a poor work record, or social isolation exist in the patient's personal history.

Etiology. An enormous amount of research has been carried out to try to determine the causes of schizophrenia. Family, twin, and adoption studies provide strong evidence to support an important genetic contribution, but the mode of inheritance is not known. Stressful life events are known to trigger or quicken the onset of schizophrenia or to cause relapse. Some abnormal neurological signs have been found in schizophrenics, and it is possible that brain damage, perhaps occurring at birth, may be a cause in some cases. Various biochemical abnormalities have been reported in schizophrenics, but the evidence for the causal relevance of these abnormalities is incomplete.

Much research has been carried out to determine whether the types of communication used in the families of schizophrenics or the parental care in such families help produce the disease. There has also been extensive interest in such factors as social class, place of residence, migration, and social isolation. Neither family dynamics nor social disadvantage have been proved to be causative agents.

Treatment. The most successful treatment approaches combine the use of drugs, psychotherapy, and supportive therapy. In acute schizophrenia, phenothiazine, chlorpromazine, or butyrophenone drugs such as haloperidol are of proven efficacy in relieving or eliminating such symptoms as delusions, hallucinations, thought disorders, agitation, and violent behaviour. Long-term maintenance on such drugs also reduces the rate of relapse. Psychotherapy serves to relieve the patient's feelings of helplessness and isolation, buttress his healthy or positive tendencies, and help him to distinguish between his psychotic perceptions and reality and to deal with any underlying emotional conflicts that might be exacerbating his condition. Occupational therapy for those in day care and regular visits from a social worker or community psychiatric nurse for outpatients are beneficial. It is sometimes useful to counsel the relatives of schizophrenic patients living at home in their way of dealing with the patient's symptoms.

Paranoid disorders. Paranoia is a syndrome in which a person thinks or believes, without justification, that other people are plotting or conspiring against him, are harassing him, or are otherwise persecuting or trying to harm him in some way. Paranoid thinking frequently causes a person to interpret or exaggerate innocuous or trivial incidents in a self-referrent way; *e.g.,* to see two people talking at a distance and to irrationally assume that they are plotting against or criticizing him. Grandiosity or delusions of grandeur, which consist of exaggerated and unjustified ideas of a person's own importance, wealth, or power, frequently coexist with the classic persecutory orientation in paranoia. Paranoia or paranoid thinking can be a prominent or primary feature in schizophrenia (paranoid schizophrenia), personality disorders, senile dementias, affective disorders, and manic-depressive psychoses, and indeed it is difficult to demarcate strictly what the DSM-III defines as paranoid disorders proper. Persons with paranoid disorders may be otherwise normal people who are simply abnormally suspicious, or they may have an unshakable and highly elaborate delusional system involving worldwide conspiracies against them. A special type of paranoia is delusional jealousy, in which a person delusionally believes or suspects that his spouse is having sexual relations with someone else. A paranoid disorder can seriously impair an individual's social or marital functioning, but his thinking remains clear and orderly, his intellectual functioning is impaired only minimally or not at all, and the core of his personality remains intact. Many people with paranoid disorders can have normal or near-normal careers. The treatment of persons with paranoid disorders involves the use of antipsychotic drugs, frequently on a long-term maintenance basis.

Affective disorders. These disorders are usually restricted to just two abnormalities of mood—depression and elation, or mania. (Mood is a predominant emotion that colours the individual's entire psychic life.)

Depression and mania

Depression is characterized by a sad or hopeless mood, pessimistic thinking, a loss of enjoyment and interest in one's usual activities and pastimes, reduced energy and vitality, increased fatigue, slowness of thought and action, loss of appetite, and disturbed sleep or insomnia. Depression must be distinguished from the grief and low spirits felt in reaction to the death of a loved one or some other unfortunate circumstance. The most dangerous consequence of severe depression is suicide.

Mania is characterized by an elated or euphoric mood, quickened thought and accelerated, loud or voluble speech, overoptimism and heightened enthusiasm and confidence, inflated self-esteem, heightened motor activity, irritability, excitement, and a decreased need for sleep. The manic individual may become injured, commit illegal acts, or suffer financial losses due to the poor judgment and risk-taking behaviour he displays when in the manic state.

There are enormous problems in the classification of affective disorders, particularly of depression, and the various clinical distinctions made by different authorities are difficult to correlate with particular sets of symptoms or particular causes. An important distinction, however, is made between depressions that are endogenous (*i.e.,* arising independently of environmental influences and presumably caused by a biochemical imbalance) and those that are reactive (*i.e.,* arising in response to external stresses or trauma).

Major affective disorders. The DSM-III defines two major, or severe, affective disorders: bipolar disorder and major depression. A person with bipolar disorder, which has traditionally been called manic-depressive psychosis, typically experiences discrete episodes of depression and then of mania lasting for a few weeks or months, with intervening periods of complete normality. The sequence of depression and mania can vary extremely from patient to patient and within one individual, with either mood abnormality predominating in duration and intensity. Depressive mood swings typically occur more often and last longer than manic ones, though there are patients who have episodes only of mania. Patients with bipolar disorder frequently also show such psychotic symptoms as delusions, hallucinations, paranoia, or grossly bizarre behaviour.

The lifetime risk for developing bipolar disorder is about 0.7 percent and is about the same for men and women. The onset of the illness often occurs around the age of 30, and the illness persists over the long term. The predisposition to develop bipolar disorder is partly genetically inherited. Antipsychotic drugs such as chlorpromazine or haloperidol are used for the treatment of acute mania. Lithium carbonate has proved effective in both treating and preventing recurrent attacks of mania.

Severe and long-lasting depression without the presence of mania is classified by the DSM-III as major depression. Depression is a much more common illness than mania, and there are indeed many sufferers from depression who have never experienced mania. Major depression may occur as a single episode or it may be recurrent. It may also exist with or without melancholia and with or without psychotic features. Melancholia implies the so-called biological symptoms of depression: early-morning waking; daily variations of mood with depression most severe in the morning; loss of appetite and weight; constipation; and loss of interest in love and sex. Melancholia is a particular depressive syndrome that is relatively more responsive to physical methods of treatment, such as drugs and electroconvulsive therapy.

It is estimated that the annual incidence of major depression is about 140 for men and 4,000 for women per 100,000 population. While the rates for major depression in men increase with age, the peak for women is between the ages of 35 and 45. There is a serious risk of suicide with the illness; of those who have a severe depressive disorder, about one-sixth eventually kill themselves. The loss of one's parents or other childhood traumas or deprivations can increase a person's vulnerability to depression later in life, and stressful life events, especially where some type of loss is involved, are in general potent precipitating causes of the illness. It seems that both psychosocial and biochemical mechanisms are important in causing depression. Of the latter factor, the best-supported hypotheses suggest that the faulty regulation of the release of one or more naturally occurring amines at sites in the brain where the transmission of nerve impulses takes place is the basic cause, with a deficiency of the amines resulting in depression and an excess causing mania. The most likely candidates for the suspect amines are the biological monoamines (norepinephrine, dopamine, and 5-hydroxytryptamine). The treatment of major depressive episodes usually requires antidepressant drugs; electroconvulsive therapy may also be helpful, as may cognitive psychotherapy.

Minor affective disorders. A less severe manifestation of the manic-depressive syndrome, in which the mood swings are present but not as extreme, is termed cyclothymic disorder. This illness is better considered a personality disorder of affective type; the prevailing mood swings are established in adolescence and continue throughout adult life.

Dysthmic disorder, or depressive neurosis, may occur on its own, but it more commonly appears along with other neurotic symptoms such as anxiety, phobia, and hypochondriasis. Where there are clear external grounds for a person's unhappiness, a dysthymic disorder is considered to be present when the depressed mood is disproportionately severe or prolonged in regard to the distressing experience, when there is a preoccupation with the precipitating situation, when the depression continues even after removal of the provocation, and when it impairs the individual's ability to cope with the specific stress.

Prevalence of minor depression

At any time, depressive symptoms may be found in one-sixth of the population, more commonly in women than men. Social factors are important etiologically, as evidenced in the high rates of depression found in urban women living without a male cohabitant, having three or more children, and lacking employment outside the home. Loss of self-esteem, feelings of helplessness and hopelessness, and losses of various types of "loved objects" are also seen as important causes of minor depression. The course and severity of dysthymic disorder is extremely variable—from a few weeks or months to several decades and from the mild impairment of social functioning to almost total incapacitation. Psychotherapy is the treatment of choice, although antidepressant medication may prove beneficial.

Anxiety disorders. Anxiety has been defined as a feeling of fear, dread, or apprehension that arises without a clear or appropriate real-life justification. Some authorities differentiate anxiety from true fear in that the latter is experienced in response to an actual threat or danger, such as those to one's physical safety. Anxiety, on the other hand, may arise in response to apparently innocuous situations or may be out of proportion to the actual degree of the external stress. Anxiety also frequently arises as a result of subjective emotional conflicts of whose nature the person himself may be unaware. Generally, intense, persistent, or chronic anxiety that is not justified in response to real-life stresses and that interferes with the individual's functioning is regarded as a manifestation of mental disorder. Anxiety is a symptom in many mental disorders, including schizophrenia, obsessive-compulsive disorders, posttraumatic stress disorders, and so on, but in phobias and other anxiety disorders proper, anxiety is the primary and frequently the only symptom.

The symptoms of anxiety are physical, psychological, and behavioral. Anxiety, especially during panic attacks, can manifest itself in a distinctive set of physical signs that arise from overactivity of the sympathetic nervous system or from tension in skeletal muscles. The sufferer experiences palpitations, dry mouth, dilatation of the pupils, shortness of breath, sweating, abdominal symptoms, tightness in the throat, trembling, and dizziness. Aside from the actual feelings of dread and apprehension, the psychological symptoms include irritability, difficulty with concentration, and restlessness. Anxiety may also be manifested in avoidance behaviour—running away from the feared object or situation.

Phobic disorder or neurosis. Phobias are neurotic states accompanied by intense dread of certain objects or situations that would not normally have such an effect. This type of anxiety is associated with a strong desire to avoid the dreaded object or situation. About six per 1,000 of the population suffer from a phobic disorder. There is a tendency for phobic symptoms, whatever their nature, to persist for many years unless treated, and the avoidance behaviour they produce can seriously limit the affected individual's movements and his social or occupational functioning.

Prevalence of phobias

People can have phobias about many different kinds of objects or situations, but three main divisions of phobic syndromes are made by the DSM-III: simple phobia, agoraphobia, and social phobia. Individuals with simple phobias may intensely fear a specific object or situation, for example, cats or thunderstorms; they have anxious thoughts upon anticipating contact with an object or event, for instance, upon hearing the weather forecast, and they try to avoid the object, as in staying indoors in order not to encounter a cat. Typically, agoraphobic patients have an intense fear of being alone in or being unable to escape from a public place or some other setting outside the home, such as a crowded bus or a supermarket. A social phobia is present when the individual has extreme anxiety in a social situation where he is under the scrutiny of others, such as eating in a restaurant or speaking at a meeting.

The treatment of phobic disorders is best approached by the use of behavioral therapy; dynamic psychotherapy and antianxiety drugs may be effective in some cases.

Anxiety states or neuroses. Anxiety disorders in which the anxiety is not aroused by any specific object or situation can basically be subsumed under the headings of panic disorder and generalized anxiety disorder. Panic attacks are characterized by the sudden onset of intense or overwhelming anxiety accompanied by any of the aforementioned physical signs, such as difficulty in breathing, sweating, palpitations, and so on. The fear and apprehension experienced in such attacks sometimes mount to what are known as feelings of doom. Clear precipitating circumstances may produce the initial feelings of intense anxiety. The panic attack may last for about a quarter of an hour and frequently recurs, either infrequently or several times a week. The disorder usually starts in young adults and may persist for many years.

A diffuse and persistent feeling of anxiety associated with no particular object or situation is termed general, or free-floating, anxiety and is classified by the DSM-III as generalized anxiety disorder. General anxiety is usually milder and less intense than in panic attacks, but it is longer lasting and may persist for several months or years, or on a recurrent basis. The most effective treatments vary according to the type of disorder and the individual patient. Psychotherapy and antianxiety drugs are often useful in treating generalized anxiety and panic attacks.

Obsessive-compulsive disorder or neurosis. In this condition an individual experiences obsessions or compulsions or both. Obsessions are recurring words, thoughts, ideas, or images that, rather than being experienced as voluntarily produced, seem to invade a person's consciousness despite his attempts to ignore, control, or suppress them. The obsessional thought or topic is perceived by the sufferer as inappropriate or senseless; the idea is recognized both as alien to his nature and yet as coming from inside himself. An obsession can take the form of a recurrent and vivid fantasy that is often obscene, disgusting, repugnant, or senseless. The patient with obsessional ruminations holds endless debates over mundane matters inside his head; *e.g.*, "Did I forget to lock the front door behind me?"

Obsessions in turn are frequently linked to compulsions. These are urges or impulses to perform repetitive acts that are apparently meaningless, unnecessary, stereotyped, or ritualistic. The compulsive person knows that the act to be performed is meaningless or unnecessary, but his failure or refusal to perform it brings on a mounting tension or anxiety that is temporarily relieved once the act is performed. Obsessional ruminations thus directly produce compulsive behaviour; *e.g.,* repeatedly checking and relocking an already secure front door. Most compulsive acts have a simple, ritualistic character and can involve checking, touching, hand-washing, or the repetition of particular words or phrases.

Drugs, psychotherapy, and behavioral therapy are selectively successful in treating obsessive-compulsive disorders, depending on the individual patient. The drug clomipramine has proved to be notably effective in reducing or even eliminating the symptoms in a large proportion of patients tested.

Posttraumatic stress disorder. In this condition symptoms develop in an individual after he has experienced a psychologically traumatic event. It is a category in the DSM-III classification but is not different in its symptomatology from certain other neurotic conditions; the distinctive feature is the presence of external trauma. The traumatic events can include serious automobile accidents, rape or assault, military combat, torture, incarceration in a concentration or death camp, and such natural disasters as floods, fires, or earthquakes.

A feature of this condition is the person's reexperiencing of the traumatic event in nightmares and in intrusive daytime fantasies. Sometimes an insignificant event, like a knock at the door, will precipitate a sudden terrifying recollection and an exaggerated startle response. Other symptoms include emotional numbing, a diminished ability to enjoy activities or relationships that were previously pleasurable, and difficulty with sleeping. Long-term symptoms of distress, marital and family problems, difficulties at work, and the abuse of alcohol and other drugs are characteristic impairments caused by the disorder.

The marked emotional symptoms may persist long after the traumatic event actually occurs. Some persons are more liable than others to develop the disorder, depending on personality traits, previous psychological disturbances, age, and genetic predisposition. Psychotherapy is the basic approach used in treating this disorder.

Somatoform disorders. In these conditions, the physical symptoms of the person suggest the presence of organic disease but no such organic disorder can be found upon physical examination and investigation, and instead there is positive evidence that the symptoms are caused by psychological factors. The production of these symptoms is not under voluntary control. The terms hypochondriasis and hysteria that traditionally designated these disorders are still widely used by psychiatrists.

Somatization disorder. This disorder was previously designated Briquet's syndrome; its essential features consist of multiple, recurrent physical complaints made over many years and starting in young adult life or adolescence. The sufferer demands medical attention, but no organic cause is found. The symptoms invariably occur in many different bodily systems, for instance back pains, painful menstruation, dizziness, indigestion, difficulty with vision, and partial paralysis; and the symptoms may follow fashions in health concerns among the public.

Physical pain caused by psychological conflict

The condition is relatively common and occurs in about 1 percent of adult women. It is very unusual to see this disorder in males. There are no clear etiological factors. Treatment involves not colluding with the patient's inclination to attribute organic causes to the symptoms and insuring that physicians and surgeons do not cooperate with the patient in seeking excessive diagnostic procedures or surgical remedies for the complaints.

Conversion disorder or hysterical neurosis, conversion type. This disorder was traditionally labeled hysteria. Its symptoms are a loss of or alteration in physical functioning, typically the paralysis suggesting neurological disease. The physical symptoms occur in the absence of organic pathology and are instead apparently the expression of an underlying emotional conflict. The characteristic motor symptoms of hysteria include the paralysis of the voluntary muscles of an arm or leg, tremor, tics, and other disorders of movement or gait. The neurological symptoms may be widely distributed and may not conform with medical knowledge of physical nerve distribution. Blindness, deafness, loss of sensation in arms or legs, the feeling of "pins and needles," an increased sensitivity to pain in a limb, and many other symptoms have been described.

Hysterical symptoms usually occur in a setting of extreme psychological stress and appear suddenly. The course is variable, with recovery often occurring in a few days but with symptoms persisting for years or decades in chronic cases that remain untreated.

The causation of hysteria has been linked with fixations; *i.e.,* arrested stages in the individual's early psychosexual development. Freud's theory that threatening or emotionally charged thoughts are repressed out of consciousness and converted into physical symptoms is still widely accepted. The treatment of hysteria thus requires psychological rather than pharmacological methods, notably the exploration of the sufferer's underlying emotional conflicts. Hysteria (and hypochondriasis) can also be considered as different forms of "illness behaviour"; *i.e.,* the patient uses the hysterical symptoms to gain a psychological advantage in social relationships, either by gathering sympathy or by being relieved of burdensome or stressful obligations and withdrawing from emotionally disturbing or threatening situations. Thus it may be advantageous to the patient, in a psychological sense, to have the consequences of the symptoms.

Hypochondriasis or hypochondriacal neurosis. Hypochondriasis is a preoccupation with physical signs or symptoms that the patient unrealistically interprets as abnormal, leading to the fear or belief that he is seriously ill. There may be fears about the development of physical or mental symptoms without any such existing, a belief that actual but minor symptoms are of dire consequence, or an experience of normal bodily sensations as threatening symptoms. A thorough physical examination may find no organic cause for the physical signs the patient is concerned about, but the examination fails to relieve his unrealistic fears about having a serious disease. The symptoms of hypochondriasis may occur with mental illnesses other than neuroses, for instance, depression or schizophrenia.

Hypochondriacal neurosis occurs in both sexes. The onset may be associated with precipitating factors such as an actual organic disease with physical and psychological aftereffects; *e.g.,* coronary thrombosis in a previously fit man. It often begins during the fourth and fifth decades of life but is also common at other times, during pregnancy, for example. Treatment aims to provide understanding and support and to reinforce healthy behaviour; antidepressant drugs may be used when there are depressive symptoms.

Other somatoform disorders. In psychogenic pain disorder the main feature is the persistent complaint of pain in the absence of organic disease and with evidence of a psychological cause. The pattern of pain may not conform to the known anatomic distribution of the nervous system. Psychogenic pain may occur as part of hypochondriasis or as a symptom of a depressive disorder. Appropriate treatment depends on the context of the symptom.

These somatoform disorders may occur together in one patient. Alternatively, they may occur in atypical form or in association with another physical or mental illness.

Dissociative disorders or hysterical neurosis, dissociative type. Dissociation is a syndrome in which one or a group of mental processes are split off, or dissociated, from the rest of the psychic apparatus so that their function is lost, altered, or impaired. Dissociative symptoms have often been regarded as the mental counterparts of the physical symptoms displayed in conversion disorders. Since the dissociation may be an unconscious mental attempt to protect the individual from threatening impulses or emotions that are repressed, the conversion into physical symptoms and the dissociation of mental processes can be seen as related defense mechanisms arising in response to emotional conflict.

In dissociative disorders there is a sudden, temporary alteration in the person's consciousness, sense of identity, or motor behaviour. There may be an apparent loss of memory of previous activities or important personal events, with amnesia for the episode itself after recovery. These are rare conditions, and it is important to exclude organic causes.

Amnesia, multiple personality, and depersonalization

In psychogenic or hysterical amnesia there is a sudden loss of memory which may appear total; the patient can remember nothing about his previous life or even his name. The amnesia may be localized to a short period of time associated with a traumatic event or it may be selective, affecting the person's recall of some, but not all, of the events during a particular time. In psychogenic fugue, the individual wanders away from his home or place of work and assumes a new identity; he cannot remember his previous identity and upon recovering cannot recall the events that occurred while he was in the fugue state. In many cases the disturbance lasts only a few hours or days and involves only limited travel. Severe stress frequently triggers this disorder.

Multiple personality is a rare and remarkable dissociative disorder in which two or more distinct and independent personalities develop in a single individual. Each of these personalities inhabits the person's conscious awareness to the exclusion of the others at particular times. This disorder frequently arises as a result of traumas suffered during childhood and is best treated by psychotherapy, which seeks to reunite the various personalities into a single, integrated one.

In depersonalization a person feels or perceives his body or self as being unreal, strange, altered in quality, or distant. This state of self-estrangement may take the form of feeling as if one is machinelike, is living in a dream, or is not in control of one's actions. Derealization, or feelings of unreality concerning objects outside one's self, often occurs at the same time. Depersonalization may occur alone in neurotic patients but is more often associated with phobic, anxiety, or depressive symptoms. It most commonly occurs in younger married women and may persist for many years. Patients find the experience of depersonalization intensely difficult to describe and often fear that people will think them insane. Organic conditions, especially temporal lobe epilepsy, must be excluded before making a diagnosis of neurosis when depersonalization occurs. As with other neurotic syndromes, it is more common to see a mixed picture with many different symptoms than depersonalization alone.

The causes of depersonalization are obscure, and there is no specific treatment for it. When the symptom arises in the context of another psychiatric condition, treatment is aimed at that illness.

Personality disorders. Personality is the characteristic way in which an individual thinks, feels, and behaves; it accounts for the ingrained behaviour patterns of the individual and allows the prediction of how he will act in particular circumstances. Personality embraces a person's moods, attitudes, and opinions, and is most clearly expressed in his interactions with other people. A personality disorder is a deeply ingrained, long-enduring, maladaptive, and inflexible pattern of thinking, feeling, and behaving that either significantly impairs an individual's social or occupational functioning or causes him subjective distress. Personality disorders are not illnesses but rather are pronounced accentuations or variations of personality in one or more of its traits.

A personality disorder may occur with another psychiatric condition or on its own, and it is particularly likely to be associated with neurotic conditions. The causes of personality disorders are obscure. There is undoubtedly a constitutional and therefore hereditary element in determining personality type. Psychological and environmental factors are also important in causation, for instance, the association of antisocial personality disorders with other features of social deviance found in some families and in members of lower socioeconomic groups.

Some generally accepted types of personality disorder are listed below. It is important to recognize that simply exhibiting the trait or even having it to an abnormal extent is not enough to constitute disorder—for that, the degree of abnormality must cause disturbance to the individual or to society.

Paranoid personality disorder. In this disorder there is a pervasive and unjustified suspiciousness and mistrust of others, whose words and actions are misinterpreted as having special significance for, and as being directed against, the individual. Sometimes such people are guarded, secretive, aggressive, quarrelsome and litigious, and excessively sensitive to the implied criticism of others.

Affective personality disorder. Three particular types of persistent mood disturbance can be described under this heading: (1) the trait of anxiety may be persistent and highly developed, so that the person encounters all new circumstances with fearful anticipation; (2) the chronic depressive personality is a gloomy pessimist who is skeptical in outlook and who may regard suffering as meritorious; and (3) the cyclothymic personality shows excessive swings of mood as a persistent lifelong trait.

Schizoid personality disorder. In this disorder there is a disinclination to mix with others, the individual appearing aloof, withdrawn, indifferent, unresponsive, and disinterested. Such a person prefers solitary to gregarious pursuits, involvement with things rather than with people, and often appears humourless or dull.

Schizotypal personality disorder. This category has been used to describe people who show various oddities or eccentricities of thought, speech, perception, or behaviour (such as bizarre fantasies or persecutory delusions) but whose symptoms are not severe enough to be labeled as schizophrenic.

Explosive personality disorder. Such people have a tendency to sudden emotional rages or tantrums that result in their physically assaulting others or impulsively attempting to commit suicide. The emotional outburst may be precipitated by a minor frustration that is disproportionate to the degree of reaction.

Anankastic, or compulsive, personality disorder. A person with this disorder shows prominent overscrupulous, perfectionistic traits that are expressed in feelings of insecurity, self-doubt, meticulous conscientiousness, indecisiveness, and rigidity of behaviour. The person is preoccupied with rules, procedures, and efficiency, is overly devoted to work and productivity, and is usually deficient in the ability to express warm or tender emotions. This disorder is more common in men and is in many ways the antithesis of antisocial personality disorder.

Histrionic personality disorder. Overly dramatic and intensely expressed behaviour, a tendency to call attention to oneself, a craving for novelty and excitement, egocentricity, highly reactive and excitable behaviour, and tendencies toward dependency and suggestibility are characteristic of this condition, which is more common in women than men.

Asthenic, dependent, or inadequate personality disorder. In this condition the person lacks the mental energy and ability to act on his own initiative and therefore passively allows others to assume responsibility for major aspects of his life.

Antisocial, asocial, sociopathic, or psychopathic personality disorder. This disorder is marked by a personal history of chronic and continuous antisocial behaviour, in which the rights of others are violated, and by poor or nonexistent job performance. It is manifested in persistent criminality, sexual promiscuity or aggressive sexual behaviour, and drug use. People with this disorder are impulsive, mendacious, irresponsible, and callous; they feel no guilt over their antisocial acts and fail to learn from their mistakes. The symptoms usually appear in adolescence. Antisocial personalities are less liable to criminal acts as they grow older, but there remains a high risk of suicide, accidental death, drug or alcohol abuse, and a tendency toward interpersonal problems.

The antisocial personality

Other categories of personality disorder. In the narcissistic personality disorder, there is a grandiose sense of self-importance and a preoccupation with fantasies of success, power, and achievement. Avoidant personalities are excessively sensitive to social rejection, humiliation, and shame, have low self-esteem, and are deeply upset by

the slightest disapproval of others; they are consequently unwilling to enter into relationships but crave affection and acceptance. Passive-aggressive personality disorder is the term applied to people who respond aggressively and negatively to demands made upon them by using such passive means as procrastination, dawdling, intentional inefficiency, or deliberate forgetfulness.

Personality traits are, by definition, virtually permanent, and so these disorders are only partially, if at all, amenable to treatment. The most effective treatment combines various types of group, behavioral, and cognitive psychotherapy. The behavioral manifestations of personality disorders often tend to diminish in their intensity in middle and old age.

Psychosexual disorders. Homosexuality and psychosexual dysfunctions such as impotence are treated in the article SEX AND SEXUALITY. The following section is concerned with disorders of gender identity and with preferences for unusual or bizarre sexual practices or objects.

Transsexualism and disorders of gender identity. In transsexualism the person feels a discrepancy between anatomical sex and the gender the person ascribes to himself. This disorder is much more common in biological males than females. The sufferer claims that he is a member of the other sex: "a female spirit trapped in a male body." He may assume the dress and behaviour and participate in activities commonly associated with the other sex and may even use hormones and surgery to achieve "restitution to my rightful appearance"; *i.e.,* to achieve the physical characteristics of the other sex. The cause of the condition is unknown. Once established, transsexualism persists for many years, perhaps for the rest of life. There is a risk of developing depression and an increased risk of suicide. Psychiatric treatment is generally supportive in type.

Paraphilias. Paraphilias, or sexual deviations, may be classified into disorders of sexual object and of the sexual act. Disorders of sexual object include the following. (1) In fetishism, inanimate objects are the repeated sexual preference and means of sexual arousal. (2) In transvestism, the recurrent wearing of clothes of the opposite sex is carried out to achieve sexual excitement. (3) In zoophilia, or bestiality, an animal is used as the repeated and preferred means of achieving sexual excitement. (4) In pedophilia, an adult has sexual fantasies about or engages in sexual acts with a prepubertal child of the same or opposite sex.

Disorders of the sexual act include the following. (1) In exhibitionism, repeated exposure of the genitals to an unsuspecting stranger is used to achieve sexual excitement. (2) In voyeurism, observing the sexual activity of others repeatedly is the preferred means of sexual arousal. (3) In sexual masochism, the individual achieves sexual excitement from being made to suffer. (4) In sexual sadism, the individual achieves sexual excitement by inflicting suffering upon another person.

There are, of course, other unusual sexual objects or acts that may be used for gratification. The causes of these conditions are generally not known. Behavioral, psychodynamic, and pharmacological methods have been used with varying efficacy to treat these disorders.

Disorders usually first evident in infancy, childhood, or adolescence. Children are usually referred to a psychiatrist or therapist because of complaints or concern over the child's behaviour or development by a parent or some other adult. Family problems, particularly difficulties in the parent–child relationship, are often an important causative factor in the symptomatic behaviour of the child. For the practice of child psychiatry, the observation of behaviour is especially important as the child may not be able to express his feelings in words. Isolated psychological symptoms are extremely common in children, but in one survey, disturbance amounting to psychiatric disorder was found to be present in 7 percent of all 10 and 11 year olds; boys were affected to twice the extent of girls.

Attention disorders. Children with these disorders show a degree of inattention and impulsiveness that is markedly inappropriate for their stage of development. Gross overactivity in children has many causes, including anxiety, conduct disorder, or the effects of living in institutions. One type of overactivity, the hyperkinetic, or hyperactive, syndrome, is characterized by extreme restlessness and by sustained and prolonged motor overactivity such as running around. Learning difficulties and antisocial behaviour may occur secondarily. This syndrome is 10 times more common in boys than in girls.

The hyperactive child

Conduct disorders. These are the most common psychiatric disorders in older children and adolescents, accounting for nearly two-thirds of disorders in those aged 10 and 11 years. Abnormal conduct more serious than ordinary childlike mischief persistently occurs; lying, disobedience, and aggression may be shown at home, and truancy, delinquency, and deterioration of work may occur at school. Vandalism, drug and alcohol abuse, and early sexual promiscuity may also occur. The most important causative factors are the family background; broken homes, unstable and rejecting families, institutional care in childhood, and a poor social environment are frequently present in such cases.

Anxiety disorders. Neurotic or emotional disorders in children are similar to the adult conditions except that they are often less clearly differentiated. In anxiety disorders of childhood, the child is fearful, timid with other children, and overdependent and clinging toward the parents. Aches and other physical symptoms, sleep disturbance, and nightmares occur. Separation from the parent or from the home environment is a major cause of this neurotic anxiety.

Eating disorders. Anorexia nervosa usually starts in late adolescence and is about 20 times more common in girls than boys. This disorder is characterized by a body weight more than 25 percent below standard, amenorrhea, a fear of loss of control of eating, and an intense desire to be thin. Though grossly thin, patients nevertheless believe themselves to be fat. They go to enormous lengths to resist eating food and to lose weight, including food avoidance, purging, self-induced vomiting, and vigorous exercise.

The condition appears to start with the patient's voluntary control of food intake in response to social pressures such as peer conformity. The disorder is exacerbated by troubled relations within the family. It is much more common in developed, wealthy societies and in girls of higher socioeconomic class. There is evidence that it has become more common in such countries since the 1960s. Patient management includes three stages: persuading the patient to accept and cooperate with treatment, achieving weight gain by medical methods of care, and helping the patient maintain weight by psychological and social therapy. Bulimia nervosa refers to episodic grossly excessive overeating binges. These may alternate with episodes of self-induced vomiting. The disorder is a variant of anorexia nervosa.

Anorexia nervosa and modern values

Disorders with physical manifestation. Stereotyped movement disorders involve the exhibition of tics in differing patterns. A tic is an involuntary, purposeless jerking movement of a group of muscles or the involuntary production of noises or words. Tics may affect the face, head, and neck or, less commonly, the limbs or trunk. Gilles de la Tourette's syndrome is typified by multiple tics and involuntary vocalization, especially the uttering of obscenities.

Other physical symptoms that are often listed among psychiatric disorders of childhood include stuttering, enuresis (the repeated involuntary voiding of urine by day or night), encopresis (the repeated voiding of feces into inappropriate places), sleepwalking, and night terror. These symptoms are not necessarily evidence of emotional disturbance or of some other mental illness. Behavioral methods of treatment may sometimes be effective.

Infantile autism. Psychotic disorders are very rare in childhood, and of these about one-half are cases of infantile autism; boys are affected three times as often as girls. Infantile autism begins in the first two years of life and is more common in the upper socioeconomic classes. The child shows an inability to make warm emotional relationships, has a severe speech and language disorder, and exhibits a desire for sameness in which he shows distress if thwarted from his stereotyped behaviour. There is some evidence to support genetic and organic factors in causation. Treatment involves management of the ab-

normal behaviour, training in life skills and occupational activities, and counseling for the family.

Other mental disorders. *Factitious disorders.* These are characterized by physical or psychological symptoms that are self-induced and are produced under voluntary control by the person; they are distinguished from hysteria, in which the physical symptoms are produced unconsciously. In factitious disorders, although the person's attempts to create or exacerbate the symptoms of an illness are voluntary, such behaviour is neurotic in that the individual is unable to refrain from it; *i.e.,* his goals, whatever they may be, are involuntarily adopted. In malingering, by contrast, the person simulates or exaggerates an illness or disability in order to obtain some kind of discernible personal gain or to avoid an unpleasant situation; *e.g.,* a prison inmate may simulate madness to obtain more comfortable living conditions. It is important to recognize factitious disorders as evidence of psychological disturbance. Treatment is of the underlying conflicts.

Disorders of impulse control. These conditions are usually associated with a disorder of personality. There is a failure to resist desires, impulses, or temptations to perform an act that is harmful to the individual or to others. The individual experiences a feeling of tension before committing the act and a feeling of release or gratification upon completing it. The behaviours involved include pathological gambling, setting fires (pyromania), and impulsive stealing (kleptomania).

Adjustment disorders. These are neurotic conditions in which there is an inappropriate reaction to an external stress occurring within three months of the stress. The symptoms may be out of proportion to the degree of stress, or they may be maladaptive in the sense that they prevent the individual from coping adequately in his social or occupational setting. These disorders are often associated with other neurotic conditions such as anxiety neurosis or minor depression. (A.C.P.S.)

Treatment of mental disorders

HISTORICAL OVERVIEW

Early history. References to mental disorders in early Egyptian, Indian, Greek, and Roman writings show that the physicians and philosophers who contemplated problems of human behaviour regarded mental illnesses as a reflection of the displeasure of the gods or as evidence of demoniac possession. Only a few realized that sufferers from mental illnesses should be treated humanely rather than exorcised, punished, or banished. Certain Greek medical writers, however, notably Hippocrates (flourished 400 BC), regarded mental disorders as diseases to be understood in terms of disturbed physiology. Hippocrates and his followers emphasized natural causes, clinical observation, and brain pathology in the study of mental disorders. Later Greek medical writers, including those who practiced in Imperial Rome, set out treatment programs for mental illness, including quiet, occupation, and the use of drugs such as the purgative hellebore. It is probable that most psychotic people during ancient times were cared for by their families and that those who were thought to be dangerous to themselves or others were detained at home by relatives or hired keepers.

During the early Middle Ages in Europe, primitive thinking about mental illness reemerged, and witchcraft and demonology were used to account for the symptoms and behaviour of psychotic people. At least some of the insane were looked after by the religious orders, who offered care for the sick generally. The empirical and quasi-scientific Greek tradition in medicine was maintained not by the Europeans but by the Muslim Arabs, who are usually credited with the establishment of asylums for the mentally ill in the Middle East as early as the 8th century. In medieval Europe in general it seems that the madman was allowed his liberty, provided he was not regarded as dangerous. The founding of the first hospital in Europe devoted entirely to the care of the insane occurred in Granada, in Spain, in 1365–67.

From the 17th century onward in Europe there was a growing tendency to isolate deviant people, including the insane, from the rest of society. Thus, such socially unwanted people as the mentally ill were confined together with the handicapped, vagrants, and delinquents. Those of the insane who were regarded as violent were often chained to the wall and were treated in a barbarous and inhumane way.

Brutal treatment of mentally ill

In the 17th and 18th centuries the development of European medicine and the rise of empirical methods of medicoscientific inquiry were paralleled by an improvement in public attitudes toward the mentally ill, which only began to emerge toward the end of that period. By the end of the 18th century, however, concern over the care of the insane had become so great among educated people in Europe and North America that governments were forced to act. After the French Revolution the physician Philippe Pinel was placed in charge of the Bicêtre, the hospital for the mentally ill in Paris. Under Pinel's supervision a completely new approach to the handling of mental patients was introduced. Chains and shackles were removed from the patients, and in place of dungeons they were provided with sunny rooms and were permitted to exercise on the hospital grounds. Among other reformers were the British Quaker layman William Tuke, who established the York Retreat for the humane care of the mentally ill in 1796, and the physician Vincenzo Chiarugi, who published a humanitarian regime for his hospital in Florence in 1788. In the mid-19th century Dorothea Dix carried on a campaign to arouse the public to the inhumane conditions that prevailed in American mental hospitals, and her efforts led to widespread reforms both in the United States and elsewhere.

The mental hospital era. Many hospitals for the insane were built in the latter half of the 18th century. Some of them, like the York Retreat in England, were run on humane and enlightened lines, while others, like the York Asylum, gave rise to great scandal because of their brutal methods and filthy living conditions. In the mid-19th century an extensive program of mental hospital building was carried out in North America, Britain, and many of the countries of continental Europe. The hospitals housed the insane poor, and their aim was to care for patients humanely and to relieve their families of the burden of caring for them. The approach was that of moral treatment, including occupation, the avoidance of physical methods of restraint, and respect for the individual patient. A widespread belief in the curability of mental illness at this time was a principal motivating factor behind such reform.

The mental hospital era was an age of reform, and there is no doubt that patients were treated much more humanely. The era produced a large number of segregated institutions in which a much higher proportion of the mentally ill were confined than previously. But the medical reformers' early hopes of successful cures were not vindicated, and by the end of the 19th century the hospitals had become overcrowded, and custodial care had replaced moral treatment.

The biological movement. Along with humanitarian reforms in hospital practice and treatment methods during the late 18th and 19th centuries, there was a resurgence of medical and scientific interest in psychiatric theory and practice. Fundamental strides were made during this period in establishing a scientific basis for the study of mental disorders. A long series of observations by clinicians in France, Germany, and England culminated in 1883 in a comprehensive classification of mental disorders by the German psychiatrist Emil Kraepelin. His classification system served as the basis for all subsequent ones, and the cardinal distinction he made between schizophrenia and manic-depressive psychosis still stands.

Rapid advances in various branches of medicine led in the later 19th century to the expectation of discovering specific brain lesions that were thought to cause the various forms of mental disorder. While these researches did not attain the results that were expected, the scientific emphasis was productive in that it did elucidate the gross and microscopic pathology of many brain disorders that can produce psychiatric disabilities. Nevertheless, certain of the psychotic disorders, notably schizophrenia and manic-

depressive psychosis, frustrated the effort to find causative agents in cellular pathology. Other explanations were needed for the many puzzling aspects of mental disorders in general, and these explanations emerged in a wave of psychological rather than physical explanations.

Development of psychotherapy. Foremost among these was that of psychoanalysis, which originated in the work of the Viennese neurologist Sigmund Freud. Having studied under the French neurologist Jean-Martin Charcot, Freud originally used well-known techniques of hypnosis to treat patients suffering from hysterical paralysis and other neurotic syndromes. Freud and his colleague, Josef Breuer, observed that their patients tended to relive earlier life experiences that could be associated with the symptomatic expression of their illnesses. When these memories and the emotions associated with them were brought to consciousness during the hypnotic state, the patients showed improvement. Observing that most of his patients proved able to talk about such memories without being under hypnosis, Freud evolved the technique of free association (the production by the patient, aloud and without suppression or self-censorship of any kind, of the thoughts and feelings about whatever was uppermost in his mind) as a means of access to the unconscious. From this beginning Freud gradually developed what became known as psychoanalysis. Other features of the new procedure included the study of dreams, the interpretation of "resistances" on the part of the patient, and the handling by the therapist of transference (the patient's feelings toward the analyst that reflected previously experienced feelings toward parents and other important figures in his early life). Freud's work, though complex and controversial in many of its aspects, laid the basis for modern psychotherapy in its use of free association and its emphasis on unconscious and irrational mental processes as causative factors in mental illness. This emphasis on purely psychological factors as a basis for both causation and treatment was to become the cornerstone of most subsequent psychotherapies.

Freud's discoveries

Variations of the original psychoanalytic technique were introduced by several of Freud's colleagues who parted company with him. Analytic psychology, devised by Carl Jung, placed less emphasis on free association and more on the interpretation of dreams and fantasies. Special importance was given to the collective unconscious, a reservoir of shared unconscious wisdom and ancestral experience that entered consciousness only in symbolic form to influence thought and behaviour. Jungian analysts sought clues to their patients' problems in the archetypal nature of myths, stories, and dreams. Individual psychology, devised by Alfred Adler, emphasized the importance of the individual's drive toward power and of his unconscious feelings of inferiority. The therapist was concerned with the patient's compensations for his inferiority, as well as with his social relationships.

Development of physical and pharmacological treatments. During the early decades of the 20th century the principal approaches to the treatment of mental disorders were psychoanalytically derived psychotherapies, used to treat people with neuroses, and custodial care in mental hospitals, for those with psychoses. But beginning in the 1930s these methods began to be supplemented by physical approaches using drugs, electroconvulsive therapy, and surgery. The first successful physical treatment in psychiatry was the induction of malaria in patients with a fatal form of neurosyphilis called general paresis. The malarial treatment stemmed from the observation that some psychotic patients improved during febrile illnesses. In 1933 the Polish psychiatrist Manfred Sakel reported the treatment of schizophrenia by repeated insulin-induced comas. (Neither of these physical treatments is in use today.) The treatment of schizophrenia by convulsions, originally induced by the injection of camphor, was reported in 1935 by the psychiatrist Ladislaus Joseph von Meduna in Budapest. An improvement in this approach was the induction of convulsions by the passage of an electrical current through the brain, a technique introduced by the Italian psychiatrists Ugo Cerletti and Lucio Bini in 1938. It soon became apparent that electroconvulsive treatment was more successful in alleviating states of severe depression than in treating schizophrenia. Psychosurgery, or surgery performed to treat mental illness, was introduced by the Portuguese neurologist António Egas Moniz in the 1930s. The operation Moniz originated, leucotomy, or lobotomy, was widely performed during the next two decades in the treatment of schizophrenia, intractable depression, severe anxiety, and severe obsessional states. The procedure was later abandoned, however, largely because its therapeutic effects could be better obtained by the use of newly developed drugs.

Accidental discoveries of pharmacotherapeutic agents

The decades after World War II were marked by the first safe and effective applications of drugs in the treatment of mental disorders. Prior to the 1950s such sedative compounds as bromides and barbiturates had been used to quiet or sedate patients, but these drugs were general in their effect and did not target the specific symptoms of mood disturbances or psychotic disorders. Many of the drugs that subsequently proved effective in treating such conditions were recognized serendipitously; *i.e.*, when researchers administered them to patients just to see what would happen, or when they were administered to treat one mental condition and were instead found to be helpful in alleviating the symptoms of an entirely different condition.

The first effective pharmacological treatment of psychosis was the treatment of mania with lithium, introduced by the Australian psychiatrist J.F.J. Cade in 1949. Lithium, however, excited little interest until its dramatic effectiveness in the maintenance treatment of bipolar affective disorder was reported in the mid-1960s. Chlorpromazine, the first of a long series of highly successful antipsychotic drugs, was synthesized in France in 1950 during work on antihistamines. It was used in anesthesia before its antipsychotic and tranquilizing effects were reported in France in 1952. The first tricyclic antidepressant drug, imipramine, was originally designed as an antipsychotic drug and was investigated by the Swiss psychiatrist Roland Kuhn. He found it ineffective in treating schizophrenia but observed its antidepressant effect, which he reported in 1957. A drug used in the treatment of tuberculosis, iproniazid, was found to be effective as an antidepressant in the mid-1950s. It was the first monoamine oxidase inhibitor to be used in psychiatry. The first modern anxiety-relieving drug was meprobamate, which was originally introduced as a muscle relaxant. It was soon overtaken by the pharmacologically rather similar but clinically more effective chlordiazepoxide, which was synthesized in 1957 and marketed as Librium in 1960. This drug was the first of the extensively used benzodiazepines. These and other drugs had a revolutionary impact not only on psychiatry's ability to relieve the symptoms and suffering of people with a wide range of mental disorders but also on the institutional care of the mentally ill.

Deinstitutionalization. Between about 1850 and 1950 there was a steady increase in the number of patients staying in mental hospitals. In England and Wales, for example, there were just over 7,000 such patients in 1850, nearly 120,000 in 1930, and nearly 150,000 in 1954. Then a steady decline in the number of occupied beds began, reaching just over 100,000 in 1970 and 75,000 in 1980, a decrease of almost 50 percent. The same process began in the United States in 1955 but continued at a more rapid rate. The decrease, from just under 560,000 in 1955 to just over 130,000 in 1980, was one of more than 75 percent. In both countries it became official policy to replace mental hospital treatment with community care, involving district general hospital psychiatric units in Britain and local mental health centres in the United States. This dramatic change can be partly attributed to the introduction of antipsychotic drugs, which quieted many psychotic patients and drastically changed the atmosphere of mental hospital wards for the better. With the recovery of lucidity and calmness, many psychotic patients could return to their homes and live at least a partially normal existence instead of spending their lives sequestered in mental hospitals. The wholesale release of mental patients into the community was not without problems, however, since many areas lacked the facilities to support and maintain such patients, many of whom thus received inadequate care.

Development of behaviour therapy. In the 1950s and '60s a new type of therapy, called behaviour therapy, was developed. In contrast to the existing psychotherapies, its techniques were based on theories of learning derived from research on classical conditioning carried out by Ivan Pavlov and others, and from the work of such American behaviourists as John B. Watson and B.F. Skinner. Behavioral therapy developed when the theoretical principles that were originally developed from experiments with animals were applied to the treatment of patients.

In 1920 Watson experimentally induced a phobia of rats in a small boy, and in 1924 Mary Cover Jones reported the extinction of phobias in children by gradual desensitization. Modern behaviour therapy began with the description by the South African psychiatrist Joseph Wolpe of his technique of systematically desensitizing a patient with phobias, beginning by exposing him to the least and gradually progressing to the most feared object or situation. Behavioral therapies were more quickly adopted in Europe than in the United States, where psychoanalytic precepts had exercised a particular dominance over psychiatry, but by the 1980s behavioral therapies were also well established in the United States.

The mental health profession in the late 20th century. There has been a great increase in the number of mental health professionals since World War II. In the United States the number of psychiatrists was 3,000 in 1939 but had increased to more than 25,000 by the early 1970s. Nonmedical mental health professionals have also increased in number and have achieved increasing independence from medical control, acquiring new roles in the process. Clinical psychologists, for instance, who were at one time largely confined to carrying out psychometric tests at the doctor's request, have become increasingly concerned with psychotherapy and behaviour therapy. Psychiatric social workers are no longer confined to casework with individual patients or their relatives but have also become psychotherapists and play prominent roles in mental health centres. There are new roles for nurses, including behaviour therapy and the management of chronic mental illness in the community. The greatest beneficiaries of this expansion of mental health professionals have been patients with neurotic and other less severe disorders.

Psychotherapy retains a major role in the mental health profession, and since the development of psychoanalysis the varieties of psychotherapy have increased and multiplied to the extent that a 1980 handbook listed "more than 250 different therapies in use today." The repertoire of drugs used in the treatment of mental illness has continued to grow as new drugs are developed or new applications of existing ones are discovered, and research on the biochemical and genetic causes of mental disease continues to make gradual headway in explicating the causes of various disorders. The triad of psychotherapy, drugs, and behaviour therapy afford an unprecedented array of approaches, techniques, and procedures for alleviating symptoms or curing people altogether of mental disorders.

PHYSIOLOGICAL TREATMENTS

Pharmacological treatments. *Antipsychotic agents.* Antipsychotic drugs, which are also known as neuroleptics and major tranquilizers, belong to several different chemical groups but are similar in their therapeutic effects. They have a calming effect that is valuable in the relief of agitation, excitement, and violent behaviour in psychotic patients. The drugs are quite successful in reducing the symptoms of schizophrenia, mania, and delirium, and they are used in combination with antidepressants to treat psychotic depression. The drugs suppress hallucinations and delusions, alleviate disordered or disorganized thinking, improve the patient's lucidity, and generally make him more receptive to psychotherapy. Patients who have previously been agitated, intractable, or grossly delusional become noticeably calmer, quieter, and more rational when maintained on these drugs. The drugs have enabled many episodically psychotic patients to have shorter stays in hospitals, and many other patients who would have been permanently confined to institutions are able to live in the outside world because of the drugs. The antipsychotics differ in their unwanted effects: some are more likely to make the patient drowsy; some to alter blood pressure or heart rate; and some to cause tremor or slowness of movement. Some of the most widely used antipsychotic drugs are shown in Table 1.

Revolutionary effects of antipsychotic drugs

Table 1: Examples of Antipsychotic Drugs

chemical group	drug
phenothiazines	chlorpromazine
	trifluoperazine
	fluphenazine decanoate
thioxanthenes	flupentixol decanoate
butyrophenones	haloperidol
diphenylbutylpiperidines	pimozide
	fluspirilene
substituted benzamides	sulpiride

In the treatment of schizophrenia, antipsychotic drugs partially or completely control such symptoms as delusions and hallucinations. They also protect the patient who has recovered from an acute episode from suffering a relapse. Unfortunately, the drugs have less of an effect on such symptoms as social withdrawal, apathy, blunted emotional capacity, and the other psychological deficits characteristic of the chronic stage of the illness.

No single drug seems to be outstanding in the treatment of schizophrenia. In an individual patient, one drug may be preferred to another because it produces less severe unwanted effects, and the dose of any one drug needed to produce a therapeutic effect varies widely from patient to patient. Because of these individual differences it is common for psychiatrists to substitute a drug of a different chemical group when one drug has been shown to be ineffective despite its use in adequate dosage for several weeks.

In an acute psychotic episode a drug such as chlorpromazine, trifluoperazine, or haloperidol usually has a calming effect within a day or two. The control of psychotic symptoms such as hallucinations or disordered thinking may take weeks. The appropriate dosage has to be determined for each patient by cautiously increasing the dose until a therapeutic effect is achieved without unacceptable side effects.

Antipsychotic drugs are thought to work by blocking dopamine receptors in the brain. Dopamine is a neurotransmitter; *i.e.,* a chemical messenger produced by certain nerve cells that influence the function of other nerve cells by interacting with receptors in their cell membranes. Since schizophrenia may be caused by either the excessive release of or an increased sensitivity to dopamine in the brain, the effects of antipsychotic drugs may be due to their ability to block or inhibit dopamine transmission.

Unwanted side effects of antipsychotic drugs

Dopamine receptor blockade is responsible for the drug's main unwanted side effects, which are termed extrapyramidal symptoms (EPS). These resemble the symptoms of Parkinson's disease and include tremor (shakiness) of the limbs; bradykinesia—slowness of movement with loss of facial expression, absence of arm-swinging during walking, and a general muscular rigidity; dystonia—sudden, sustained contraction of muscle groups causing abnormal postures; akathisia—a subjective feeling of restlessness leading to an inability to keep still; and tardive dyskinesia—involuntary movements, particularly involving the lips and tongue. Most extrapyramidal symptoms disappear when the drug is withdrawn. Tardive dyskinesia occurs late in the drug treatment and in about half of the cases persists even after the drug is no longer used. There is no satisfactory treatment.

Antianxiety agents. The drugs most commonly used in the treatment of anxiety are the benzodiazepines, or minor tranquilizers, which have replaced the barbiturates because of their vastly greater safety. The advantage of benzodiazepines is that they calm the patient without the marked sleep-inducing effects of barbiturates, so that a degree of wakeful alertness is maintained and the individual can carry on his daily activities. A large number of benzodiazepines have been marketed, and the more common ones are listed in Table 2. They differ from each other in duration of action rather than effectiveness. Smaller doses

Table 2: Commonly Used Benzodiazepines

duration of effect	drug
short-acting	lorazepam
	oxazepam
	alprazolam
long-acting	chlordiazepoxide
	diazepam
	clorazepate

have a calming effect and alleviate both the physical and psychological symptoms of anxiety. Larger doses induce sleep, and some benzodiazepines are marketed as hypnotics. The benzodiazepines have become among the most widely prescribed drugs in the developed world, and controversy has arisen over their excessive use by the public.

The side effects of these drugs are usually few—most often drowsiness and unsteadiness. The drugs themselves are not lethal even in very large overdoses, but they increase the sedative effects of alcohol and other drugs. The benzodiazepines are basically intended for short- or medium-term use, since the body develops a tolerance to them that reduces their effectiveness and necessitates the use of progressively larger doses. Dependence on them also occurs, even in moderate dosages, and withdrawal symptoms have been observed in those who have used the drugs for only four to six weeks. In patients who have taken a benzodiazepine for many months or longer, withdrawal symptoms occur in between 15 and 40 percent of the cases and may take weeks or months to subside.

The withdrawal symptoms are of three kinds. Such severe symptoms as delirium or convulsions are rare. Frequently the withdrawal symptoms represent a renewal or increase of the anxiety itself. Many patients also experience other symptoms such as hypersensitivity to noise and light, as well as muscle twitching. As a result, many long-term users continue to take the drug not because of persistent anxiety but because the withdrawal symptoms are too unpleasant.

Because of the danger of dependence, benzodiazepines should be taken in the lowest possible dose for no more than a few weeks. For longer periods they should be taken intermittently, and only when the anxiety is severe.

Benzodiazepines act on specialized receptors in the brain that are adjacent to receptors for a neurotransmitter called gamma-aminobutyric acid (GABA), which inhibits anxiety. It is possible that the interaction of benzodiazepines with these receptors facilitates the inhibitory (anxiety-suppressing) action of GABA within the brain.

Antidepressant agents. Many patients suffering from depressive illness gain symptomatic relief from treatment with one of the tricyclic drugs (so called because of their three-ringed chemical structure) or one of the newer drugs with similar properties. Some of the most widely used antidepressant drugs are shown in Table 3.

Table 3: Some Commonly Used Antidepressant Drugs

variety	drug
tricyclic	imipramine
	amitriptyline*
	nortriptyline
	clomipramine
	doxepin*
	lofepramine
newer tricyclic-like drugs	mianserin*
	trazodone*
monoamine oxidase inhibitors	phenelzine
	tranylcypromine

*Marked sedative effect.

Patients with melancholia who are not delusional benefit the most from the antidepressants. The drug is given in reduced dosage for the first few days to lessen the severity of any side effects, which are often more severe in the first days of treatment, and the drug is then given in full dosage for at least three to four weeks. A sedative effect may be apparent within the first two to three days, but it usually takes two to three weeks for the drug to significantly improve the patient's depressed mood. If there is no improvement after four weeks of adequate dosage, the drug should be discontinued over the course of a few days. There is no convincing evidence that any one tricylic antidepressant is superior to the others in therapeutic effectiveness. If a sedative antidepressant produces troublesome daytime drowsiness, it should be replaced by a less sedative drug. If the side effects of a classical tricyclic are severe, it can be replaced by lofepramine, a newer drug with fewer unwanted effects, or with one of the tricyclic-like drugs.

The spectrum of antidepressants

Successful treatment with such drugs relieves all the symptoms of depressive illness, including disturbances of sleep and appetite, loss of sexual desire, and decreased energy, interest, and concentration. Once a good response has been achieved, the drug should be continued in reduced dosage for a further six months. Research has shown that such maintenance treatment greatly reduces the risk of relapse during that time.

It is widely theorized that depression is partly caused by reduced quantities or reduced activity of the monoamine neurotransmitters serotonin and norepinephrine. Tricyclics are thought to act by inhibiting the body's physiological inactivation of the monoamine neurotransmitters in the brain. This results in the buildup or accumulation of the neurotransmitters there and allows them to remain in contact longer with their receptors, changes that may be important in elevating mood.

The side effects of these drugs are mostly due to their interference with the function of the autonomic nervous system. The side effects include dryness of the mouth, blurred vision, constipation, and, particularly in older men, difficulty in passing urine. Weight gain can be a distressing side effect in patients taking a tricyclic for a long period. Such patients often report an increase in appetite for carbohydrate-rich foods. In elderly patients, in whom lower doses should be used, antidepressants can cause delirium. The classic tricyclics interfere with conduction in heart muscle, and so they are best avoided in patients with heart disease. They are also dangerous if a patient takes a large overdose of them. Drug interactions occur with tricyclics, the most important being their interference with the action of certain drugs used in the treatment of high blood pressure.

Monoamine oxidase inhibitors (MAOIs), of which phenelzine is the best researched, are generally ineffective in patients with melancholia or depressive delusions but are more effective in patients whose depressed mood can be temporarily lightened by a change in the environment. They are used in the same way as tricyclics, with a low initial dose increased after a few days. Their antidepressant effect is also delayed, and treatment needs to continue for four to six weeks before its effectiveness can be assessed.

As their name implies, the drugs interfere with the action of monoamine oxidase, an enzyme involved in the breakdown of norepinephrine and serotonin. As a result, these neurotransmitters accumulate within nerve cells and presumably leak out onto receptors.

The side effects of these drugs include daytime drowsiness, difficulty in getting to sleep, and a fall in blood pressure when rising to one's feet. The MAOIs interact dangerously with various other drugs, including narcotics and some over-the-counter drugs used in treating colds. Patients taking an MAOI must avoid certain foods containing tyramine or other naturally occurring amines, which can cause a severe rise in blood pressure leading to headaches and even to intracranial bleeding. Tyramine occurs in several foodstuffs, of which the most important are cheese, Chianti wine, and well-cured meats. The MAOI drugs are safe in normal dosages but are dangerous in overdosage.

Certain antidepressants seem to be effective in treating other mental disorders. This is true of imipramine in some cases of panic disorder, monoamine oxidase inhibitors in agoraphobia, and clomipramine in agoraphobia and in some cases of obsessive-compulsive disorder.

Mood-stabilizing drugs. Lithium, usually administered as its carbonate in several small doses per day, is effective

in the treatment of an episode of mania. It can drastically reduce the elation, overexcitement, grandiosity, paranoia, irritability, and flights of ideas typical of people in the manic state. It has little or no effect for several days, however, and a therapeutic dose is rather close to a toxic dose. In severe episodes haloperidol or chlorpromazine may also be used. Lithium also has an antidepressant action in some patients with melancholia.

Importance of lithium treatment

The most important use of lithium is in the maintenance treatment of patients with bipolar affective disorder (manic-depressive illness) or with recurrent depression. When given while the patient is well, lithium may prevent further mood swings, or it may reduce either their frequency or their severity. Its mode of action is unknown. Treatment begins with a small dose that is gradually increased until a specified concentration of lithium in the blood is reached. Blood tests to determine this are carried out weekly in the early stages of treatment and later every two to three months. It may take as long as a year for lithium to become fully effective.

The toxic effects of lithium, which usually occur when there are high concentrations of it in the blood, include drowsiness, coarse tremors, vomiting, diarrhea, incoordination of movement, and, with still higher blood concentrations, convulsions, coma, and death. At therapeutic blood concentrations, lithium's side effects include fine tremors (which can be alleviated by propranolol), weight gain, passing increased amounts of urine with consequent increased thirst, and reduced thyroid function.

Carbamazepine, an anticonvulsant drug with a chemical composition similar to that of the tricyclic antidepressants, has been shown to be effective in the treatment of mania and in the maintenance treatment of bipolar affective disorder. It may be combined with lithium in bipolar affective patients who fail to respond to either drug alone.

Electroconvulsive treatment. In electroconvulsive treatment (ECT) a convulsion is produced in a person by passing an electric current through his brain. The duration of the convulsive activity in the brain appears to determine its therapeutic effects, while the intensity of the electrical stimulus plays a role in determining its unwanted side effects, particularly the short-term memory impairment in the patient immediately after treatment. Several controlled trials have shown that ECT is effective in treating patients suffering from a depressive illness with melancholia.

Prior to the administration of ECT, the patient is given an intravenous injection of a short-acting anaesthetic to put him to sleep and then is given an injection of a muscle-relaxant in order to reduce the force of his muscular contractions during the convulsion. The electrical current is then applied to the brain. In bilateral ECT this is done by applying an electrode to each side of the head; in unilateral ECT both electrodes are placed over the nondominant cerebral hemisphere—*i.e.,* the right side of the head in a right-handed person. Unilateral ECT produces noticeably less confusion and memory impairment in patients, but more treatments may be needed. Patients recover consciousness rapidly after the treatment but may be confused and may experience a mild headache for an hour or two.

ECT treatments are normally given two or three times a week in the treatment of depressed patients. Once a program of ECT has been successfully completed, maintenance treatment with a tricyclic such as imipramine for the next few months significantly decreases the risk of relapse.

Electroconvulsive treatment as an alternative therapy

Whenever rapid improvement is important, ECT is preferred to the treatment of depression with drugs. This is so in cases of severe depression when the patient's life is endangered because of refusal of food and fluids or because of serious risk of suicide, as well as in cases of depression with psychosis after childbirth, when it is desirable to reunite the mother and baby as soon as possible. ECT is also used in treating patients with melancholia who have failed to respond to adequate dosages of antidepressants, and to treat elderly depressed patients who are unable to tolerate antidepressant drugs.

The number of electroconvulsive treatments required to treat depression is usually between four and eight, with more needed by some elderly patients. Some patients improve after the first treatment, others only after several. The mode of action of ECT is not understood. It has been shown that electrically induced convulsions in animals alter the sensitivity of norepinephrine and serotonin receptors in the brain. Conceivably this action could alter depressed mood in human patients.

The chief unwanted effect of ECT is impairment of memory. Some patients report memory gaps covering the period just before treatment, others during several years before. Many patients have memory difficulties for a few days or even a few weeks after completion of the treatment so that they forget appointments, phone numbers, and the like. These difficulties are transient and disappear rapidly in the vast majority of patients. Occasionally, however, patients complain of permanent memory impairment after ECT; these are almost always patients who did not recover from their depression as a result of the treatment.

Psychosurgery. Psychosurgery is the destruction of groups of nerve cells or nerve fibres in the brain by surgical techniques in an attempt to relieve psychiatric symptoms that are not due to structural brain disease. The removal of a brain tumour that is causing psychiatric symptoms is not an example of psychosurgery.

The classical technique of bilateral prefrontal leucotomy (lobotomy) is no longer performed because of its frequent undesirable effects on physical and mental health, in particular the development of epilepsy and the appearance of permanent, undesirable changes in personality. The latter included increased apathy and passivity, lack of initiative, and a generally decreased depth and intensity of the person's emotional responses to life. In its heyday the operation was used to quiet chronically tense, delusional, agitated, or violent psychotic patients. Stereotaxic surgical techniques have been developed that enable the surgeon to produce small areas of nerve cell or fibre destruction by means of metal probes inserted into accurately located parts of the brain. The nerve tissue is then destroyed by the implantation of a radioactive substance (usually yttrium) or by the application of heat or cold.

The proponents of psychosurgery claim that it is effective in treating some patients with severe and intractable depression and with anxiety or obsessional neuroses and that it may improve the behaviour of abnormally aggressive patients. There is no compelling evidence to support these claims, and many of the therapeutic effects that were claimed for psychosurgery by its adherents are in fact now attainable by the use of antipsychotic and antidepressant drugs. Many physicians believe that psychosurgery is never justified. Others accept that it has a very small part to play in psychiatric treatment when the prolonged use of other forms of treatment has been unsuccessful and the patient is chronically and severely distressed or tormented by his symptoms. Whereas ECT is a routine treatment in certain specified conditions, psychosurgery is at best a last resort.

THE PSYCHOTHERAPIES

Psychotherapy implies the treatment of mental discomfort, dysfunction, or disease by psychological means by a trained therapist who adheres to a particular theory of both symptom causation and symptom relief. The American psychotherapist Jerome Frank has classified psychotherapies into religio-magical and empirico-scientific forms. The former depend on the shared beliefs of the therapist and client in magic, spirits, or other supernatural processes or powers. This article is concerned, however, with the latter forms of psychotherapy, which have been developed by modern medicine and which are carried out by a member of one of the mental health professions such as a psychiatrist or a clinical psychologist.

It is usual to contrast two main forms of psychotherapy, dynamic and behavioral. They are conceptually different; behaviour therapy concentrates on alleviating a patient's overt symptoms, which are attributed to faulty learning, while dynamic therapy concentrates on understanding the meaning of symptoms and understanding the emotional conflicts within the patient that may be causing those symptoms. In their pure forms the two approaches are very different, but in practice many therapists use elements of both.

Dynamic psychotherapies. There are many variants of dynamic psychotherapy, all of which ultimately derive from the basic precepts of psychoanalysis. The fundamental approach of most dynamic psychotherapies can be traced to three basic theoretical principles or assertions: (1) Human behaviour is prompted chiefly by emotional considerations, but insight and self-understanding are necessary to modify and control such behaviour and its underlying aims; (2) A significant proportion of human emotion is not normally accessible to one's personal awareness or introspection, being rooted in the unconscious, those portions of the mind beneath the level of consciousness; (3) Any process that makes available to a person's conscious awareness the true significance of emotional conflicts and tensions that were hitherto held in the unconscious will thereby produce heightened awareness and increased stability and emotional control. The classic dynamic psychotherapies are relatively intensive talking treatments that are aimed at providing the person with insight into his conscious and unconscious mental processes, with the goal of enabling him ultimately to achieve a better understanding of himself. Dynamic psychotherapy attempts to enhance personality growth as well as to remove symptoms. The main therapeutic forces are activated in the relationship between patient and therapist and depend both upon the empathy, understanding, integrity, and concern demonstrated by the therapist and upon the motivation, intelligence, and capacity for achieving insight manifested by the patient. The attainment of a therapeutic alliance—*i.e.*, a working relationship between patient and therapist that is based on mutual respect, trust, and confidence—provides the context in which the patient's problems can be worked through and resolved. Several of the most important forms are treated below.

The role of unconscious mental processes

Psychoanalytic psychotherapy. Classical psychoanalysis is the most demanding of all the psychotherapies in terms of time, cost, and effort. It is conducted with the patient lying on a couch and with the analyst seated out of his sight but close enough to hear what the patient says. The treatment sessions last 50 minutes and are usually held four or five times a week for at least three years. The primary technique used in psychoanalysis and in other dynamic psychotherapies to enable unconscious material to enter the patient's consciousness is that of "free association." In free association, according to Freud, the patient "is to tell us not only what he can say intentionally and willingly, what will give him relief like a confession, but everything else as well that his self-observation yields him, everything that comes into his head, even if it is disagreeable for him to say it, even if it seems to him unimportant or actually nonsensical." Such a procedure is rendered difficult, first, because for a person to speak of his innermost (and often socially unacceptable) thoughts is a departure from years of practice in which he has selected what he has said to others. Free association is also difficult because the patient resists remembering repressed experiences or feelings that are connected with intense or conflicting emotions that have never been finally resolved or settled. Such repressed emotions or memories usually revolve around the patient's important personal relationships and his innermost feelings of self, and the release or recollection of such emotions in the course of treatment can be itself intensely disturbing.

The role of free association

Attentive listening and "empathy" on the part of the therapist allows the patient to express thoughts and feelings that in turn permit the uncovering of his underlying emotional conflicts. In the course of treatment, the patient often seeks to project (attribute to something other than himself) the disturbing emotions he feels in the process of recollection and free association, and the person who is almost invariably selected for the focus of such projection is the psychoanalyst; that is, the patient is likely to blame his emotional distress on the analyst. In this way, the patient comes to feel love or hatred, dependence or rebellion, and rivalry or rejection toward the analyst. These are the same attitudes the patient has felt but has never consciously acknowledged toward his parents or other people with whom he shared important relations earlier in life. The patient's projection onto the therapist of these feelings and behaviours that originated in his earlier relationships is called the transference. To facilitate the development of the transference, the analyst endeavours to maintain a neutral stance toward the patient in order to serve as a "blank screen" onto which the patient can project his inner feelings. The analyst's handling of the transference situation is of vital importance in psychoanalysis or, indeed, in any form of dynamic psychotherapy. It is through the transference that the patient discovers the nature of his unconscious feelings and then becomes able to acknowledge them. Once this has been done, he often finds himself able to regard them in a far more dispassionate and tolerant light and often feels himself liberated from their influence upon his future behaviour.

A major therapeutic tool in the course of treatment is interpretation. This technique helps the patient to become aware of previously repressed aspects of his emotional conflicts and to uncover the meaning of uncomfortable feelings evoked by the transference. Interpretation, in turn, is used to determine the underlying psychological meaning of the patient's dreams, which are held to have a hidden or latent content that symbolize and indirectly express aspects of the patient's emotional conflicts.

Individual dynamic psychotherapy. Although psychoanalysis has had a profound influence, particularly on American psychiatry, that influence waned significantly during the 1970s and '80s. Fewer patients now enter psychoanalysis, and many analysts carry out short-term individual dynamic psychotherapy. This form of therapy is much more readily available and usually requires 50 minutes a week for six to 18 months. The aim of treatment, as in psychoanalysis, is to increase the patient's insight (understanding of himself), to relieve his symptoms, and to improve his psychological functioning. Suitable patients include those with a wide range of neurotic disorders and personal or social problems who wish to change and who are able to view their problems in psychological terms.

As in psychoanalysis, the patient learns to trust the therapist and becomes able to talk candidly and honestly about his most intimate thoughts and feelings. The treatment setting is less formal than in psychoanalysis, with the therapist and patient seated so that eye contact can be achieved if desired.

Treatment techniques include free association and the use of interpretation by the therapist to analyze the transference, the patient's unconscious defense mechanisms, and his dreams. The therapist may ask the patient to clarify or enlarge on some point on which the therapist is not clear if this seems important in the development of the patient's symptoms. The therapist directs the patient's attention to important links, of which he seems unaware, between the present and the past, between his emotional responses to the therapist and to people important to him, and so on. The therapist may challenge the patient with the likely consequences of his resistant or maladaptive behaviour and stress instead the importance of confronting and trying to resolve his psychological difficulties.

Brief focal psychotherapy. This is a form of short-term dynamic therapy in which a time limit to the duration of the therapy is often agreed upon with the patient at the outset. Sessions lasting 30 to 60 minutes are held weekly for, typically, five to 15 weeks. At the beginning of treatment the therapist helps identify the patient's problem or problems, and these are made the focus of the treatment. The problem should be an important source of distress to the patient and it should be modifiable within the time limit. The therapist is more active, directive, and confrontational than in long-term dynamic therapy and ensures that the patient keeps to the focus of treatment and is not diverted by subsidiary problems or concerns. Some therapists deliberately aim to produce considerable emotional arousal in the patient during each session as a way to activate or highlight specific problems. Research suggests that brief therapy can produce as good results as long-term therapy, and more quickly.

Psychotherapy over the short-term

Group psychotherapy. Many types of psychological treatment may be provided for groups of patients with psychiatric disorders. This is true, for example, of relaxation training and anxiety-management training. There

are also self-help groups, of which Alcoholics Anonymous is perhaps the best known. A considerable number of group experiences have been devised for people who are not suffering from any psychiatric disorder; encounter groups are a well-known example. This discussion, however, is concerned with long-term dynamic group therapy, in which six to 10 psychiatric patients meet with a trained group therapist, or sometimes two therapists, for 60 to 90 minutes a week for up to 18 months. Often the group is closed, *i.e.,* confined to the original group membership, even if one or more members drop out before the treatment ends. In an open group patients who have stopped attending, whether by default or because of the relief of symptoms, are replaced by new members.

The types of mental disorders considered suitable for group therapy are much the same as those suitable for individual therapy. Again the patient must want to change and must be psychologically minded. In addition, it is important that he not consider group therapy as a poor second to individual therapy.

There are many varieties of dynamic group therapy, and they differ in their theoretical background and technique. The influential model of the American psychiatrist Irvin D. Yalom provides a good example of such therapies, however. The therapist continually encourages the patients to direct their attention to the personal interactions occurring within the group rather than to what happened in the past to individual members or what is currently happening outside the group, although both of these areas may be considered when they are relevant. The therapist regularly draws attention to what is happening among members of the group as they learn more about themselves and test out different ways of behaving with one another. The goal in group therapy is to create a climate in which the participants can shed their inhibitions. When the members come to trust one another, they are able to provide feedback and to respond to other group members in ways that might not be possible in ordinary social interactions owing to the constraints of social conventions.

Several factors appear to be important in group therapy. The most important is group cohesion, which gives the patient a feeling of belonging, identification, and security and enables him to be frank and open and to take risks without the danger of rejection. Universality refers to the patient's realization that he is not unique, that all the other group members have problems, some of them similar to his. Optimism about what can be achieved in the group, fostered by the perception of change in others, combats demoralization. Guidance, the giving of advice and explanation, is important in the early meetings of the group and is largely a function of the therapist. What has been called vicarious learning later becomes more important; the patient observes how other group members evolve solutions to common problems and emulates desirable qualities he sees in fellow members. Catharsis, or the release of highly charged emotion, occurs within the group setting and is helpful provided that the patient can come to understand it and appreciate its significance. Another factor that is helpful in improving self-esteem is altruism, the opportunity to give assistance to another group member.

Internal dynamics of the group

Family therapy. Family therapists view the family as the "patient" or "client" and as more than the sum of its members. The family as a focus for treatment usually comprises the members who live under the same roof, sometimes supplemented by relatives who live elsewhere or by nonrelatives who share the family home. Therapy with couples—marital therapy—may be considered as a special type of family therapy. Family therapy may be appropriate when the person referred for treatment has symptoms clearly related to such disturbances in family function as marital discord, distorted family roles, and parent–child conflict, or when the family as a unit asks for help. It is not appropriate when the patient has a severe disorder needing specific treatment in its own right.

The many theoretical approaches include psychoanalytic, systems-theory, and behavioral models. The analyst is concerned with the family's past as the cause of the present; he pays attention to psychodynamic aspects of the individual members as well as of the family unit; and he makes numerous interpretations and aims at increasing the insight of the members. The systems therapist, on the other hand, is interested in the present rather than the past and is often not concerned with promoting insight but rather with changing the family system, perhaps by altering the implicit and fixed rules under which it functions so that it can do so more effectively. The behaviour therapist is concerned with behaviour patterns and with pinpointing the types of reinforcement that maintain behaviour that other family members regard as undesirable. Members specify the changes in behaviour that they wish to see in one another, and strategies are devised to reinforce the desired behaviours. This approach has been shown to be effective in work with couples, when one partner promises some particular change provided that the other reciprocates.

Treatment sessions in family therapy are rarely held more often than once a week and often only once every three or four weeks. Termination commonly occurs when the therapist considers that treatment has succeeded—or failed irretrievably—or when the family firmly decides to withdraw from treatment. There seems no doubt that family therapy can produce marked change within a family.

Behavioral psychotherapy. This is an approach to the treatment of mental disorders that uses a variety of methods based upon principles derived from experimental psychology, mainly that of learning theory. In the words of Joseph Wolpe, "behavior therapy, or conditioning therapy, is the use of experimentally established principles of learning for the purpose of changing unadaptive behavior. Unadaptive habits are weakened and eliminated; adaptive habits are initiated and strengthened" (*The Practice of Behavior Therapy,* 1973).

In the treatment of phobias, behaviour therapy seeks to modify and eliminate the avoidance response that the patient manifests when confronted with a phobic object or situation. This is crucial because, since the patient's avoidance of the anxiety-producing situation does indeed reduce his anxiety, his conditioned association of the phobic situation with the experience of anxiety remains unchallenged and frequently persists. Behaviour therapy interrupts this self-reinforcing pattern of avoidance behaviour by presenting the feared situation to the patient in a controlled manner such that it eventually ceases to produce anxiety; the strong associative link that has been built up within him between the feared situation, the experience of anxiety, and his subsequent avoidance behavior will thus be broken down and replaced by a less maladaptive set of responses.

The behavioral therapist is concerned with the forces and mechanisms that perpetuate the patient's present symptoms or abnormal behaviours, not with experiences in the past that may have caused them nor with any postulated intrapsychic conflict. Behavioral therapy focuses on observable phenomena; *i.e.,* what is done and what is said, rather than on what must be inferred (unconscious motives and processes and symbolic meanings).

The behavioral therapist carries out a detailed analysis of the patient's behaviour problems, paying particular attention to the circumstances in which they occur, to the patient's attempts to cope with his symptoms, and to his wish for change. The goals of treatment are precisely defined and usually do not include aims such as personal growth or personality change. The relationship between patient and therapist is sometimes said to be unimportant in behaviour therapy, and it is quite true that a patient may successfully follow a behavioral therapeutic program that he learns from a book or a personal computer. Nevertheless, a patient is more likely to complete an arduous program if he trusts and respects the therapist.

Behaviour therapy has become the preferred treatment for phobic states and for some obsessive-compulsive disorders, and it is effective in many cases of sexual dysfunction and deviation. It also has an important role in the rehabilitation of patients with chronic, disabling disorders.

Treatment of neurotic disorders. The essence of the treatment of phobias is the controlled exposure of the patient to the objects or situations that he fears. Behaviour therapy tries to eliminate the phobia by teaching the patient how to face those situations that clearly trigger his

discomfort so that he can learn to tolerate them. The exposure of the patient to the feared situation can be gradual (sometimes called desensitization) or rapid (sometimes known as flooding). Contrary to popular belief, the anxiety that is produced during exposure is not usually harmful. Even if severe panic does strike the sufferer, it will gradually evaporate and will be less likely to return in the future. Effective exposure treatments developed only as therapists learned to endure the phobic anxiety exhibited by their patients and when the therapists grew secure in the knowledge that such anxiety is much more likely to lead to the patients' improvement than to be harmful. The important point in this therapy is to persevere until the phobic anxiety starts to lessen and to be prepared to go on until it does. In general, the more rapidly and directly the worst fears are embraced by the patient, the more quickly the phobic terror fades to a tolerable mild tension.

Exposure techniques used to treat phobias

In the technique of desensitization, the patient is first taught how to practice muscular relaxation. He then reviews the situations of which he is afraid and lists them in order of increasing fearfulness, called a "hierarchy." Finally, the patient faces the various fear-producing situations in ascending order by means of vividly imagining them, countering any anxiety he feels by using relaxation techniques. This treatment is prolonged, and its use is restricted to feared situations that patients cannot regularly confront in real life, such as fear of lightning.

One of the most common phobic disorders treated by exposure techniques is agoraphobia (fear of open or public places). The patient, who is commonly a young woman, is encouraged to practice exposure daily, staying in a phobic situation for at least an hour, so that anxiety has time to reach a peak and then subside. The patient must be determined to get the better of the fears and not to run away from them. Agoraphobics cannot simply enter a dreaded crowded store, feel the familiar surge of panic, and rush out again. They must devote a full afternoon to a shopping trip. When panic strikes, the patient can sit in a corner of the store and ride out the terror. When feeling better, the patient can continue shopping. Persistence and patience are essential to conquering phobias in this way.

There is considerable evidence that exposure techniques work in most cases. Even phobias present for as long as 20 years can be overcome in a treatment program requiring no more than three to 15 hours of therapist time per patient. There is also considerable evidence that many phobics can treat themselves perfectly adequately by self-exposure without a therapist, using carefully devised self-help manuals.

Some patients with obsessive-compulsive disorders can also be helped by behaviour therapy. Several different techniques may need to be employed. For instance, a patient with an obsessional fear of contamination is treated by exposure, being taught to "soil" his hands with dirt and then to avoid washing them for longer and longer periods. Anxiety-management training enables the patient to withstand the anxiety he feels during the period of exposure.

This and other techniques have been shown to be effective in the treatment of compulsive rituals, with improvement occurring in more than two-thirds of patients. There is also a reduction in the frequency and intensity of obsessional thoughts that accompany the rituals. The treatment of obsessional thoughts that occur alone is much less satisfactory, however.

Other therapies. Many other types of psychotherapy have been developed in the second half of the 20th century, each with its own emphasis on symptom causation and its own particular approach to treatment. Many of these therapies use classical dynamic and behavioral models in modified forms, and they may also stress the understanding and modification of cognition and the ways in which people "process" their experiences, moods, and emotions. Among these relatively recent psychotherapies are client-centred psychotherapy, developed by the American psychologist Carl R. Rogers; transactional analysis, originated by the American psychiatrist Eric Berne; the interpersonal therapies developed by the American psychiatrists Adolf Meyer and Harry Stack Sullivan; cognitive therapy, developed by the American psychiatrist Aaron T. Beck; rational-emotive psychotherapy, developed by the American psychologist Albert Ellis; and Gestalt therapy, which stems from the work of the German psychiatrist Frederick S. Perls.

Another class of therapies consists of those used to care for psychotic patients, both those in hospitals and those who live in the community. Supportive psychotherapy consists of the long-term help of patients who are chronically handicapped by schizophrenia or other mental disorders. Such a therapy uses reassurance, guidance, and encouragement to help the patient cope with his disabilities and live as satisfactory a life as possible. Rehabilitation programs for chronic or episodically psychotic patients include drug maintenance; training in social skills that they may have lost while sick; occupational training to improve the patient's skills in cooking, shopping, and other domestic tasks; and industrial therapy, which usually offers the patient gainful employment under conditions of minimal stress. Family therapy is sometimes used to help relatives learn to cope with a schizophrenic patient who has returned home from the hospital.

Community care for released schizophrenics or other psychotic patients must provide them with drug maintenance and a minimum of psychiatric monitoring; appropriate housing facilities; some type of employment; and training in such skills as using public transport, preparing their own food, and looking after their finances. Each patient should have a case manager, a professional worker who ensures that the patient does not drop out of sight and who secures from governmental or social agencies the assistance that the patient needs. When provisions like these are not made, some formerly hospitalized patients stop taking their medicine and in effect drop out of the mental health care system, becoming unemployed and even homeless. This phenomenon became particularly evident in the United States and to a lesser extent in western Europe, when massive numbers of mental patients were released from hospitals during the 1950s and '60s after the effectiveness of antipsychotic drugs had been verified. These releases were also motivated by the concerns of civil libertarians over the abuse of patients' rights in keeping them committed to mental hospitals. However, the support network of community-based mental health clinics that would have been necessary to cope with the released patients was either inadequately established or nonexistent. The result was that many psychotic patients received inadequate outpatient care and supervision or encountered severe difficulties in obtaining housing or employment. Some of these people dropped out of the mental health care system altogether and became homeless wanderers in large urban areas. (J.L.Gi.)

Unforeseen effects of deinstitutionalization

BIBLIOGRAPHY

General works: The following works provide descriptions of the syndromes, causes, epidemiology, and methods of treatment of mental disorders: J.L. GIBBONS (ed.), *Psychiatry* (1983); SIR WILLIAM TRETHOWAN and A.C.P. SIMS, *Psychiatry,* 5th ed. (1983); ROBERT G. PRIEST and GERALD WOOLFSON, *Handbook of Psychiatry,* 8th ed. (1986); ROBERT J. WALDINGER, *Fundamentals of Psychiatry* (1986); and DAVID STAFFORD-CLARK and ANDREW C. SMITH, *Psychiatry for Students,* 6th ed. (1983). More detailed and comprehensive sources are MICHAEL GELDER, DENNIS GATH, and RICHARD MAYOU, *Oxford Textbook of Psychiatry* (1983); HAROLD I. KAPLAN and BENJAMIN J. SADOCK (eds.), *Comprehensive Textbook of Psychiatry/IV,* 4th ed., 2 vol. (1985); SILVANO ARIETI (ed.), *American Handbook of Psychiatry,* 2nd rev. ed., 8 vol. (1974–86); and M. SHEPHERD (ed.), *Handbook of Psychiatry,* 5 vol. (1982–85).

Classification and epidemiology: For the two classificatory systems mentioned in the text, see AMERICAN PSYCHIATRIC ASSOCIATION, *Diagnostic and Statistical Manual of Mental Disorders,* 3rd rev. ed. (1987); and WORLD HEALTH ORGANIZATION, *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death,* rev. ed., 2 vol. (1977–78). J.K. WING, J.E. COOPER, and N. SARTORIUS, *Measurement and Classification of Psychiatric Symptoms* (1974), provides an example of research diagnostic classification. See also HUGH L. FREEMAN (ed.), *Mental Health and the Environment* (1984); ALAN KERR and PHILIP SNAITH (eds.), *Contemporary Issues in Schizophrenia* (1986); EINAR KRINGLEN, *Heredity and Environment in the Functional Psychoses* (1968); and ANDREW SIMS, *Neurosis in Society* (1983).

Causation: Different aspects and theories of causation are considered in P. MCGUFFIN, M.F. SHANKS, and R.J. HODGSON, *The Scientific Principles of Psychopathology* (1984); DAVID M. SHAW, A.M.P. KELLAM, and R.F. MOTTRAM, *Brain Sciences in Psychiatry* (1982); TOM COX, *Stress* (1978); and GEORGE W. BROWN and TIRRIL HARRIS, *Social Origins of Depression: A Study of Psychiatric Disorder in Women* (1978). GORDON CLARIDGE, *Origins of Mental Illness: Temperament, Deviance, and Disorder* (1985), is a study of the interdependence of mental disorders and temperament, based on classic sources. H.P. LAUGHLIN, *The Ego and Its Defenses,* 2nd ed. (1979), is a detailed study of defense mechanisms.

Diagnostic categories: WILLIAM ALWYN LISHMAN, *Organic Psychiatry: The Psychological Consequences of Cerebral Disorder* (1978), deals comprehensively with the organic psychiatric syndromes; JOHN CUTTING, *The Psychology of Schizophrenia* (1985), studies theoretical and psychological aspects of the disorder; SILVANO ARIETI, *Interpretation of Schizophrenia,* 2nd rev. ed. (1974), presents major notions on schizophrenia. See also AARON T. BECK, *Depression: Clinical, Experimental, and Theoretical Aspects* (1967; reissued as *Depression: Cases and Treatment,* 1970); FRANZ ALEXANDER, *Psychosomatic Medicine: Its Principles and Applications* (1950, reprinted with a new introduction, 1987); and PHILIP SNAITH, *Clinical Neurosis* (1981).

Treatment: A comprehensive discussion of current treatment methods for various mental disorders is offered in JOHN H. GREIST, JAMES W. JEFFERSON, and ROBERT L. SPITZER (eds.), *Treatment of Mental Disorders* (1982). Works concerned with the theoretical concepts underlying psychotherapy include HENRI F. ELLENBERGER, *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry* (1970); and JEROME D. FRANK, *Persuasion and Healing: A Comparative Study of Psychotherapy,* rev. ed. (1973). A good account of the main forms of psychological treatment is given in SIDNEY BLOCH (ed.), *An Introduction to the Psychotherapies,* 2nd ed. (1986). See also PETER E. SIFNEOS, *Short-Term Dynamic Psychotherapy,* 2nd ed. (1987); IRVIN D. YALOM, *The Theory and Practice of Group Psychotherapy,* 3rd ed. (1985); and DAVID C. RIMM and JOHN C. MASTERS, *Behavior Therapy: Techniques and Empirical Findings,* 2nd ed. (1979).

Pharmacological and physical methods of treatment are dealt with in LOTHAR B. KALINOWSKY, HANNS HIPPIUS, and HELMFRIED E. KLEIN, *Biological Treatments in Psychiatry* (1982); ROSS J. BALDESSARINI, *Chemotherapy in Psychiatry: Principles and Practice,* rev. ed. (1985); and PETER DALLY and JOSEPH CONNOLLY, *An Introduction to Physical Methods of Treatment in Psychiatry,* 6th rev. ed. (1981). ELLIOT S. VALENSTEIN, *Great and Desperate Cures: The Rise and Decline of Psychosurgery and Other Radical Treatments for Mental Illness* (1986), presents a history of the methods and personalities involved.

(J.L.Gi./A.C.P.S.)

Metabolism

Living organisms are unique in that they can extract energy from their environments and use it to carry out activities such as movement, growth and development, and reproduction. But how do living organisms—or, their cells—extract energy from their environments, and how do cells use this energy to synthesize and assemble the components from which the cells are made?

The answers to these questions lie in the enzyme-mediated chemical reactions that take place in living matter (metabolism). Hundreds of coordinated, multistep reactions, fueled by energy obtained from nutrients and/or solar energy, ultimately convert readily available materials into the molecules required for growth and maintenance.

The physical and chemical properties of the components of living things dealt with in this article are found in the articles BIOCHEMICAL COMPONENTS OF ORGANISMS; CELLS; and PHOTOSYNTHESIS.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 322.

The article is divided into the following sections:

A summary of metabolism 976
- The unity of life 976
- Biological energy exchanges 976
  - The carrier of chemical energy
  - Catabolism
  - Anabolism
- Integration of catabolism and anabolism 978
  - Fine control
  - Coarse control
- The study of metabolic pathways 979

The fragmentation of complex molecules 980
- The catabolism of glucose 980
  - Glycolysis
  - The phosphogluconate pathway
- The catabolism of sugars other than glucose 983
- The catabolism of lipids (fats) 984
  - Fate of glycerol
  - Fate of fatty acids
- The catabolism of proteins 985
  - Removal of nitrogen
  - Disposal of nitrogen
  - Oxidation of the carbon skeleton

The combustion of food materials 987
- The oxidation of molecular fragments 987
  - The oxidation of pyruvate
  - The tricarboxylic acid (TCA) cycle
- Biological energy transduction 989
  - Adenosine triphosphate as the currency of energy exchange
  - Energy conservation

The biosynthesis of cell components 992
- The nature of biosynthesis 992
  - The stages of biosynthesis
  - Utilization of ATP
- The supply of biosynthetic precursors 992
  - Anaplerotic routes
  - Growth of microorganisms on TCA cycle intermediates
- The synthesis of building blocks 993
  - Gluconeogenesis
  - Lipid components
  - Amino acids
  - Mononucleotides
- The synthesis of macromolecules 998
  - Carbohydrates and lipids
  - Nucleic acids and proteins

Regulation of metabolism 1001
- Fine control 1001
  - End-product inhibition
  - Positive modulation
  - Energy state of the cell
- Coarse control 1002

Metabolic diseases 1003
- General considerations 1003
- The derivation of specific metabolic disorders 1004
- Disorders of carbohydrate metabolism 1005
  - Glycogen storage diseases
  - Galactosemia
  - Fructose disorders
  - Pyruvate disorders
  - Intestinal carbohydrate malabsorption
  - Mucopolysaccharidoses
- Disorders of lipid metabolism 1006
  - Blood lipid disorders
  - Lipid oxidative disorders
  - Tissue lipid disorders
- Disorders of amino-acid metabolism 1007
  - Phenylalanine and tyrosine
  - Urea cycle enzymes
  - Branched chain amino acids
  - Disorders of amino-acid transport
- Disorders of porphyrin metabolism 1009
- Disorders of purine and pyrimidine metabolism 1010

Bibliography 1010

A summary of metabolism

THE UNITY OF LIFE

At the cellular level of organization, the main chemical processes of all living matter are similar, if not identical. This is true for animals, plants, fungi, or bacteria; where variations occur (such as, for example, in the secretion of antibodies by some molds), the variant processes are but variations on common themes. Thus, all living matter is made up of large molecules called proteins, which provide support and coordinated movement, as well as storage and transport of small molecules, and, as catalysts, enable chemical reactions to take place rapidly and specifically under mild temperature, relatively low concentration, and neutral conditions (*i.e.,* neither acidic nor basic). Proteins are assembled from some 20 amino acids, and, just as the 26 letters of the alphabet can be assembled in specific ways to form words of various lengths and meanings, so may tens or even hundreds of the 20 amino-acid "letters" be joined to form specific proteins. Moreover, those portions of protein molecules involved in performing similar functions in different organisms often comprise the same sequences of amino acids.

Proteins

There is the same unity among cells of all types in the manner in which living organisms preserve their individuality and transmit it to their offspring. For example, hereditary information is encoded in a specific sequence of bases that make up the DNA (deoxyribonucleic acid) molecule in the nucleus of each cell. Only four bases are used in synthesizing DNA: adenine, guanine, cytosine, and thymine. Just as the Morse Code consists of three simple signals—a dash, a dot, and a space—the precise arrangement of which suffices to convey coded messages, so the precise arrangement of the bases in DNA contains and conveys the information for the synthesis and assembly of cell components. Some primitive life-forms, however, use RNA (ribonucleic acid; a nucleic acid differing from DNA in containing the sugar ribose instead of the sugar deoxyribose and the base uracil instead of the base thymine) in place of DNA as a primary carrier of genetic information. The replication of the genetic material in these organisms must, however, pass through a DNA phase. With minor exceptions, the genetic code used by all living organisms is the same.

The chemical reactions that take place in living cells are similar as well. Green plants use the energy of sunlight to convert water (H_2O) and carbon dioxide (CO_2) to carbohydrates (sugars and starches), other organic (carbon-containing) compounds, and molecular oxygen (O_2). The process of photosynthesis requires energy, in the form of sunlight, to split one water molecule into one-half of an oxygen molecule (O_2; the oxidizing agent) and two hydrogen atoms (H; the reducing agent), each of which dissociates to one hydrogen ion (H^+) and one electron. Through a series of oxidation-reduction reactions, electrons (denoted e^-) are transferred from a donating molecule (oxidation), in this case water, to an accepting molecule (reduction) by a series of chemical reactions; this "reducing power" may be coupled ultimately to the reduction of carbon dioxide to the level of carbohydrate. In effect, carbon dioxide accepts and bonds with hydrogen, forming carbohydrates ($C_n[H_2O]_n$).

Living organisms that require oxygen reverse this process: they consume carbohydrates and other organic materials, using oxygen synthesized by plants to form water, carbon dioxide, and energy. The process that removes hydrogen atoms (containing electrons) from the carbohydrates and passes them to the oxygen is an energy-yielding series of reactions.

In plants, all but two of the steps in the process that converts carbon dioxide to carbohydrates are the same as those steps that synthesize sugars from simpler starting materials in animals, fungi, and bacteria. Similarly, the series of reactions that take a given starting material and synthesize certain molecules that will be used in other synthetic pathways are similar, or identical, among all cell types. From a metabolic point of view, the cellular processes that take place in a lion are only marginally different from those that take place in a dandelion.

BIOLOGICAL ENERGY EXCHANGES

The energy changes associated with physicochemical processes are the province of thermodynamics, a subdiscipline of physics. The first two laws of thermodynamics state, in essence, that energy can be neither created nor destroyed and that the effect of physical and chemical changes is to increase the disorder, or randomness (*i.e.,* entropy), of the universe. Although it might be supposed that biological processes—through which organisms grow in a highly ordered and complex manner, maintain order and complexity throughout their life, and pass on the instructions for order to succeeding generations—are in contravention of these laws, this is not so. Living organisms neither consume nor create energy: they can only transform it from one form to another. From the environment they absorb energy in a form useful to them; to the environment they return an equivalent amount of energy in a biologically less useful form. The useful energy, or free energy, may be defined as energy capable of doing work under isothermal conditions (conditions in which no temperature differential exists); free energy is associated with any chemical change. Energy less useful than free energy is returned to the environment, usually as heat. Heat cannot perform work in biological systems because all parts of cells have essentially the same temperature and pressure.

Thermodynamics

Free energy and heat

The carrier of chemical energy. At any given time, a neutral molecule of water dissociates into a hydrogen ion (H^+) and a hydroxide ion (OH^-), and the ions are continually re-forming into the neutral molecule. Under normal conditions (neutrality), the concentration of hydrogen ions (acidic ions) is equal to that of the hydroxide ions (basic ions); each are at a concentration of 10^{-7} moles per litre, which is described as a pH of 7.

All cells either are bounded by membranes or contain organelles that have membranes. These membranes do not permit water or the ions derived from water to pass into or out of the cells or organelles. In green plants, sunlight is absorbed by chlorophyll and other pigments in the chloroplasts of the cells, called photosystem II. As shown previously, when a water molecule is split by light energy, one-half of an oxygen molecule and two hydrogen atoms (which dissociate to two electrons and two hydrogen ions, H^+) are formed. When excited by sunlight, chlorophyll loses one electron to an electron carrier molecule but quickly recovers it from a hydrogen atom of the split water molecule, which sends H^+ into solution in the process. Two oxygen atoms come together to form a molecule of oxygen gas (O_2). The free electrons are passed to photosystem I, but, in doing so, an excess concentration of positively charged hydrogen ions (H^+) appears on one side of the membrane in the chloroplast, whereas an excess of negatively charged hydroxide ions (OH^-) builds up on the other side. The free energy released as H^+ ions move through a specific "pore" in the membrane, to equalize the concentrations of ions, is sufficient to make some biological processes work, such as the uptake of certain nutrients by bacteria and the rotation of the whiplike protein-based propellers that enable such bacteria to move. Equally important, however, is that this gradient across the membrane powers the formation of adenosine triphosphate (ATP) from inorganic phosphate (HPO_4^{2-}, abbreviated P_i) and adenosine diphosphate (ADP). It is ATP (Figure 1) that is the major carrier of biologically utilizable energy in all forms of living matter. The interrelationships of energy-yielding and energy-requiring metabolic reactions may be considered largely as processes that couple the formation of ATP with its breakdown.

Synthesis of ATP by green plants is similar to the synthesis of ATP that takes place in the mitochondria of animal, plant, and fungus cells, and in the plasma membranes of bacteria that use oxygen (or other inorganic electron acceptors, such as nitrate) to accept electrons from the removal of hydrogen atoms from a molecule of food (see below *The combustion of food materials: Biological energy transduction*). Through these processes most of the energy stored in food materials is released and converted into the molecules that fuel life processes. It must also be remembered, however, that many living organisms (usually bacteria and protozoa) cannot tolerate oxygen; they form

Plant synthesis of ATP

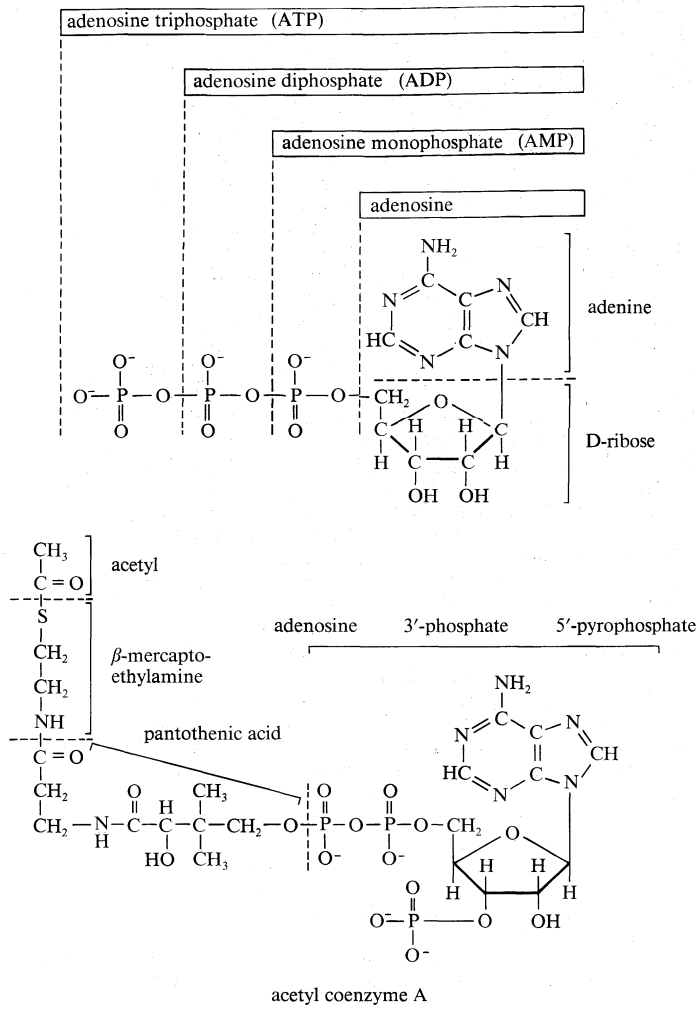


Figure 1: *Biological energy carriers.*
(Top) ATP, ADP, and AMP. (Bottom) Acetyl coenzyme A (acetyl CoA).

ATP from inorganic phosphate and ADP by substrate-level phosphorylations (the addition of a phosphate group) that do not involve the establishment and collapse of proton gradients across membranes. Such processes are discussed in detail below (*The fragmentation of complex molecules: The catabolism of glucose*). It must also be borne in mind that the fuels of life and the cellular "furnace" in which they are "burned" are made of the same types of material: if the fires burn too brightly, not only the fuel but also the furnace is consumed. It is therefore essential to release energy at small, discrete, readily utilizable intervals. The relative complexity of the catabolic pathways (by which food materials are broken down) and the complexity of the anabolic pathways (by which cell components are synthesized) reflect this need and offer the possibility for simple feedback systems to control the rate at which materials travel along these sequences of enzymic reactions.

Catabolism. *Formation of small molecules.* The release of chemical energy from food materials essentially occurs in three phases. In the first phase (phase I), the large molecules that make up the bulk of food materials are broken down into small constituent units: proteins are converted to the 20 or so different amino acids of which they are composed; carbohydrates (polysaccharides such as starch in plants and glycogen in animals) are degraded to sugars such as glucose; and fats (lipids) are broken down into fatty acids and glycerol. The amounts of energy liberated in phase I are relatively small: only about 0.6 percent of the free, or useful, energy of proteins and carbohydrates, and about 0.1 percent of that of fats, is released during this phase. Because this energy is liberated largely as heat, it cannot be utilized by the cell. The purpose of the reactions of phase I, which can be grouped under the term digestion and which, in animals, occur mainly in the intestinal tract and in tissues in which reserve materials are prepared, or mobilized, for energy production, is to prepare the foodstuffs for the energy-releasing processes.

Phase I

Incomplete oxidation. In the second phase of the release of energy from food (phase II), the small molecules produced in the first phase—sugars, glycerol, a number of fatty acids, and about 20 varieties of amino acids—are incompletely oxidized (in this sense, oxidation means the removal of electrons or hydrogen atoms), the end product being (apart from carbon dioxide and water) one of only three possible substances: the two-carbon compound acetate, in the form of a compound called acetyl coenzyme A (Figure 1); the four-carbon compound oxaloacetate; and the five-carbon compound α-oxoglutarate. The first, acetate in the form of acetyl coenzyme A, constitutes by far the most common product—it is the product of two-thirds of the carbon incorporated into carbohydrates and glycerol; all of the carbon in most fatty acids; and approximately half of the carbon in amino acids. The end product of several amino acids is α-oxoglutarate; that of a few others is oxaloacetate, which is formed either directly or indirectly (from fumarate). These processes, represented diagrammatically in Figure 2, show what happens in the bacterium *Escherichia coli,* but essentially similar processes occur in animals, plants, fungi, and other organisms capable of oxidizing their food materials wholly to carbon dioxide and water.

Phase II

Complete oxidation. Total oxidation of the relatively few products of phase II occurs in a cyclic sequence of chemical reactions known as the tricarboxylic acid (TCA) cycle, or the Krebs cycle, after its discoverer, Sir Hans Krebs; it represents phase III of energy release from foods. Each turn of this cycle (see below *The tricarboxylic acid [TCA] cycle*) is initiated by the formation of citrate, with six carbon atoms, from oxaloacetate (with four carbons) and acetyl coenzyme A; subsequent reactions result in the reformation of oxaloacetate and the formation of two molecules of carbon dioxide. The carbon atoms that go into the formation of carbon dioxide are no longer available to the cell. The concomitant stepwise oxidations—in which hydrogen atoms or electrons are removed from intermediate compounds formed during the cycle and, via a system of carriers, are transferred ultimately to oxygen to form water—are quantitatively the most important means of generating ATP from ADP and inorganic phosphate.

Phase III

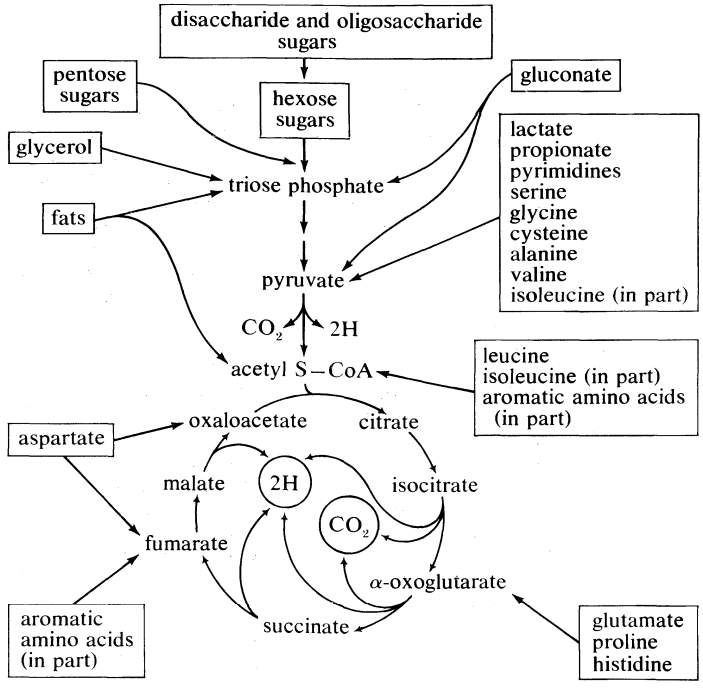


Figure 2: Pathways for the catabolism of nutrients by *Escherichia coli.*

These events are known as terminal respiration and oxidative phosphorylation (for details of this process, see below *Biological energy transduction*).

Some microorganisms, incapable of completely converting their carbon compounds to carbon dioxide, release energy by fermentation reactions, in which the intermediate compounds of catabolic routes either directly or indirectly accept or donate hydrogen atoms. Such secondary changes in intermediate compounds result in considerably less energy being made available to the cell than occurs with the pathways that are linked to oxidative phosphorylation; however, fermentation reactions yield a large variety of commercially important products. Thus, for example, if the oxidation (removal of electrons or hydrogen atoms) of some catabolic intermediate is coupled to the reduction of pyruvate or of acetaldehyde derived from pyruvate, the products formed are lactic acid and ethyl alcohol, respectively.

Anabolism versus catabolism

Anabolism. Catabolic pathways effect the transformation of food materials into the interconvertible intermediates of the pathways shown in Figure 2. Anabolic pathways, on the other hand, are sequences of enzyme-catalyzed reactions in which the component building blocks of large molecules, or macromolecules (*e.g.*, proteins, carbohydrates, and fats), are constructed from the same intermediates. Thus, catabolic routes have clearly defined beginnings but no unambiguously identifiable end products; anabolic routes, on the other hand, lead to clearly distinguishable end products from diffuse beginnings. The two types of pathway are linked through reactions of phosphate transfer, involving ADP, AMP, and ATP as described above, and also through electron transfers, which enable reducing equivalents (*i.e.*, hydrogen atoms or electrons), which have been released during catabolic reactions, to be utilized for biosynthesis. But, although catabolic and anabolic pathways are closely linked, and although the overall effect of one type of route is obviously the opposite of the other, they have few steps in common. The anabolic pathway for the synthesis of a particular molecule generally starts from intermediate compounds quite different from those produced as a result of catabolism of that molecule; for example, microorganisms catabolize aromatic (*i.e.*, containing a ring, or cyclic, structure) amino acids to acetyl coenzyme A and an intermediate compound of the TCA cycle. The biosynthesis of these amino acids, however, starts with a compound derived from pyruvate and an intermediate compound of the metabolism of pentose (a general name for sugars with five carbon atoms). Similarly, histidine is synthesized from a pentose sugar but is catabolized to α-oxoglutarate.

Even in cases in which a product of catabolism is used in an anabolic pathway, differences emerge; thus, fatty acids, which are catabolized to acetyl coenzyme A, are synthesized not from acetyl coenzyme A directly but from a derivative of it, malonyl coenzyme A (see below *The biosynthesis of cell components: Lipid components*). Furthermore, even enzymes that catalyze apparently identical steps in catabolic and anabolic routes may exhibit different properties. In general, therefore, the way down (catabolism) is different from the way up (anabolism). These differences are important because they allow for the regulation of catabolic and anabolic processes in the cell.

Eukaryotic cells

In eukaryotic cells (*i.e.*, those with a well-defined nucleus, characteristic of organisms higher than bacteria) the enzymes of catabolic and anabolic pathways are often located in different cellular compartments. This also contributes to the manner of their cellular control; for example, the formation of acetyl coenzyme A from fatty acids, referred to above, occurs in animal cells in small sausage-shaped components, or organelles, called mitochondria, which also contain the enzymes for terminal respiration and for oxidative phosphorylation. The biosynthesis of fatty acids from acetyl coenzyme A, on the other hand, occurs in the cytoplasm.

INTEGRATION OF CATABOLISM AND ANABOLISM

Fine control. Possibly the most important means for controlling the flux of metabolites through catabolic and anabolic pathways, and for integrating the numerous different pathways in the cell, is through the regulation of either the activity or the synthesis of key (pacemaker) enzymes. It was recognized in the 1950s, largely from work with microorganisms, that pacemaker enzymes can interact with small molecules at more than one site on the surface of the enzyme molecule. The reaction between an enzyme and its substrate—defined as the compound with which the enzyme acts to form a product—occurs at a specific site on the enzyme known as the catalytic, or active, site; the proper fit between the substrate and the active site is an essential prerequisite for the occurrence of a reaction catalyzed by an enzyme (see BIOCHEMICAL COMPONENTS OF ORGANISMS). Interactions at other, so-called regulatory sites on the enzyme, however, do not result in a chemical reaction but cause changes in the shape of the protein; the changes profoundly affect the catalytic properties of the enzyme, either inhibiting or stimulating the rate of the reaction. Modulation of the activity of pacemaker enzymes may be effected by metabolites of the pathway in which the enzyme acts or by those of another pathway; the process may be described as a "fine control" of metabolism. Very small changes in the chemical environment thus produce important and immediate effects on the rates at which individual metabolic processes occur.

Most catabolic pathways are regulated by the relative proportions of ATP, ADP, and AMP in the cell. It is reasonable to suppose that a pathway that serves to make ATP available for energy-requiring reactions would be less active if sufficient ATP were already present, than if ADP or AMP were to accumulate. The relative amounts of the adenine nucleotides (*i.e.*, ATP, ADP, and AMP) thus modulate the overall rate of catabolic pathways. They do so by reacting with specific regulatory sites on pacemaker enzymes necessary for the catabolic pathways, which do not participate in the anabolic routes that effect the opposite reactions. Similarly, it is reasonable to suppose that many anabolic processes, which require energy, are inhibited by ADP or AMP; elevated levels of these nucleotides may be regarded therefore as cellular distress signals indicating a lack of energy.

Since one way in which anabolic pathways differ from catabolic routes is that the former result in identifiable end products, it is not unexpected that the pacemaker enzymes of many anabolic pathways—particularly those effecting the biosynthesis of amino acids and nucleotides—are regulated by the end products of these pathways or, in cases in which branching of pathways occurs, by end products of each branch. Such pacemaker enzymes usually act at the first step unique to a particular anabolic route. If branching occurs, the first step of each branch is controlled. By this so-called negative feedback system, the cellular concentrations of products determine the rates of their formation, thus ensuring that the cell synthesizes only as much of the products as it needs.

Coarse control. A second and less immediately responsive, or "coarse," control is exerted over the synthesis of pacemaker enzymes. The rate of protein synthesis reflects the activity of appropriate genes, which contain the information that directs all cellular processes. Coarse control is therefore exerted on genetic material rather than on enzymes. Preferential synthesis of a pacemaker enzyme is particularly required to accommodate a cell to major changes in its chemical milieu. Such changes occur in multicellular organisms only to a minor extent, so that this type of control mechanism is less important in animals than in microorganisms. In the latter, however, it may determine the ease with which a cell previously growing in one nutrient medium can grow after transfer to another. In cases in which several types of organism compete in the same medium for available carbon sources, the operation of coarse controls may well be decisive in ensuring survival.

Gene involvement

Alterations in the differential rates of synthesis of pacemaker enzymes in microorganisms responding to changes in the composition of their growth medium also manifest the properties of negative feedback systems. Depending on the nature of the metabolic pathway of which a pacemaker enzyme is a constituent, the manner in which the alterations are elicited may be distinguished. Thus, an in-

crease in the rates at which enzymes of catabolic routes are synthesized results from the addition of inducers—usually compounds that exhibit some structural similarity to the substrates on which the enzymes act. A classic example of an inducible enzyme of this type is β-galactosidase. *Escherichia coli* growing in nutrient medium containing glucose do not utilize the milk sugar, lactose (glucose-4-β-D-galactoside); however, if the bacteria are placed in a growth medium containing lactose as the sole source of carbon, they synthesize β-galactosidase and can therefore utilize lactose. The reaction catalyzed by the enzyme is the hydrolysis (*i.e.,* breakdown involving water) of lactose to its two constituent sugars, glucose and galactose; the preferential synthesis of the enzyme thus allows the bacteria to use the lactose for growth and energy. Another characteristic of the process of enzyme induction is that it continues only as long as the inducer (in this case, lactose) is present; if cells synthesizing β-galactosidase are transferred to a medium containing no lactose, synthesis of β-galactosidase ceases, and the amount of the enzyme in the cells is diluted as they divide, until the original low level of the enzyme is reestablished.

In contrast, the differential rates of synthesis of pacemaker enzymes of anabolic routes are usually not increased by the presence of inducers. Instead, the absence of small molecules that act to repress enzyme synthesis accelerates enzyme formation. Similar to the fine control processes described above is the regulation by coarse control of many pacemaker enzymes of amino-acid biosynthesis. Like the end product inhibitors, the repressors in these cases also appear to be the amino-acid end products themselves.

It is useful to regard the acceleration of the enzyme-forming machinery as the consequence, metaphorically, of either placing a foot on the accelerator or removing it from the brake. Analysis of the mechanisms by which gene activity is controlled suggest, however, that the distinction between inducible and repressible enzymes may be more apparent than real (see below *Regulation of metabolism*).

THE STUDY OF METABOLIC PATHWAYS

Purpose

There are two main reasons for studying a metabolic pathway: (1) to describe, in quantitative terms, the chemical changes catalyzed by the component enzymes of the route; and (2) to describe the various intracellular controls that govern the rate at which the pathway functions.

Studies with whole organisms or organs can provide information that one substance is converted to another and that this process is localized in a certain tissue; for example, experiments can show that urea, the chief nitrogen-containing end product of protein metabolism in mammals, is formed exclusively in the liver. They cannot reveal, however, the details of the enzymatic steps involved. Clues to the identity of the products involved, and to the possible chemical changes effected by component enzymes, can be provided in any of four ways involving studies with either whole organisms or tissues.

Identification of metabolites

First, under stress or the imbalances associated with diseases, certain metabolites (*i.e.,* products of metabolism) may accumulate to a much greater extent than normal. Thus, during the stress of violent exercise, lactic acid appears in the blood, while glycogen, the form in which carbohydrate is stored in muscle, disappears. Such observations do not, however, prove that lactic acid is a normal intermediate of glycogen catabolism; rather, they show only that compounds capable of yielding lactic acid are likely to be normal intermediates. Indeed, in the example, lactic acid is formed in response to abnormal circumstances and is not directly formed in the pathways of carbohydrate catabolism. On the other hand, the abnormal accumulation of pyruvic acid in the blood of vitamin B_1-deficient pigeons was a valuable clue to the role of this vitamin in the oxidation of pyruvate.

Second, the administration of metabolic poisons may lead to the accumulation of specific metabolites. If fluoroacetic acid or fluorocitric acid is ingested by animals, for example, citric acid accumulates in the liver. This correctly suggests that fluorocitric acid administered as such, or formed from fluoroacetic acid via the tricarboxylic acid (TCA) cycle, inhibits an enzyme of citrate oxidation.

Third, the fate of any nutrient—indeed, often the fate of a particular chemical group or atom in a nutrient—can be followed with relative ease by administering the nutrient labeled with an isotope. Isotopes are forms of an element that are chemically indistinguishable from each other but differ in physical properties.

The use of a nonradioactive isotope of nitrogen in the 1930s first revealed the dynamic state of body constituents. It had previously been believed that the proteins of tissues are stable once formed, disappearing only with the death of the cell. By feeding amino acids labeled with isotopic nitrogen to rats, it was discovered that the isotope was incorporated into many of the amino acids found in proteins of the liver and the gut, even though the total protein content of these tissues did not change. This suggested that the proteins of these tissues exist in a dynamic steady state, in which relatively high rates of synthesis are counterbalanced by equal rates of degradation. Thus, although the average liver cell has a life-span of several months, half of its proteins are synthesized and degraded every five to six days. On the other hand, the proteins of the muscle or the brain, tissues that (unlike the gut or liver) need not adjust to changes in the chemical composition of their milieu, do not turn over as rapidly. The high rates of turnover observed in liver and gut tissues indicate that the coarse controls, exerted through the onset and cessation of synthesis of pacemaker enzymes, do occur in animal cells.

Rate of protein turnover

Finally, genetically altered organisms (mutants) fail to synthesize certain enzymes in an active form. Such defects, if not lethal, result in the accumulation and excretion of the substrate of the defective enzyme; in normal organisms, the substrate would not accumulate, because it would be acted upon by the enzyme. The significance of this observation was first realized in the early 20th century when the phrase "inborn errors of metabolism" was used to describe hereditary conditions in which a variety of amino acids and other metabolites are excreted in the urine (see below *Metabolic diseases*). In microorganisms, in which it is relatively easy to cause genetic mutations and to select specific mutants, this technique has been very useful. In addition to their utility in the unraveling of metabolic pathways, the use of mutants in the early 1940s led to the postulation of the one gene-one enzyme hypothesis by the Nobel Prize winners George W. Beadle and Edward L. Tatum; their discoveries opened the field of biochemical genetics and first revealed the nature of the fine controls of metabolism.

Because detailed information about the mechanisms of component enzymatic steps in any metabolic pathway cannot be obtained from studies with whole organisms or tissues, various techniques have been developed for studying these processes—*e.g.,* sliced tissues, and homogenates and cell-free extracts, which are produced by physical disruption of the cells and the removal of cell walls and other debris. The sliced-tissue technique was successfully used by the Nobel Prize winner Sir Hans Krebs in his pioneer studies in the early 1930s on the mechanism of urea formation in the liver. Measurements were made of the stimulating effects of small quantities of amino acids on both the rate of oxygen uptake and the amount of oxygen taken up; the amino acids were added to liver slices bathed in a nutrient medium. Such measurements revealed the cyclic nature of the process; specific amino acids acted as catalysts, stimulating respiration to an extent greater than expected from the quantities added. This was because the added material had been re-formed in the course of the cycle (see below *The catabolism of proteins: Disposal of nitrogen*).

The use of homogenates

Homogenates of tissue are useful in studying metabolic processes because permeability barriers that may prevent ready access of external materials to cell components are destroyed. The tissue is usually minced, blended, or otherwise disrupted in a medium that is suitably buffered to maintain the normal acid–base balance of the tissue, and contains the ions required for many life processes, chiefly sodium, potassium, and magnesium. The tissue is either used directly—as was done by Krebs in elucidating, in 1937, the TCA cycle from studies of the respiration of minced pigeon breast muscle—or fractionated (*i.e.,* bro-

ken down) further. If the latter procedure is followed, homogenization is often carried out in a medium containing a high concentration of the sugar sucrose, which provides a milieu favourable for maintaining the integrity of cellular components. The components are recovered by careful spinning in a centrifuge, at a series of increasing speeds. It is thus possible to obtain fractions containing predominantly one type of organelle: nuclei (and some unbroken cells); mitochondria, lysosomes, and microbodies; microsomes (*i.e.,* ribosomes and endoplasmic reticulum fragments); and—after prolonged centrifugation at forces in excess of 100,000 times gravity—a clear liquid that represents the soluble fraction of the cytoplasm. The fractions thus obtained can be further purified and tested for their capacity to carry out a given metabolic step or steps. This procedure was used to show that isolated mitochondria catalyze the oxidation reactions of the TCA cycle and that these organelles also contain the enzymes of fatty acid oxidation. Similarly, isolated ribosomes are used to study the pathway and mechanism of protein synthesis.

The final step in elucidating a reaction in a metabolic pathway includes isolation of the enzyme involved. The rate of the reaction and the factors that control the activity of the enzyme are then measured.

It should be emphasized that biochemists realize that studies on isolated and highly purified systems, such as those briefly described above, can do no more than approximate biological reality. The identification of the fine and coarse controls of a metabolic pathway, and (when appropriate) other influences on that pathway, must ultimately involve the study of the pathway in the whole cell or organism. Although some techniques have proved adequate for relating findings in the test tube to the situation in living organisms, study of the more complex metabolic processes, such as those involved in differentiation and development, may require the elaboration of new experimental approaches.

The fragmentation of complex molecules

Food materials must undergo oxidation in order to yield biologically useful energy. Oxidation does not necessarily involve oxygen, although it must involve the transfer of electrons from a donor molecule to a suitable acceptor molecule; the donor is thus oxidized and the recipient reduced. Many microorganisms either must live in the absence of oxygen (*i.e.,* are obligate anaerobes) or can live in its presence or its absence (*i.e.,* are facultative anaerobes).

Fermentations

If no oxygen is available, the catabolism of food materials is effected via fermentations, in which the final acceptor of the electrons removed from the nutrient is some organic molecule, usually generated during the fermentation process. There is no net oxidation of the food molecule in this type of catabolism; that is, the overall oxidation state of the fermentation products is the same as that of the starting material.

Organisms that can use oxygen as a final electron acceptor also use many of the steps in the fermentation pathways in which food molecules are broken down to smaller fragments; these fragments, instead of serving as electron acceptors, are fed into the TCA cycle, the pathway of terminal respiration.

In this cycle all of the hydrogen atoms (H) or electrons (e^-) are removed from the fragments and are channeled through a series of electron carriers, ultimately to react with oxygen (O; see below *Energy conservation*). All carbon atoms are eliminated as carbon dioxide (CO_2) in this process. The sequence of reactions involved in the catabolism of food materials may thus be conveniently considered in terms of an initial fragmentation (fermentation), followed by a combustion (respiration) process.

THE CATABOLISM OF GLUCOSE

Glycolysis. *The transformation of glucose.* Quantitatively, the most important source of energy for cellular processes is the six-carbon sugar glucose ($C_6H_{12}O_6$). Two structures of glucose are shown in Figure 3, in which the carbon atoms are numbered. (See BIOCHEMICAL COMPONENTS OF ORGANISMS for a discussion of the chemical nature of glucose and other carbohydrates.) Glucose is made available to animals through the hydrolysis of polysaccharides, such as glycogen and starch, the process being catalyzed by digestive enzymes. In animals, the sugar thus set free passes from the gut into the bloodstream and from there into the cells of the liver and other tissues. In microorganisms, of course, no such specialized tissues are involved.

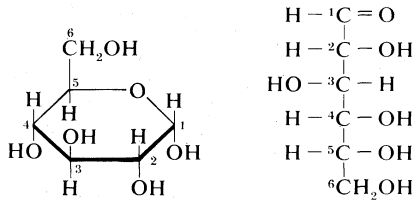
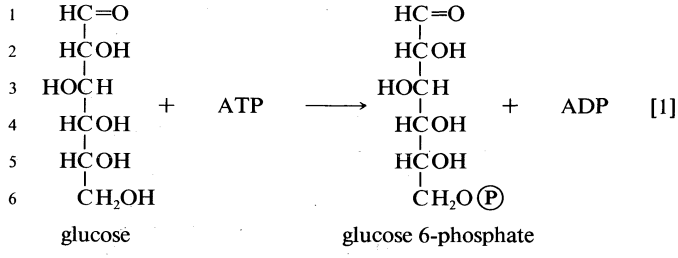


Figure 3: The structure of α-D-glucose.

The fermentative phase of glucose catabolism (glycolysis) involves several enzymes; the action of each is summarized below. In living cells many of the compounds that take part in metabolism exist as negatively charged moieties, or anions, and are named as such in most of this article; *e.g.,* pyruvate, oxaloacetate.

In order to obtain a net yield of ATP from the catabolism of glucose, it is first necessary to invest ATP. During step [1] the alcohol group at position 6 of the glucose molecule readily reacts with the terminal phosphate group of ATP, forming glucose 6-phosphate and ADP. For convenience, the phosphoryl group (PO_3^{2-}) is represented by Ⓟ. Because the decrease in free energy is so large, this reaction is virtually irreversible under physiological conditions.

The formation of glucose 6-phosphate

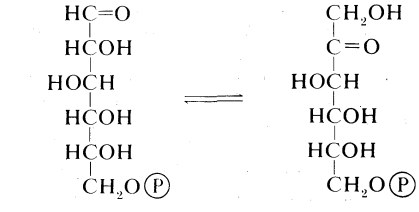


[1]

In animals, this phosphorylation of glucose, which yields glucose 6-phosphate, is catalyzed by two different enzymes. In most cells a hexokinase with a high affinity for glucose—*i.e.,* only small amounts of glucose are necessary for enzymatic activity—effects the reaction. In addition, the liver contains a glucokinase, which requires a much greater concentration of glucose before it reacts. Glucokinase functions only in emergencies, when the concentration of glucose in the blood rises to abnormally high levels.

Certain facultative anaerobic bacteria also contain hexokinases but apparently do not use them to phosphorylate glucose. In such cells, external glucose can be utilized only if it is first phosphorylated to glucose 6-phosphate via a system linked to the cell membrane that involves a compound called phosphoenolpyruvate (formed in step [9] of glycolysis), which serves as an obligatory donor of the phosphate group; *i.e.,* ATP cannot serve as the phosphate donor in the reaction.

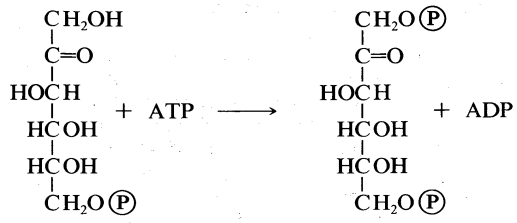
The reaction in which glucose 6-phosphate is changed to fructose 6-phosphate is catalyzed by phosphoglucoisomerase [2]. In the reaction, a secondary alcohol group



[2]

(−CHOH) at the second carbon atom is oxidized to a keto-group (*i.e.*, −C=O), and the aldehyde group (−CHO) at the first carbon atom is reduced to a primary alcohol group (−CH_2OH). Reaction [2] is readily reversible, as is indicated by the double arrows.

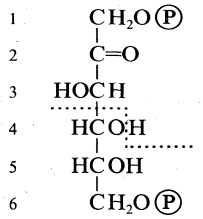
The formation of the alcohol group at the first carbon atom permits the repetition of the reaction effected in step [1]; that is, a second molecule of ATP is invested. The product is fructose 1,6-diphosphate [3]. Again, as in the hexokinase reaction, the decrease in free energy of the reaction, which is catalyzed by phosphofructokinase, is sufficiently large to make this reaction virtually irreversible under physiological conditions; ADP is also a product.



[3]

fructose 6-phosphate fructose 1,6-diphosphate

The first three steps of glycolysis have thus transformed an asymmetrical sugar molecule, glucose, into a symmetrical form, fructose 1,6-diphosphate, containing a phosphoryl group at each end; the molecule next is split into two smaller fragments that are interconvertible. This elegant simplification is achieved via steps [4] and [5], which are described below.



[4]

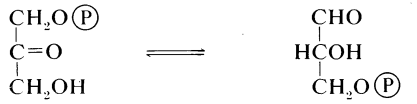
fructose 1,6-diphosphate



dihydroxyacetone phosphate glyceraldehyde 3-phosphate

The aldolase reaction. In [4], an enzyme catalyzes the breaking apart of the six-carbon sugar fructose 1,6-diphosphate into two three-carbon fragments. The molecule is split between carbons 3 and 4. Reversal of this cleavage—*i.e.*, the formation of a six-carbon compound from two three-carbon compounds—is possible. Because the reverse reaction is an aldol condensation—*i.e.*, an aldehyde (glyceraldehyde 3-phosphate) combines with a ketone (dihydroxyacetone phosphate)—the enzyme is commonly called aldolase. The two three-carbon fragments produced in step [4], dihydroxyacetone phosphate and glyceraldehyde 3-phosphate, are also called triose phosphates. They are readily converted to each other by a process [5] analogous to that in step [2]. The enzyme that catalyzes the interconversion [5] is triose phosphate isomerase, a different enzyme than that catalyzing step [2].

Aldol condensation

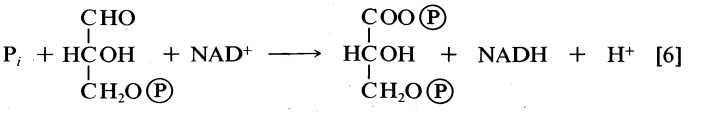


[5]

dihydroxyacetone phosphate glyceraldehyde 3-phosphate

The formation of ATP. The second stage of glucose catabolism comprises reactions [6] through [10], in which a net gain of ATP is achieved through the oxidation of one of the triose phosphate compounds formed in step [5]. One molecule of glucose forms two molecules of the triose phosphate; both three-carbon fragments follow the same pathway, and steps [6] through [10] must occur twice to complete the glucose breakdown.

Conversion of energy of oxidation into ATP

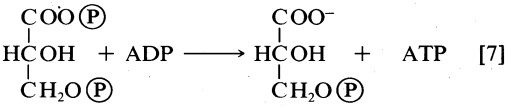


[6]

glyceraldehyde 3-phosphate 1,3-diphosphoglycerate

Step [6], in which glyceraldehyde 3-phosphate is oxidized, is one of the most important reactions in glycolysis. It is during this step that the energy liberated during oxidation of the aldehyde group (−CHO) is conserved in the form of a high-energy phosphate compound; namely, as 1,3-diphosphoglycerate, an anhydride of a carboxylic acid and phosphoric acid. The hydrogen atoms or electrons removed from the aldehyde group during its oxidation are accepted by a coenzyme (so called because it functions in conjunction with an enzyme) involved in hydrogen or electron transfer; the coenzyme, nicotinamide adenine dinucleotide (NAD^+), is reduced to form $NADH + H^+$ in the process. The NAD^+ thus reduced is bound to the enzyme glyceraldehyde 3-phosphate dehydrogenase, catalyzing the overall reaction, step [6].

The 1,3-diphosphoglycerate produced in step [6] reacts with ADP in a reaction catalyzed by phosphoglycerate kinase, with the result that one of the two phosphoryl groups is transferred to ADP to form ATP and 3-phosphoglycerate. This reaction [7] is highly exergonic (*i.e.*, it proceeds with a loss of free energy); as a result, the oxidation of glyceraldehyde 3-phosphate, step [6], is irreversible. In summary, the energy liberated during oxidation of an aldehyde group (−CHO in glyceraldehyde 3-phosphate) to a carboxylic acid group (−COO^- in 3-phosphoglycerate) is conserved as the phosphate bond energy in ATP during steps [6] and [7]. This step occurs twice for each molecule of glucose; thus the initial investment of ATP in steps [1] and [3] is recovered.

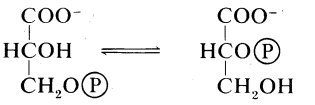


[7]

1,3-diphosphoglycerate 3-phosphoglycerate

The 3-phosphoglycerate in step [7] now forms 2-phosphoglycerate, in a reaction catalyzed by phosphoglyceromutase [8]. During step [9] the enzyme enolase reacts with 2-phosphoglycerate to form phosphoenolpyruvate (PEP), water being lost from 2-phosphoglycerate in the process. Phosphoenolpyruvate acts as the second source of ATP in glycolysis. The transfer of the phosphate group from PEP to ADP, catalyzed by pyruvate kinase [10], is also highly exergonic and is thus virtually irreversible under physiological conditions.

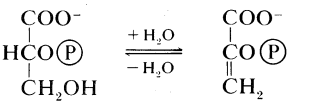
A second source of ATP



[8]

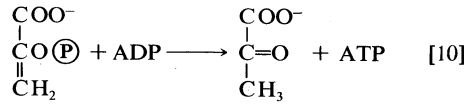
3-phosphoglycerate 2-phosphoglycerate

Reaction [10] occurs twice for each molecule of glucose entering the glycolytic sequence; thus the net yield is two molecules of ATP for each six-carbon sugar. No further molecules of glucose can enter the glycolytic pathway, however, until the $NADH + H^+$ produced in step [6] is reoxidized to NAD^+. In anaerobic systems this means that electrons must be transferred from $(NADH + H^+)$ to some



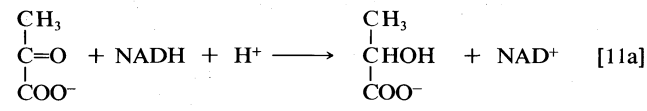
[9]

2-phosphoglycerate phosphoenolpyruvate



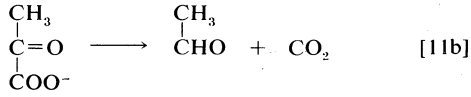
phosphoenolpyruvate pyruvate

organic acceptor molecule, which thus is reduced in the process. Such an acceptor molecule could be the pyruvate formed in reaction [10]. In certain bacteria (*e.g.*, so-called lactic acid bacteria) or in muscle cells functioning vigorously in the absence of adequate supplies of oxygen, pyruvate is reduced to lactate via a reaction catalyzed by lactate dehydrogenase (reaction [11a]); *i.e.*, NADH gives up its hydrogen atoms or electrons to pyruvate, and lactate



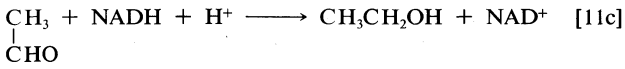
pyruvate lactate

and NAD^+ are formed. Alternatively, in organisms such as brewers' yeast, pyruvate is first decarboxylated to form acetaldehyde and carbon dioxide in a reaction catalyzed by pyruvate decarboxylase [11b]; acetaldehyde then is reduced (by $NADH + H^+$) in a reaction catalyzed by alcohol



pyruvate acetaldehyde

dehydrogenase [11c], yielding ethanol and oxidized coenzyme (NAD^+).

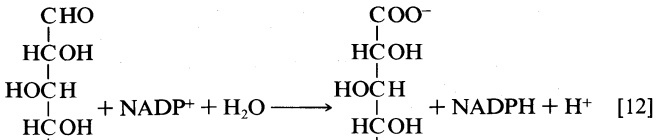


acetaldehyde ethanol

Many variations of reaction [11] occur in nature. In the heterolactic (mixed lactic acid) fermentations carried out by some microorganisms, a mixture of reactions [11a, b, and c] regenerates NAD^+ and results in the production, for each molecule of glucose fermented, of a molecule each of lactate, ethanol, and carbon dioxide. In other types of fermentation, the end products may be derivatives of acids such as propionic, butyric, acetic, and succinic; decarboxylated materials derived from them (*e.g.*, acetone); or compounds such as glycerol.

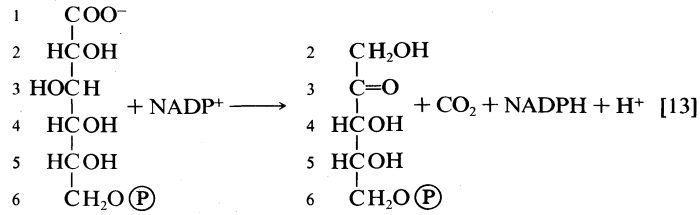
The oxidation of glucose 6-phosphate

The phosphogluconate pathway. Many cells possess, in addition to all or part of the glycolytic pathway that comprises reactions [1] through [11], other pathways of glucose catabolism that involve, as the first unique step, the oxidation of glucose 6-phosphate [12] instead of the formation of fructose 6-phosphate [2]. This is the phosphogluconate pathway, or pentose phosphate cycle. During reaction [12], hydrogen atoms or electrons are removed from the carbon atom at position 1 of glucose 6-phosphate in a reaction catalyzed by glucose 6-phosphate dehydrogenase. The product of the reaction is 6-phosphogluconate. The reducing equivalents (hydrogen atoms or electrons) are accepted by nicotine adenine dinucleotide phosphate ($NADP^+$), a coenzyme similar to but not identical with NAD^+. A second molecule of $NADP^+$ is reduced as 6-phosphogluconate is further oxidized; the reaction is catalyzed by 6-phosphogluconate dehydrogenase [13]. The



glucose 6-phosphate 6-phosphogluconate

products of the reaction also include ribulose 5-phosphate and carbon dioxide. (The numbers at the carbon atoms in step [13] indicate that carbon 1 of 6-phosphogluconate forms carbon dioxide.)



6-phosphogluconate ribulose 5-phosphate

Ribulose 5-phosphate can undergo a series of reactions in which two-carbon and three-carbon fragments are interchanged between a number of sugar phosphates; this sequence of events can lead to the formation of two molecules of fructose 6-phosphate and one of glyceraldehyde 3-phosphate from three molecules of ribulose 5-phosphate (*i.e.*, the conversion of three molecules with five carbons to two with six and one with three). Although the cycle, which is outlined in Figure 4, is the main pathway in microorganisms for fragmentation of pentose sugars, it is not of major importance as a route for the oxidation of glucose. Its primary purpose in most cells is to generate reducing power in the cytoplasm, in the form of reduced $NADP^+$. This function is especially prominent in tissues—such as the liver, mammary gland, fat tissue, and the cortex (outer region) of the adrenal gland—that actively carry out the biosynthesis of fatty acids and other fatty substances (*e.g.*, steroids). A second function of reactions [12] and [13] is to generate from glucose 6-phosphate the pentoses that are used in the synthesis of nucleic acids (see below *The biosynthesis of cell components*).

In photosynthetic organisms, some of the reactions of the phosphogluconate pathway are part of the major route for the formation of sugars from carbon dioxide; in this case, the reactions occur in a direction opposite to that in which they occur in nonphotosynthetic tissues (see PHOTOSYNTHESIS).

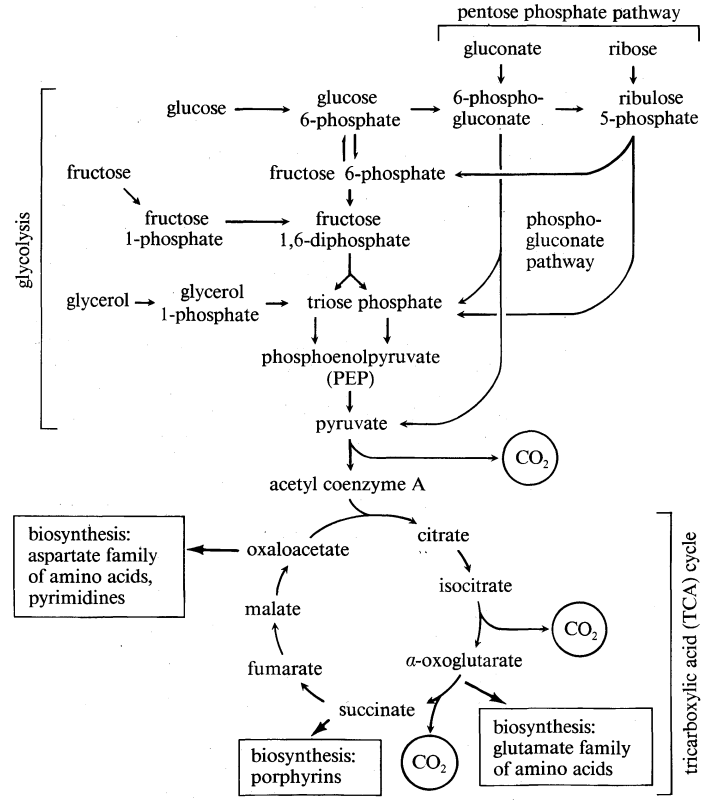
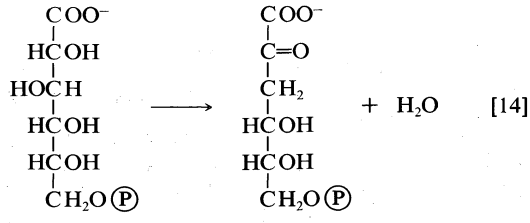


Figure 4: Pathways for the utilization of carbohydrates.

Alternative route

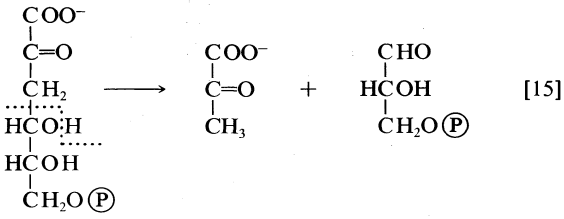
A different route for the catabolism of glucose also involves 6-phosphogluconate; it is of considerable importance in microorganisms lacking some of the enzymes necessary for glycolysis. In this route, 6-phosphogluconate (derived from glucose via steps [1] and [12]) is not oxidized to ribulose 5-phosphate via reaction [13] but, in an enzyme-catalyzed reaction [14], loses water, forming the compound 2-keto-3-deoxy-6-phosphogluconate (KDPG).



[14]

6-phosphogluconate KDPG

This is then split into pyruvate and glyceraldehyde-3-phosphate [15], both of which are intermediates of the glycolytic pathway.

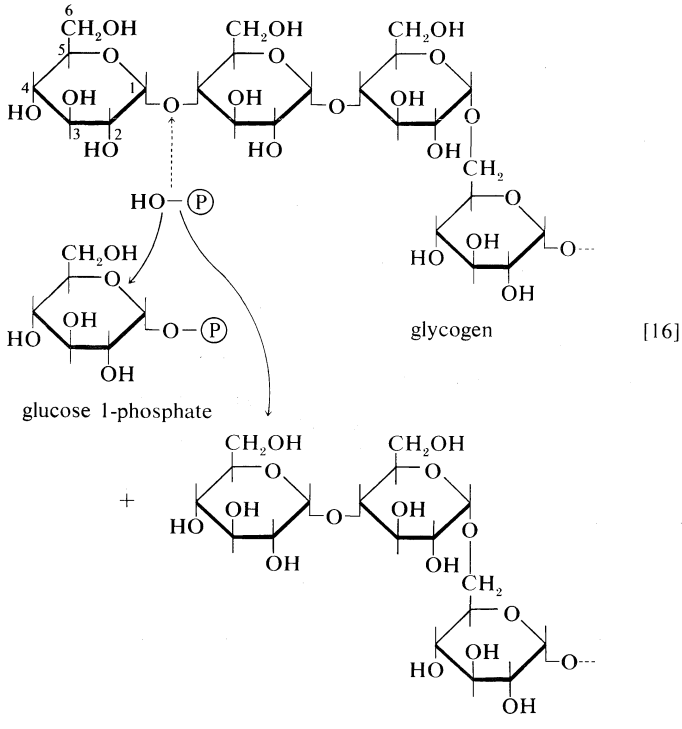


[15]

KDPG pyruvate glyceraldehyde 3-phosphate

THE CATABOLISM OF SUGARS OTHER THAN GLUCOSE

The main storage carbohydrate of animal cells is glycogen, in which chains of glucose molecules—linked end-to-end, the C1 position of one glucose being linked to the C4 position of the adjacent one—are joined to each other by occasional linkages between a carbon at position 1 on one glucose and a carbon at position 6 on another. Two enzymes cooperate in releasing glucose molecules from glycogen. Glycogen phosphorylase catalyzes the splitting of the 1,4-bonds by adding the elements of phosphoric acid at the point shown by the broken arrow in [16], rather than water, as in the digestive hydrolysis of polysaccharides such as glycogen and starch. The products of [16] are glucose 1-phosphate and chains of sugar molecules shortened by one unit; the chains are degraded further by repetition of step [16]. When a bridge linking two chains, at C1 and C6 carbon atoms of adjacent glucose units, is reached, it is hydrolyzed in a reaction involving the enzyme α (1 → 6) glucosidase. After the two chains are separated, reaction [16] can occur again. The glucose 1-phosphate thus formed from glycogen or, in plants, from starch, is converted to glucose 6-phosphate by phosphoglucomutase [78], which catalyzes a reaction very similar to that effected in step [8] of glycolysis; glucose 6-phosphate can then undergo further catabolism via glycolysis [2–10] or via either of the routes involving formation of 6-phosphogluconate [12].



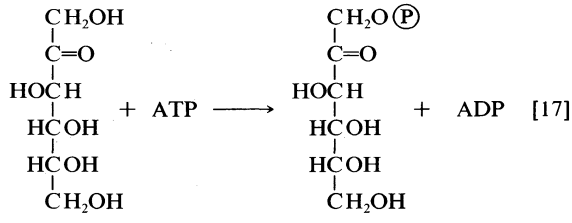
[16]

glycogen

glucose 1-phosphate

Other sugars encountered in the diet are likewise transformed to products that are intermediates of central metabolic pathways. Lactose, or milk sugar, is composed of one molecule of galactose linked to one molecule of glucose. Sucrose, the common sugar of cane or beet, is made up of glucose linked to fructose. Both sucrose and lactose are hydrolyzed to glucose and fructose or galactose, respectively. Glucose is utilized as already described, but special reactions must occur before the other sugars can enter the catabolic routes. Galactose, for example, is phosphorylated in a manner analogous to step [1] of glycolysis. The reaction, catalyzed by a galactokinase, results in the formation of galactose 1-phosphate; this product is transformed to glucose 1-phosphate by a sequence of reactions requiring as a coenzyme uridine triphosphate (UTP). Fructose may also be phosphorylated in animal cells through the action of hexokinase [1], in which case fructose 6-phosphate is the product, or in liver tissue via a fructokinase that gives rise to fructose 1-phosphate [17]. Adenosine triphosphate supplies the phosphate group in both cases.

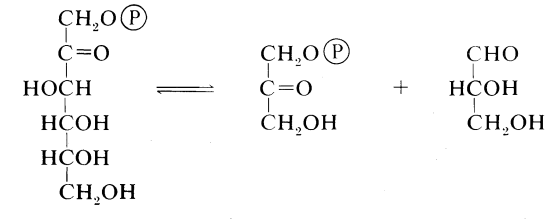
Fragmentation of lactose and sucrose



[17]

fructose fructose 1-phosphate

Fructose 1-phosphate is also formed when facultative anaerobic microorganisms use fructose as a carbon source for growth; in this case, however, the source of the phosphate is phosphoenolpyruvate rather than ATP. Fructose 1-phosphate can be catabolized by one of two routes. In the liver, it is split by an aldolase enzyme [18] abundant in that tissue (but lacking in muscle); the products are dihydroxyacetone phosphate and glyceraldehyde. It will be recalled that dihydroxyacetone phosphate is an intermediate compound of glycolysis. Although glyceraldehyde is not an intermediate of glycolysis, it can be converted to one (glyceraldehyde 3-phosphate) in a reaction involving the conversion of ATP to ADP.



[18]

fructose 1-phosphate dihydroxyacetone phosphate glyceraldehyde

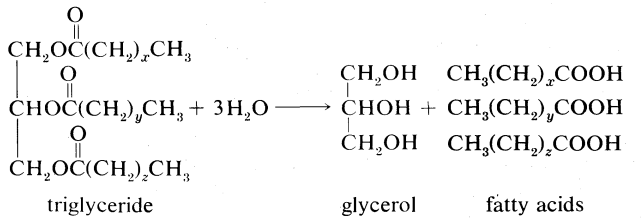
In many organisms other than mammals, fructose 1-phosphate does not have to undergo reaction [18] in order to enter central metabolic routes. Instead, a fructose 1-phosphate kinase, distinct from the phosphofructokinase that catalyzes step [3] of glycolysis, effects the direct conversion of fructose 1-phosphate and ATP to fructose 1,6-diphosphate and ADP.

THE CATABOLISM OF LIPIDS (FATS)

Although carbohydrates are the major fuel for most organisms, fatty acids are also a very important energy source. In vertebrates at least half of the oxidative energy used by the liver, kidneys, heart muscle, and resting skeletal muscle is derived from the oxidation of fatty acids; in fasting or hibernating animals or in migrating birds, fat is virtually the sole source of energy.

Triglycerides

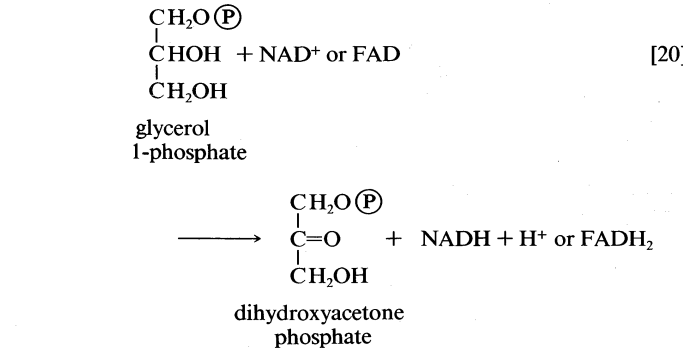
Neutral fats or triglycerides, the major components of storage fats in plant and animal cells, consist of the alcohol glycerol linked to three molecules of fatty acids. Before a molecule of neutral fat can be metabolized, it must be hydrolyzed to its component parts. Hydrolysis [19] is effected by intracellular enzymes or gut enzymes, and forms phase I of fat catabolism. Letters *x*, *y*, and *z* represent the number of $-CH_2-$ groups in the fatty acid molecules.



[19]

As is apparent from [19], the three molecules of fatty acid released from the triglyceride need not be identical; a fatty acid usually contains 16 or 18 carbon atoms but may also be unsaturated; that is, containing one or more double bonds ($-CH=CH-$). Only the fate of saturated fatty acids, of the type $CH_3(CH_2)_nCOOH$ (*n* most commonly is an even number), is dealt with here.

Fate of glycerol. It requires but two reactions to channel glycerol into a catabolic pathway (see Figure 2). In a reaction catalyzed by glycerolkinase, ATP is used to phosphorylate glycerol; the products are glycerol 1-phosphate and ADP. Glycerol 1-phosphate is then oxidized to dihydroxyacetone phosphate [20], an intermediate of glycolysis. The reaction is catalyzed by either a soluble (cytoplasmic) enzyme, glycerolphosphate dehydrogenase, or a similar enzyme present in the mitochondria. In addition to their different locations, the two dehydrogenase enzymes differ in that a different coenzyme accepts the electrons removed from glycerol 1-phosphate. In the case of the cytoplasmic enzyme, NAD^+ accepts the electrons (and is reduced to $NADH + H^+$); in the case of the mitochondrial enzyme, flavin adenine dinucleotide (FAD) accepts the electrons (and is reduced to $FADH_2$).

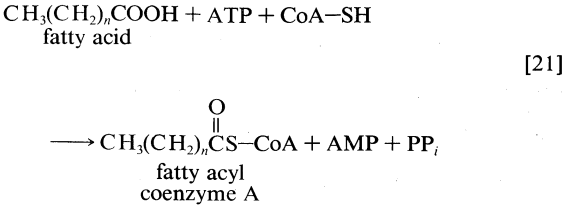


[20]

Fate of fatty acids. As with sugars, the release of energy from fatty acids necessitates an initial investment of ATP. A problem unique to fats is a consequence of the low solubility in water of most fatty acids. Their catabolism requires mechanisms that fragment them in a controlled and stepwise manner. The mechanism involves a coenzyme for the transfer of an acyl group

Coenzyme A and the fragmentation of fats

(*e.g.*, $CH_3\overset{|}{C}=O$), namely, coenzyme A. The functional portion of this complex molecule is the sulfhydryl ($-SH$) group at one end. The coenzyme is often identified as $CoA-SH$ (see step [21]). The organized and stepwise degradation of fatty acids linked to coenzyme A is ensured

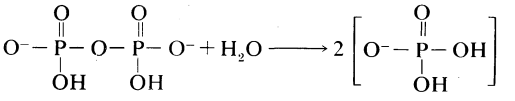


[21]

because the necessary enzymes are sequestered in particulate structures. In microorganisms these enzymes are associated with cell membranes, and in higher organisms with mitochondria.

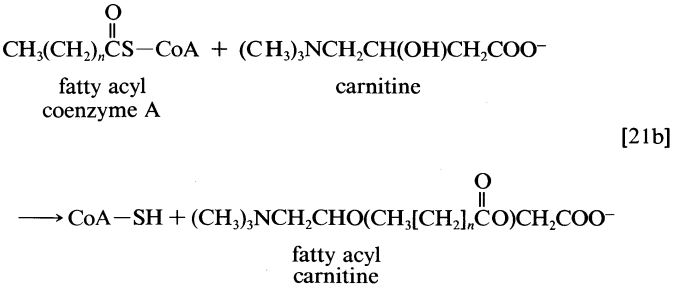
Fatty acids are linked to coenzyme A ($CoA-SH$) in one of two main ways. In higher organisms, enzymes in the cytoplasm called thiokinases catalyze the linkage of fatty acids with $CoA-SH$ to form a compound that can be called a fatty acyl coenzyme A [21]. This step requires ATP, which is split to AMP and inorganic pyrophosphate (PP_i) in the process.

In this series of reactions, *n* indicates the number of hydrocarbon units ($-CH_2-$) in the molecule. Because most tissues contain highly active pyrophosphatase enzymes [21a], which catalyze the virtually irreversible hydrolysis of inorganic pyrophosphate (PP_i) to two molecules of inorganic phosphate (P_i), reaction [21] proceeds overwhelmingly to completion; *i.e.*, from left to right.



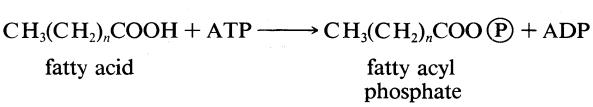
[21a]

Although fatty acids are activated in this way, the acyl coenzyme A derivatives that are formed must be transported to the enzyme complex that effects their oxidation. Activation occurs in the cytoplasm, but, in animal cells, oxidation takes place in the mitochondria. The transfer of fatty acyl coenzyme A across the mitochondrial membrane is effected by the enzyme carnitine, a nitrogen-containing small hydroxy acid of the formula $(CH_3)_3NCH_2CH(OH)CH_2COO^-$. The $-OH$ group within the carnitine molecule accepts the acyl group of fatty acyl coenzyme A, forming acyl carnitine, which can cross the inner membrane of the mitochondrion and there return the acyl group to coenzyme A.



[21b]

These reactions are catalyzed by the enzyme carnitine acyl transferase. Defects in this enzyme or in the carnitine carrier are inborn errors of metabolism. In obligate anaerobic bacteria the linkage of fatty acids to coenzyme A may require the formation of a fatty acyl phosphate, *i.e.*, the phosphorylation of the fatty acid using ATP; ADP is also a product [21c]. The fatty acyl moiety $[CH_3(CH_2)_nCOO^-]$



[21c]

is then transferred to coenzyme A [21d], forming a fatty acyl coenzyme A compound and P_i. In either case, it is the fatty acyl coenzyme A molecules that are fragmented in the sequence of events summarized in Figure 5.

$$CH_3(CH_2)_nCOO\textcircled{P} + CoA{-}SH \longrightarrow CH_3(CH_2)_n\overset{O}{\overset{\|}{C}}S{-}CoA + P_i \quad [21d]$$

fatty acyl coenzyme A

Unsaturated fatty acyl coenzyme A

Initially (step [22], Figure 5) two hydrogen atoms are lost from the fatty acyl coenzyme A, resulting in the formation of an unsaturated fatty acyl coenzyme A (*i.e.*, with a double bond, $-CH{=}CH-$) between the α- and β-carbons of the acyl moiety. (The α-carbon is the one closest to the carboxyl [$-COOH$] group of a fatty acid; the next closest is the β-, and so on to the end of the hydrocarbon chain.) The hydrogen atoms are accepted by the coenzyme FAD (flavin adenine dinucleotide), which is reduced to $FADH_2$. The product of step [22], α,β-unsaturated fatty acyl coenzyme A, is enzymatically hydrated [23]; *i.e.*, water is added across the double bond. The product, called a β-hydroxyacyl coenzyme A, can again be oxidized in an enzyme-catalyzed reaction [24]; the electrons removed are accepted by NAD^+. The product is called a β-ketoacyl coenzyme A.

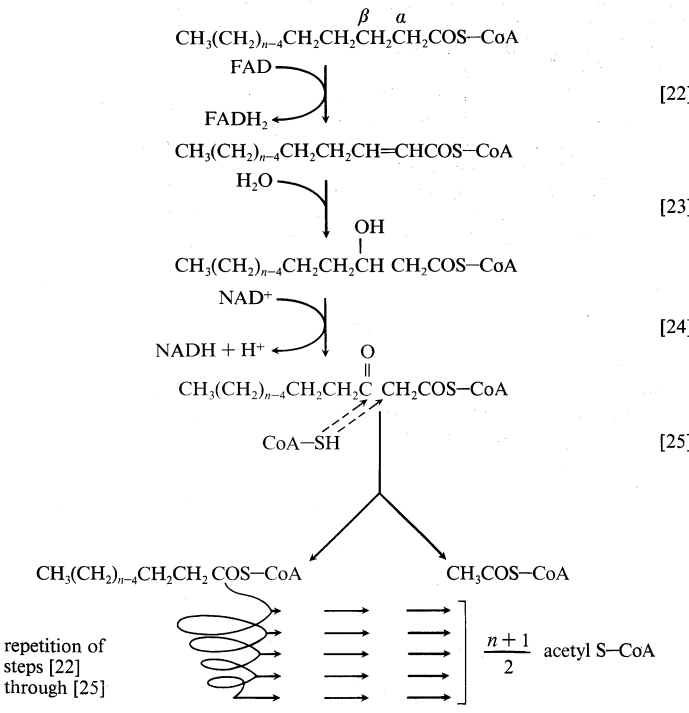


Figure 5: Fragmentation of acyl coenzyme A with an even number of carbon atoms (see text).

The next enzymatic step [25] enables the energy invested in step [21] to be conserved. The β-ketoacyl coenzyme A that is the product of reaction [24] is split, not by water but by coenzyme A. The process, called thiolysis (as distinct from hydrolysis), yields the two-carbon fragment acetyl coenzyme A and a fatty acyl coenzyme A having two fewer carbon atoms than the molecule that underwent reaction [22]; otherwise the two are similar.

The shortened fatty acyl coenzyme A molecule now undergoes the sequence of reactions again, beginning with the dehydrogenation step [22], and another two-carbon fragment is removed as acetyl coenzyme A. With each passage through the process of fatty acid oxidation, the fatty acid loses a two-carbon fragment as acetyl coenzyme A and two pairs of hydrogen atoms to specific acceptors. The 16-carbon fatty acid, palmitic acid, for example, undergoes a total of seven such cycles, yielding eight molecules of acetyl coenzyme A and 14 pairs of hydrogen atoms, seven of which appear in the form of $FADH_2$ and seven in the form of $NADH + H^+$. The reduced coenzymes, $FADH_2$ and reduced NAD^+, are reoxidized when the electrons pass through the electron transport chain, with concomitant formation of ATP (see below *Biological energy transduction*). In anaerobes, organic molecules and not oxygen are electron acceptors; thus the yield of ATP is reduced. In all organisms, however, the acetyl coenzyme A formed from the breakdown of fatty acids joins that arising from the catabolism of carbohydrates (see below *The oxidation of pyruvate*) and many amino acids (see below *The catabolism of proteins*); Figure 2 shows the interrelationships.

Fatty acids with an odd number of carbon atoms are relatively rare in nature but may arise during microbial fermentations or through the oxidation of amino acids such as valine and isoleucine. They may be fragmented through repeated cycles of steps [22] to [25] until the final five-carbon acyl coenzyme A is split into acetyl coenzyme A and propionyl coenzyme A, which has three carbon atoms. In many bacteria this propionyl coenzyme A can be transformed either to acetyl coenzyme A and carbon dioxide or to pyruvate. In other microorganisms and in animals propionyl coenzyme A has a different fate: carbon dioxide is added to propionyl coenzyme A in a reaction requiring ATP. The product, methylmalonyl coenzyme A, has four carbon atoms; the molecule undergoes a rearrangement, forming succinyl coenzyme A, which is an intermediate of the TCA cycle.

THE CATABOLISM OF PROTEINS

Role of amino acids

The amino acids derived from proteins function primarily as the precursors, or building blocks, for the cell's own proteins and (unlike lipids and carbohydrates) are not primarily a source of energy. Many microorganisms, on the other hand, can grow by using amino acids as the sole carbon and nitrogen source. Under these conditions these microorganisms derive from the amino acids all of their required energy and all of the precursors of the macromolecules that comprise the components of their cells. Moreover, it has been calculated that a man of average weight (70 kilograms, or 154 pounds) turns over about 0.4 kilogram of protein per day. About 0.1 kilogram is degraded and replaced by dietary amino acids; the remaining 0.3 kilogram is recycled as part of the dynamic state of cell constituents. The cells of plants contain and metabolize many amino acids in addition to the 20 or so that are normally found in proteins. A complete discussion of these special pathways is outside the scope of this article, however.

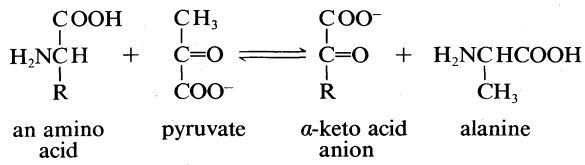
Before proteins can enter cells, the bonds linking adjacent amino acids (peptide bonds) must be hydrolyzed; this process releases the amino acids constituting the protein. The utilization of dietary proteins thus requires the operation of extracellular digestive enzymes; *i.e.*, enzymes outside the cell. Many microorganisms secrete such enzymes into the nutrient media in which they are growing; animals secrete them into the gut. The turnover of proteins within cells, on the other hand, requires the functioning of intracellular enzymes that catalyze the splitting of the peptide bonds linking adjacent amino acids; little is known about the mechanism involved.

Amino acids may be described by the general formula $RCH(NH_2)COOH$, or $RCH(NH_3^+)COO^-$, in which R represents a specific chemical moiety. The catabolic fate of amino acids involves (1) removal of nitrogen, (2) disposal of nitrogen, and (3) oxidation of the remaining carbon skeleton.

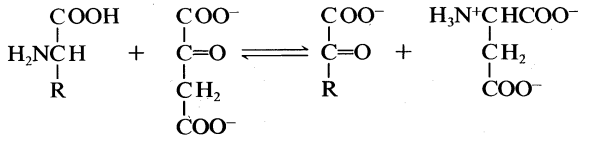
Removal of nitrogen. The removal of the amino group ($-NH_2$) generally constitutes the first stage in amino-acid catabolism. The amino group usually is initially transferred to the anion of one of three different α-keto acids (*i.e.*, of the general structure $RCOCOO^-$): pyruvate, which is an intermediate of carbohydrate fragmentation; or oxaloacetate or α-oxoglutarate, both intermediates of the TCA cycle. The products are alanine, aspartate, and glutamate (reactions [26a, b, and c]).

Since the effect of these reactions is to produce n amino acids and n keto acids from n different amino acids and n different keto acids, no net reduction in the nitrogen content of the system has yet been achieved. The elimination of nitrogen occurs in a variety of ways.

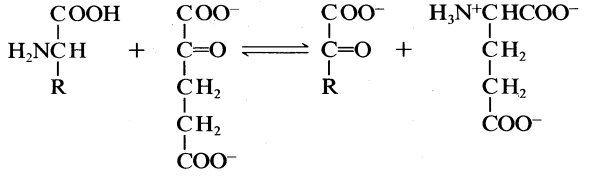
Elimination of nitrogen



an amino acid | pyruvate | α-keto acid anion | alanine [26a]

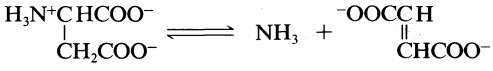


an amino acid | oxalo-acetate | α-keto acid anion | aspartate [26b]



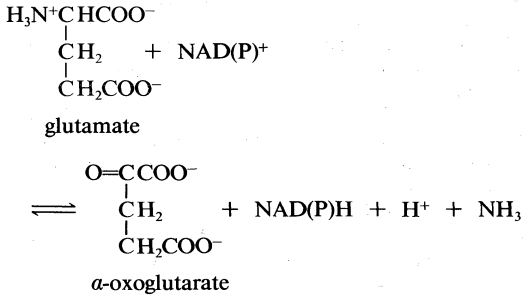
an amino acid | α-oxogutarate | α-keto acid anion | glutamate [26c]

In many microorganisms, ammonia (NH_3) can be removed from aspartate via a reaction catalyzed by aspartase [27]; the other product, fumarate, is an intermediate of the TCA cycle.



aspartate | fumarate [27]

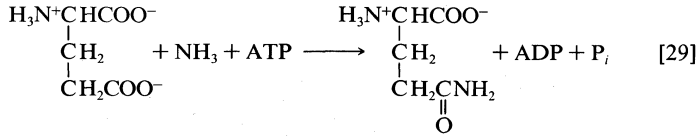
A quantitatively more important route is that catalyzed by glutamate dehydrogenase, in which the glutamate formed in [26c] is oxidized to α-oxoglutarate, another TCA cycle intermediate [28]. Either $NADP^+$ or both $NADP^+$ and NAD^+ may serve as the hydrogen or electron acceptor, depending on the organism; and some organisms synthesize two enzymes, one of which prefers $NADP^+$ and the other NAD^+. In reaction [28], $NAD(P)^+$ is used to indicate that either NAD^+, $NADP^+$, or both may serve as the electron acceptor.



glutamate | α-oxoglutarate [28]

The occurrence of the transfer reactions [26] and either step [27] or, more importantly, step [28] allows the channeling of many amino acids into a common pathway by which nitrogen can be eliminated as ammonia.

Disposal of nitrogen. In animals that excrete ammonia as the main nitrogenous waste product (*e.g.*, some marine invertebrates, crustaceans), it is derived from nitrogen transfer reactions [26] and oxidation via glutamate dehydrogenase [28] as described above for microorganisms. Because ammonia is toxic to cells, however, it is detoxified as it forms. This process involves an enzyme-catalyzed reaction between ammonia and a molecule of glutamate; ATP provides the energy for the reaction, which results in the formation of glutamine, ADP, and inorganic phosphate [29]. This reaction [29] is catalyzed by glutamine synthetase, which is subject to a variety of metabolic controls. The glutamine thus formed gives up the amide nitrogen in the kidney tubules. As a result glutamate is formed once again, and ammonia is released into the urine.

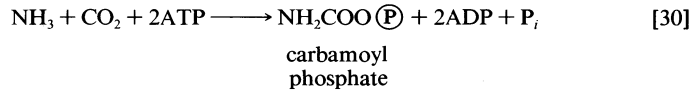


glutamate | glutamine [29]

Uric acid

In terrestrial reptiles and birds, uric acid rather than glutamate is the compound with which nitrogen combines to form a nontoxic substance for transfer to the kidney tubules. Uric acid is formed by a complex pathway that begins with ribose 5-phosphate and during which a so-called purine skeleton (see Figure 11) is formed; in the course of this process, nitrogen atoms from glutamine and the amino acids aspartic acid and glycine are incorporated into the skeleton. These nitrogen donors are derived from other amino acids via amino group transfer [26] and the reaction catalyzed by glutamine synthetase [29].

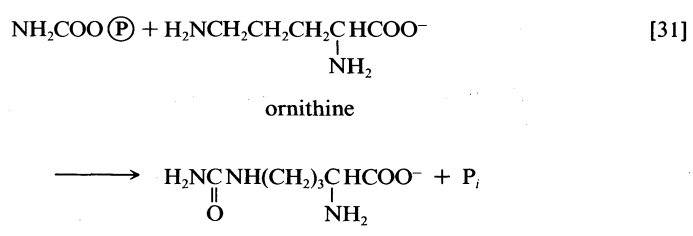
In most fishes, amphibians, and mammals, nitrogen is detoxified in the liver and excreted as urea, a readily soluble and harmless product. The sequence leading to the formation of urea, commonly called the urea cycle, is summarized as follows: Ammonia, formed from glutamate and NAD^+ in the liver mitochondria (reaction [28]), reacts with carbon dioxide and ATP to form carbamoyl phosphate, ADP, and inorganic phosphate, as shown in reaction [30]. The reaction is catalyzed by carbamoyl phosphate synthetase. The carbamoyl moiety of carbamoyl phosphate



carbamoyl phosphate [30]

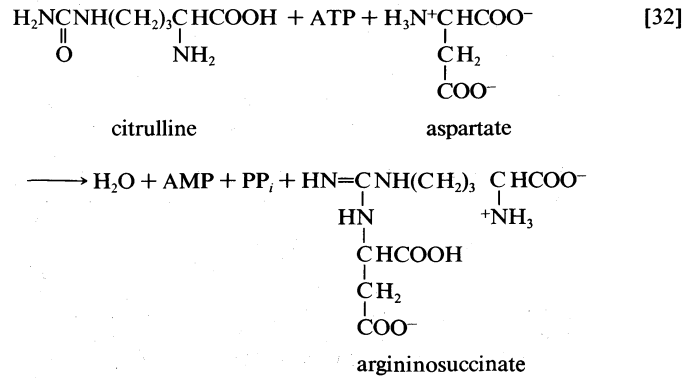
(NH_2CO-) is transferred to ornithine, an amino acid, in a reaction catalyzed by ornithine transcarbamoylase; the products are citrulline and inorganic phosphate [31].

Citrulline

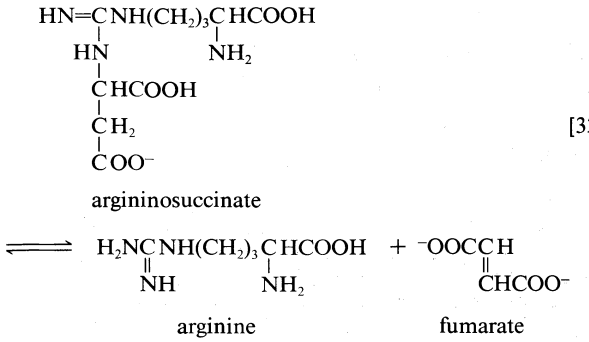


ornithine | citrulline [31]

Citrulline and aspartate formed from amino acids via step [26b] react to form argininosuccinate [32]; argininosuccinic acid synthetase catalyzes the reaction. Argininosuccinate splits into fumarate and arginine during a reaction catalyzed by argininosuccinase [32a]. In the final step



citrulline | aspartate | argininosuccinate [32]

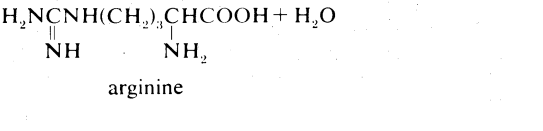


argininosuccinate

[32a]

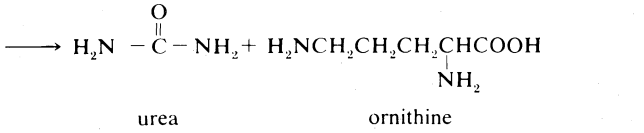
arginine fumarate

of the urea cycle, arginine, in a reaction catalyzed by arginase, is hydrolyzed [33]. Urea and ornithine are the products; ornithine thus is available to initiate another cycle beginning at step [31].



arginine

[33]



urea ornithine

Oxidation of the carbon skeleton. As indicated in Figure 2, the carbon skeletons of amino acids (*i.e.*, the portion of the molecule remaining after the removal of nitrogen) are fragmented to form only a few end products; all of them are intermediates of either glycolysis or the TCA cycle. The number and complexity of the catabolic steps by which each amino acid arrives at its catabolic end point reflects the chemical complexity of that amino acid. Thus, in the case of alanine, only the amino group must be removed to yield pyruvate; the amino acid threonine, on the other hand, must be transformed successively to the amino acids glycine and serine before pyruvate is formed. The fragmentation of leucine to acetyl coenzyme A involves seven steps; that of tryptophan to the same end product requires 11. (A detailed discussion of the events that enable each of the 20 commonly occurring amino acids to enter central metabolic pathways is beyond the scope of this article.)

End products as intermediates

The combustion of food materials

Although the pathways for fragmentation of food materials effect the conversion of a large variety of relatively complex starting materials into only a few simpler intermediates of central metabolic routes—mainly pyruvate, acetyl coenzyme A, and a few intermediates of the TCA cycle—their operation releases but a fraction of the energy contained in the materials. The reason is that, in the fermentation process, catabolic intermediates serve also as the terminal acceptors of the reducing equivalents (hydrogen atoms or electrons) that are removed during the oxidation of food; the end products thus may be at the same oxidation level and may contain equivalent numbers of carbon, hydrogen, and oxygen atoms, as the material that was catabolized by a fermentative route. The necessity for pyruvate, for example, to act as hydrogen acceptor in the fermentation of glucose to lactate (see reactions [1] through [11a]) results in the conservation of all the component atoms of the glucose molecule in the form of lactate. The consequent release of energy as ATP (in steps [7] and [10]) is thus small.

A more favourable situation arises if the reducing equivalents formed by oxidation of nutrients can be passed on to an inorganic acceptor such as oxygen. In this case, the products of fermentation need not act as "hydrogen sinks," in which the energy in the molecule is lost when they leave the cell; instead, the products of fermentation can be degraded further, during phase III of catabolism, and all the usable chemical energy of the nutrient can be transformed into ATP.

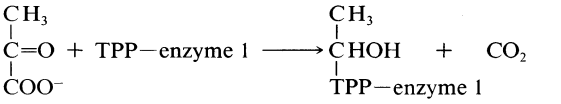
Importance of complete oxidation

This section describes the manner in which the products obtained by the fragmentation of nutrients are oxidized (*i.e.*, the manner in which hydrogen atoms or electrons are removed from them) and the manner in which these reducing equivalents react with oxygen, with concomitant formation of ATP.

THE OXIDATION OF MOLECULAR FRAGMENTS

The oxidation of pyruvate. The oxidation of pyruvate involves the concerted action of several enzymes and coenzymes collectively called the pyruvate dehydrogenase complex; *i.e.*, a multienzyme complex in which the substrates are passed consecutively from one enzyme to the next, and the product of the reaction catalyzed by the first enzyme immediately becomes the substrate for the second enzyme in the complex. The overall reaction is the formation of acetyl coenzyme A and carbon dioxide from pyruvate, with concomitant liberation of two reducing equivalents in the form of NADH + H^+. The individual reactions that result in the formation of these end products are as follows.

Pyruvate first reacts with the coenzyme of pyruvic acid decarboxylase (enzyme 1), thiamine pyrophosphate (TPP); in addition to carbon dioxide a hydroxyethyl–TPP–enzyme complex ("active acetaldehyde") is formed [34]. Thiamine is vitamin B_1; the biological role of TPP was first revealed by the inability of vitamin B_1-deficient animals to oxidize pyruvate.

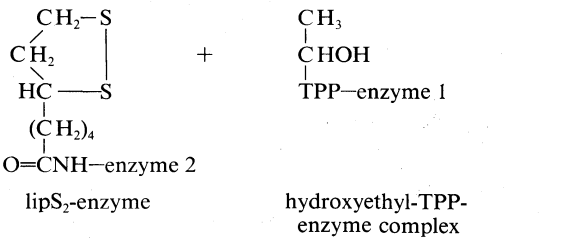


pyruvate hydroxyethyl-TPP-enzyme complex

[34]

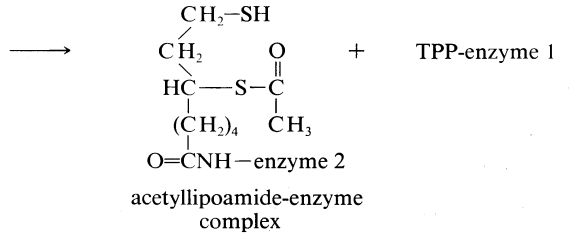
The hydroxyethyl moiety formed in [34] is immediately transferred to one of the two sulfur atoms (S) of the coenzyme (6,8-dithio-*n*-octanoate or $lipS_2$) of the second enzyme in the complex, dihydrolipoyl transacetylase (enzyme 2). The hydroxyethyl group attaches to $lipS_2$ at one of its sulfur atoms, as shown in [35]; the result is that coenzyme $lipS_2$ is reduced and the hydroxyethyl moiety is oxidized.

Enzyme 2



$lipS_2$-enzyme hydroxyethyl-TPP-enzyme complex

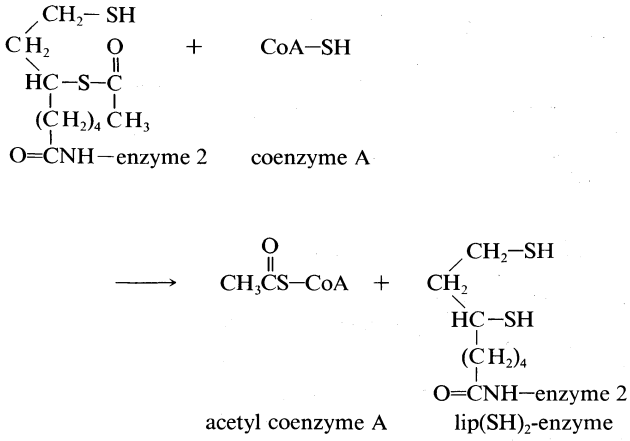
[35]



acetyllipoamide-enzyme complex

The acetyl group ($CH_3C{=}O$) then is transferred to the sulfhydryl (−SH) group of coenzyme A, thereby completing the oxidation of pyruvate (reaction [36]).

The coenzyme $lipS_2$ that accepted the hydroxyethyl moiety in step [35] of the sequence, now reduced, must be reoxidized before another molecule of pyruvate can be oxidized. The reoxidation of the coenzyme is achieved by the enzyme-catalyzed transfer of two reducing equivalents initially to the coenzyme flavin adenine dinucleotide

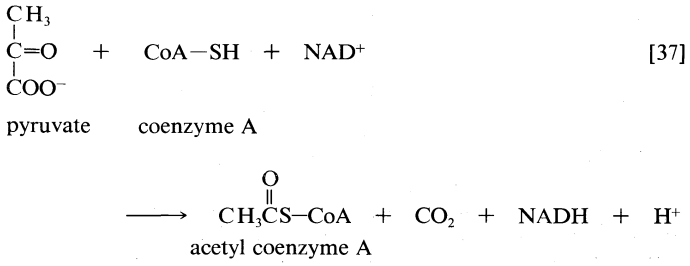


[36]

acetyl coenzyme A lip(SH)$_2$-enzyme

(FAD) and thence to the NAD^+ that is the first carrier in the so-called electron transport chain. The passage of two such reducing equivalents from reduced NAD^+ to oxygen is accompanied by the formation of three molecules of ATP (see *Biological energy transduction*).

The overall reaction may be written as shown in [37], in which pyruvate reacts with coenzyme A in the presence of TPP and lipS_2 to form acetyl coenzyme A and carbon dioxide, and to liberate two hydrogen atoms (in the form of NADH + H^+) that can subsequently yield energy by the reduction of oxygen to water. The lipS_2 reduced during this process is reoxidized in the presence of the enzyme lipoyl dehydrogenase, with the concomitant reduction of NAD^+.

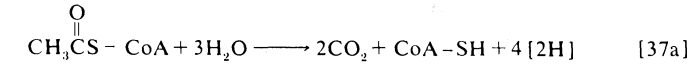


[37]

pyruvate coenzyme A acetyl coenzyme A

Acetyl coenzyme A

The tricarboxylic acid (TCA) cycle. Acetyl coenzyme A arises not only from the oxidation of pyruvate but also from that of fats (Figure 5) and many of the amino acids comprising proteins (Figure 2); the sequence of enzyme-catalyzed steps that effects the total combustion of the acetyl moiety of the coenzyme represents the terminal oxidative pathway for virtually all food materials. The balance of the overall reaction of the TCA cycle [37a] is that three molecules of water react with acetyl coenzyme A to form carbon dioxide, coenzyme A, and reducing equivalents. The oxidation by oxygen of the reducing equivalents is accompanied by the conservation (as ATP) of most of the energy of the food ingested by aerobic organisms.

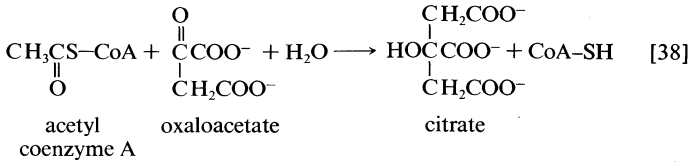


[37a]

acetyl coenzyme A reducing equivalents

The relative complexity and number of chemical events that constitute the TCA cycle, and their location as components of spatially determined structures such as cell membranes in microorganisms and mitochondria in plants and higher animals, reflect the problems involved chemically in "dismembering" a compound having only two carbon atoms and releasing in a controlled and stepwise manner the reducing equivalents ultimately to be passed to oxygen. These problems have been overcome by the simple but effective device of initially combining the two-carbon compound with a four-carbon acceptor; it is much less difficult chemically to dismember and oxidize a compound having six carbon atoms.

In the TCA cycle, acetyl coenzyme A initially reacts with

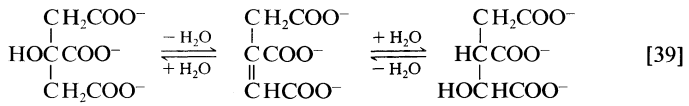


[38]

acetyl coenzyme A oxaloacetate citrate

oxaloacetate to yield citrate and to liberate coenzyme A. This reaction [38] is catalyzed by citrate synthase. (As mentioned above, many of the compounds in living cells that take part in metabolic pathways exist as charged moieties, or anions, and are named as such.) Citrate undergoes isomerization (*i.e.*, a rearrangement of certain atoms comprising the molecule) to form isocitrate [39]. The reaction involves first the removal of the elements of water from citrate to form *cis*-aconitate, and then the re-addition of water to *cis*-aconitate in such a way that isocitrate is formed. It is probable that all three reactants—citrate, *cis*-aconitate, and isocitrate—remain closely associated with aconitase, the enzyme that catalyzes the isomerization process, and that most of the *cis*-aconitate is not released from the enzyme surface but is immediately converted to isocitrate.

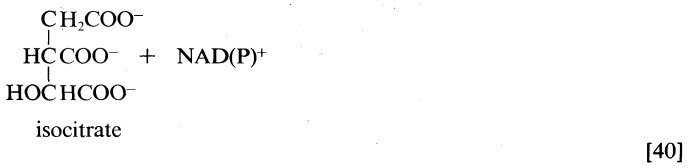
Isomerization



[39]

citrate *cis*-aconitate isocitrate

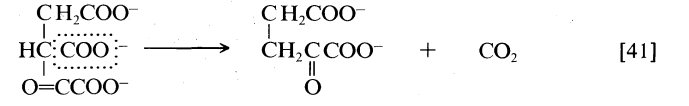
Isocitrate is oxidized—*i.e.*, hydrogen is removed—to form oxalosuccinate; the two hydrogen atoms are usually transferred to NAD^+, thus forming reduced NAD^+ [40]; in some microorganisms, and during the biosynthesis of glutamate in the cytoplasm of animal cells, however, the hydrogen atoms may also be accepted by $NADP^+$. Thus the enzyme controlling this reaction, isocitrate dehydrogenase, differs in specificity for the coenzymes; various forms occur not only in different organisms but even within the same cell. In [40] $NAD(P)^+$ indicates that either NAD^+ or $NADP^+$ can act as a hydrogen acceptor.



[40]

isocitrate oxalosuccinate

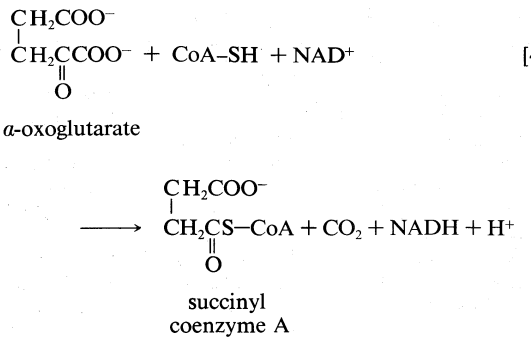
The position of the carboxylate ($-COO^-$) that is "sandwiched" in the middle of the oxalosuccinate molecule renders it very unstable; as a result, the carbon of this group is lost as carbon dioxide (note the dotted rectangle) in a reaction [41] that can occur spontaneously but may be further accelerated by an enzyme.



[41]

oxalosuccinate α-oxoglutarate

The five-carbon product of reaction [41], α-oxoglutarate, has chemical properties similar to pyruvate (free-acid forms of both are so-called α-oxoacids), and the chemical events involved in the oxidation of α-oxoglutarate are analogous to those already described for the oxidation of pyruvate (see reaction [37]). Reaction [42] is effected by a multienzyme complex; TPP, lipS_2 (6,8-dithio-*n*-octanoate), and coenzyme A are required as coenzymes. The products are carbon dioxide and succinyl coenzyme A. As was noted



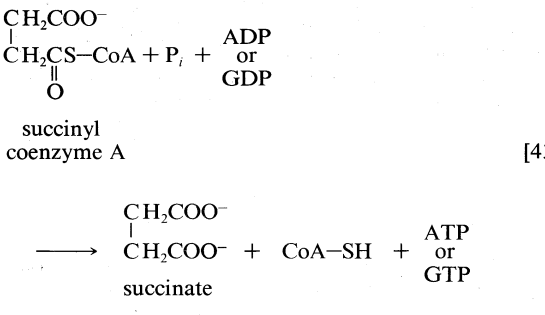
α-oxoglutarate

succinyl coenzyme A

with reaction [37], this oxidation of α-oxoglutarate results in the reduction of $lipS_2$, which must be reoxidized. This is done by transfer of reducing equivalents to FAD and thence to NAD^+. The resultant $NADH + H^+$ is reoxidized by the passage of the electrons, ultimately, to oxygen, via the electron transport chain.

Phosphorolysis

Unlike the acetyl coenzyme A produced from pyruvate in reaction [37], succinyl coenzyme A undergoes a phosphorolysis reaction—*i.e.*, transfer of the succinyl moiety from coenzyme A to inorganic phosphate. The succinyl phosphate thus formed is not released from the enzyme surface; an unstable, high-energy compound called an acid anhydride, it transfers a high-energy phosphate to



succinyl coenzyme A

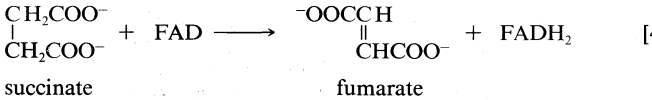
[43]

succinate

ADP, directly or via guanosine diphosphate (GDP) [43]. If guanosine triphosphate (GTP) forms, ATP can readily arise from it in an exchange involving ADP [43a]:

$$GTP + ADP \rightleftharpoons ATP + GDP \qquad [43a]$$

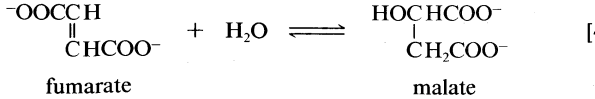
The remainder of the reactions of the TCA cycle serve to regenerate the initial four-carbon acceptor of acetyl coenzyme A (oxaloacetate) from succinate, the process requiring in effect the oxidation of a methylene group ($-CH_2-$) to a carbonyl group ($-CO-$), with concomitant release of $2 \times [2H]$ reducing equivalents. It is therefore similar to, and is effected in like manner to, the oxidation of fatty acids (steps [22–24]; see Figure 5). As is the case with fatty acids, hydrogen atoms or electrons are initially removed from the succinate formed in [43] and are accepted by FAD; the reaction, catalyzed by succinate dehydrogenase [44], results in the formation of fumarate and reduced FAD.



succinate fumarate

[44]

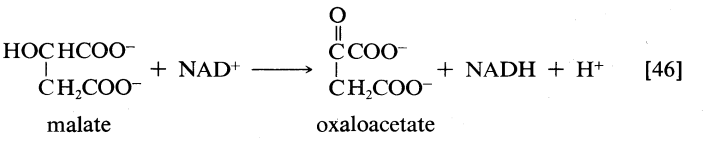
The elements of water are added across the double bond ($-CH{=}CH-$) of fumarate in a reaction catalyzed by fumarase [45]; this type of reaction also occurred in step [39] of the cycle. The product of reaction [45] is malate.



fumarate malate

[45]

Malate can be oxidized to oxaloacetate by removal of two hydrogen atoms, which are accepted by NAD^+. This type of reaction, catalyzed by malate dehydrogenase in reaction [46], also occurred in step [40] of the cycle. The formation of oxaloacetate completes the TCA cycle, which can now begin again with the formation of citrate [38].



malate oxaloacetate

[46]

Energy conserved during the TCA cycle

The loss of the two molecules of carbon dioxide in steps [41] and [42] does not yield biologically useful energy. The substrate-linked formation of ATP accompanies step [43], in which one molecule of of ATP, is formed during each turn of the cycle. The hydrogen ions and electrons that result from steps [40], [42], [44], and [46] are passed down the chain of respiratory carriers to oxygen, with the concomitant formation of three molecules of ATP, per [2H], as $NADH + H^+$ (see below). Similarly, the oxidation of the reduced FAD formed in [44] results in the formation of two ATP. Each turn of the cycle thus leads to the production of a total of 12 ATP. It will be recalled that the anaerobic fragmentation of glucose to two molecules of pyruvate yielded two ATP; the aerobic oxidation via the TCA cycle of two molecules of pyruvate thus makes available to the cell at least 15 times more ATP per molecule of glucose catabolized than is produced anaerobically. If, in addition, the $2 \times [NADH + H^+]$ generated per glucose in reaction [6] are passed on to oxygen, a further six ATP are generated. The advantage to living organisms is to be able to respire rather than merely to ferment.

BIOLOGICAL ENERGY TRANSDUCTION

Adenosine triphosphate as the currency of energy exchange. When the terminal phosphate group is removed from ATP by hydrolysis, two negatively charged products are formed, ADP^{3-} and HPO_4^{2-} (a phosphate group) [47].

$$ATP^{4-} + H_2O \longrightarrow ADP^{3-} + HPO_4^{2-} + H^+ \qquad [47]$$

These products are electrically more stable than the parent molecule and do not readily recombine. The total free energy (G) of the products is much less than that of ATP; hence energy is liberated (*i.e.*, the reaction is exergonic).

Standard free energy

The amount of energy liberated under strictly defined conditions is called the standard free energy change ($\Delta G'$); this value for the hydrolysis of ATP is relatively high, at −8 kilocalories per mole. (One kilocalorie is the amount of heat required to raise the temperature of 1,000 grams of water one degree centigrade.) Conversely, the formation of ATP from ADP and inorganic phosphate (P_i) is an energy-requiring (*i.e.*, endergonic) reaction with a standard free energy change of +8 kilocalories per mole.

The hydrolysis of the remaining phosphate-to-phosphate bond of ADP is also accompanied by a considerable liberation of free energy (the standard free energy change is −6.5 kilocalories per mole); AMP hydrolysis liberates less energy (the standard free energy change is −2.2 kilocalories per mole).

The free energy of hydrolysis of a compound thus is a measure of the difference in energy content between the starting substances (reactants) and the final substances (products). Adenosine triphosphate does not have the highest standard free energy of hydrolysis of all the naturally occurring phosphates but instead occupies a position at approximately the halfway point in a series of phosphate compounds with a wide range of standard free energies of hydrolysis. Compounds such as 1,3-diphosphoglycerate or phosphoenolpyruvate (PEP), which are above ATP on the scale (see Figure 6), have large negative $\Delta G'$ values on hydrolysis and are often called high-energy phosphates; they are said to exhibit a high phosphate group transfer potential because they have a tendency to lose their phosphate groups. Compounds such as glucose 6-phosphate or fructose 6-phosphate, which are below ATP on the scale because they have smaller negative $\Delta G'$ values on hydrolysis, have a tendency to hold on to their phosphate groups and thus act as low-energy phosphate acceptors.

Both ATP and ADP act as intermediate carriers for the transfer of phosphate groups (which are more precisely

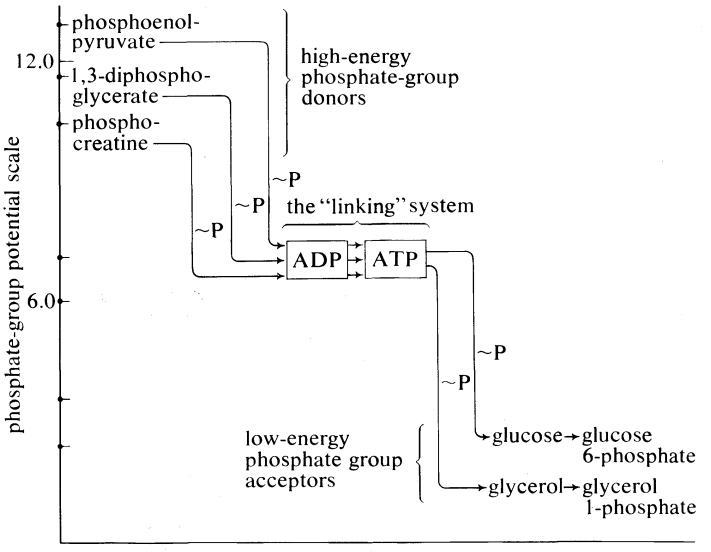


Figure 6: The transfer of phosphate groups from high-energy donors to low-energy acceptors by way of the ATP–ADP system (see text).

called phosphoryl groups) and hence of energy, from compounds lying above ATP to those lying beneath it. Thus, in glycolysis, ADP acts as an acceptor of a phosphate group during the synthesis of ATP from PEP (see reaction [10]), and ATP functions as a donor of a phosphate group during the formation of fructose 1,6-diphosphate from fructose 6-phosphate (see reaction [3]).

ATP coupling

The first step in glycolysis, the formation of glucose 6-phosphate (G6P), illustrates how an energetically unfavourable reaction may become feasible under intracellular conditions by coupling it to ATP.

$$\text{glucose} + P_i \longrightarrow \text{G6P} \qquad \Delta G' = +3.3 \text{ kcal} \qquad [48]$$

$$\text{ATP} \longrightarrow \text{ADP} + P_i \qquad \Delta G' = -7.3 \text{ kcal} \qquad [48a]$$

$$\text{glucose} + \text{ATP} \longrightarrow \text{G6P} + \text{ADP} \qquad \Delta G' = -4 \text{ kcal} \qquad [48b]$$

Reaction [48] has a positive $\Delta G'$ value, indicating that the reaction tends to proceed in the reverse direction. It is therefore necessary to use the standard free energy generated by the breaking of the first phosphate bond in ATP (reaction [48a]), which is −7.3 kilocalories per mole, to move reaction [48] in the forward direction. Combining these reactions and their standard free energies gives reaction [48b] and a standard free energy value of −4 kilocalories per mole, indicating that the reaction will proceed in the forward direction. There are many intracellular reactions in which the formation of ADP or AMP from ATP provides energy for otherwise unfavourable biosyntheses. Some cellular reactions use equivalent phosphorylated analogues of ATP, for example, guanosine triphosphate (GTP) for protein synthesis and cytidine triphosphate (CTP) for lipid synthesis.

The utilization of ATP to perform work

The function of ATP as a common intermediate of energy transfer during anabolism is further dealt with below (see *The biosynthesis of cell components*). In certain specialized cells or tissues the chemical energy of ATP is used to perform work other than the chemical work of anabolism; for example, mechanical work—such as muscular contraction, or the movement of contractile structures called cilia and flagella, which are responsible for the motility of many small organisms. The performance of osmotic work also requires ATP; *e.g.*, the transport of ions or metabolites through membranes against a concentration gradient, a process that is basically responsible for many physiological functions, including nerve conduction, the secretion of hydrochloric acid in the stomach, and the removal of water from the kidneys.

Energy conservation. The amount of ATP in a cell is limited, and it must be replaced continually to maintain repair and growth. This is achieved by using the energy liberated during the oxidative stages of catabolism to synthesize ATP from ADP and phosphate. The synthesis of ATP linked to catabolism occurs by two distinct mechanisms: substrate-level phosphorylation and oxidative, or respiratory-chain, phosphorylation. Oxidative phosphorylation is the major method of energy conservation under aerobic conditions in all nonphotosynthetic cells.

Substrate-level phosphorylation. In substrate-level phosphorylation a phosphoryl group is transferred from an energy-rich donor (*e.g.*, 1,3-diphosphoglycerate) to ADP to yield a molecule of ATP. This type of ATP synthesis (see reactions [7], [10], and [43]) does not require molecular oxygen (O_2), although it is frequently, but not always, preceded by an oxidation (*i.e.*, dehydrogenation) reaction. Substrate-level phosphorylation is the major method of energy conservation in oxygen-depleted tissues and during fermentative growth of microorganisms.

Oxidative, or respiratory-chain, phosphorylation. In oxidative phosphorylation the oxidation of catabolic intermediates by molecular oxygen occurs via a highly ordered series of substances that act as hydrogen and electron carriers. They constitute the electron transfer system, or respiratory chain. In most animals, plants, and fungi, the electron transfer system is fixed in the membranes of mitochondria; in bacteria (which have no mitochondria) this system is incorporated into the plasma membrane. Sufficient free energy is released to allow the synthesis of ATP by a process described below. First, however, it is necessary to consider the nature of the respiratory chain.

The electron transfer system

Four types of hydrogen or electron carriers are known to participate in the respiratory chain, in which they serve to transfer two reducing equivalents (2H) from reduced substrate (AH_2) to molecular oxygen (see reaction [49]); the products are the oxidized substrate (A) and water (H_2O).

$$AH_2 + \tfrac{1}{2}O_2 \longrightarrow A + H_2O \qquad [49]$$

The carriers are NAD^+ and, less frequently, $NADP^+$; the flavoproteins FAD and FMN (flavin mononucleotide); ubiquinone (or coenzyme Q); and several types of cytochromes. Each carrier has an oxidized and reduced form (*e.g.*, FAD and $FADH_2$, respectively), the two forms constituting an oxidation-reduction, or redox, couple. Within the respiratory chain each redox couple undergoes cyclic oxidation-reduction—*i.e.*, the oxidized component of the couple accepts reducing equivalents from either a substrate or a reduced carrier preceding it in the series and in turn donates these reducing equivalents to the next oxidized carrier in the sequence. Reducing equivalents are thus transferred from substrates to molecular oxygen by a number of sequential redox reactions.

Most oxidizable catabolic intermediates initially undergo a dehydrogenation reaction, during which a dehydrogenase enzyme transfers the equivalent of a hydride ion ($H^+ + 2e^-$, with e^- representing an electron) to its coenzyme, either NAD^+ or $NADP^+$ (as examples, see [40] and [46]). The reduced NAD^+ (or $NADP^+$) thus produced (usually written as $NADH + H^+$ or $NADPH + H^+$) diffuses to the membrane-bound respiratory chain to be oxidized by an enzyme known as NADH dehydrogenase; the enzyme has as its coenzyme FMN. There is no corresponding NADPH dehydrogenase in mammalian mitochondria; instead, the reducing equivalents of $NADPH + H^+$ are transferred to NAD^+ in a reaction catalyzed by a transhydrogenase enzyme, with the products being reduced $NADH + H^+$ and $NADP^+$. A few substrates (*e.g.*, acyl coenzyme A and succinate; see reactions [22] and [44]) bypass this reaction and instead undergo immediate dehydrogenation by specific membrane-bound dehydrogenase enzymes. During the reaction, the coenzyme FAD accepts two hydrogen atoms and two electrons ($2H + 2e^-$). The reduced flavoproteins (*i.e.*, $FMNH_2$ and $FADH_2$) donate their two hydrogen atoms to the lipid carrier ubiquinone, which is thus reduced.

The fourth type of carrier, the cytochromes, consists of hemoproteins—*i.e.*, proteins with a nonprotein component, or prosthetic group, called heme (or a derivative of heme), which is an iron-containing pigment molecule. The iron atom in the prosthetic group is able to carry one electron and oscillates between the oxidized, or ferric (Fe^{3+}), and the reduced, or ferrous (Fe^{2+}), forms. The

five cytochromes present in the mammalian respiratory chain, designated cytochromes *b*, c_1, *c*, *a*, and a_3, act in sequence between ubiquinone and molecular oxygen. The terminal cytochrome of this sequence (a_3, also known as cytochrome oxidase) is able to donate electrons to oxygen rather than to another electron carrier; a_3 is also the site of action of two substances that inhibit the respiratory chain, potassium cyanide and carbon monoxide. Special Fe-S complexes play a role in the activity of NADH dehydrogenase and succinate dehydrogenase. The sequence of carriers, from substrates to oxygen, is shown schematically in Figure 7.

Sequence of carriers

In each redox couple the reduced form has a tendency to lose reducing equivalents (*i.e.*, to act as an electron or hydrogen donor); similarly, the oxidized form has a tendency to gain reducing equivalents (*i.e.*, to act as an electron or hydrogen acceptor). The oxidation-reduction characteristics of each couple can be determined experimentally under well-defined, standard conditions. The value thus obtained is the standard oxidation-reduction (redox) potential (E_6). Values for respiratory chain carriers range from $E_6 = -320$ millivolts (one millivolt = 0.001 volt) for NAD^+/reduced NAD^+ to $E_6 = +820$ millivolts for $^1/_2O_2/H_2O$; the values for intermediate carriers lie between. Reduced NAD^+ is the most electronegative carrier, oxygen the most electropositive acceptor. During respiration reducing equivalents undergo stepwise transfer from the reduced form of the most electronegative carrier (reduced NAD^+) to the oxidized form of the most electropositive couple (oxygen). Each step is accompanied by a decline in standard free energy ($\Delta G'$) proportional to the difference in the standard redox potentials (ΔE_0) of the two carriers involved.

Overall oxidation of reduced NAD^+ by oxygen ($\Delta E_0 = +1{,}140$ millivolts) is accompanied by the liberation of free energy ($\Delta G' = -52.4$ kilocalories per mole); in theory this energy is sufficient to allow the synthesis of six or seven molecules of ATP. In the cell, however, this synthesis of ATP, called oxidative phosphorylation, proceeds with an efficiency of about 46 percent; thus only three molecules of ATP are produced per atom of oxygen consumed—this being the so-called $P/2e^-$, P/O, or ADP/O ratio. The energy that is not conserved as ATP is lost as heat. The oxidation of succinate by molecular oxygen ($\Delta E_0 = +790$ millivolts), which is accompanied by a smaller liberation of free energy ($\Delta G' = -36.5$ kilocalories per mole), yields only two molecules of ATP per atom of oxygen consumed (P/O = 2).

Oxidative phosphorylation

In order to understand the mechanism by which the energy released during respiration is conserved as ATP, it is necessary to appreciate the structural features of mitochondria. These are organelles in animal and plant cells in which oxidative phosphorylation takes place. There are many mitochondria in animal tissues; for example, in heart and skeletal muscle, which require large amounts of energy for mechanical work, in the pancreas, where there is biosynthesis, and in the kidney, where the process of excretion begins. Mitochondria have an outer membrane, which allows the passage of most small molecules and ions, and a highly folded inner membrane (crista), which does not even allow the passage of small ions and so maintains a closed space within the cell. The electron-transferring molecules of the respiratory chain and the enzymes responsible for ATP synthesis are located in and on this inner membrane, while the space inside (matrix) contains the enzymes of the TCA cycle (reactions [34] to [46]; see also CELLS). The enzyme systems primarily responsible for the release and subsequent oxidation of reducing equivalents are thus closely related so that the reduced coenzymes formed during catabolism ($NADH + H^+$ and $FADH_2$) are available as substrates for respiration. The movement of most charged metabolites into the matrix space is mediated by special carrier proteins in the crista that catalyze exchange-diffusion (*i.e.*, a one-for-one exchange). The oxidative phosphorylation systems of bacteria are similar in principle but show a greater diversity in the composition of their respiratory carriers.

The mechanism of ATP synthesis appears to be as follows. During the transfer of hydrogen atoms from $FMNH_2$ or $FADH_2$ to oxygen (Figure 7), protons (H^+ ions) are pumped across the crista from the inside of the mitochondrion to the outside. Thus, respiration generates an electrical potential (and in mitochondria a small pH gradient) across the membrane corresponding to 200 to 300 millivolts, and the chemical energy in the substrate is converted into electrical energy. Attached to the crista is a complex enzyme (ATP synthetase) that binds ATP, ADP, and P_i. It has nine polypeptide chain subunits of five different kinds in a cluster and a unit of at least three more membrane proteins composing the attachment point of ADP and P_i. This complex forms a specific proton pore in the membrane. When ADP and P_i are bound to ATP synthetase, the excess of protons (H^+) that has formed outside of the mitochondria (an H^+ gradient) moves back into the mitochondrion through the enzyme complex. The energy released is used to convert ADP and P_i to ATP. In this process, electrical energy is converted to chemical energy, and it is the supply of ADP that limits the rate of this process. The precise mechanism by which the ATP synthetase complex converts the energy stored in the electrical H^+ gradient to the chemical bond energy in ATP is not well understood. The H^+ gradient may power other endergonic (energy-requiring) processes besides ATP synthesis, such as the movement of bacterial cells and the transport of carbon substrates or ions.

Photosynthesis generates ATP by a mechanism that is similar in principle, if not in detail. The organelles responsible are different from mitochondria, but they also

Photosynthesis

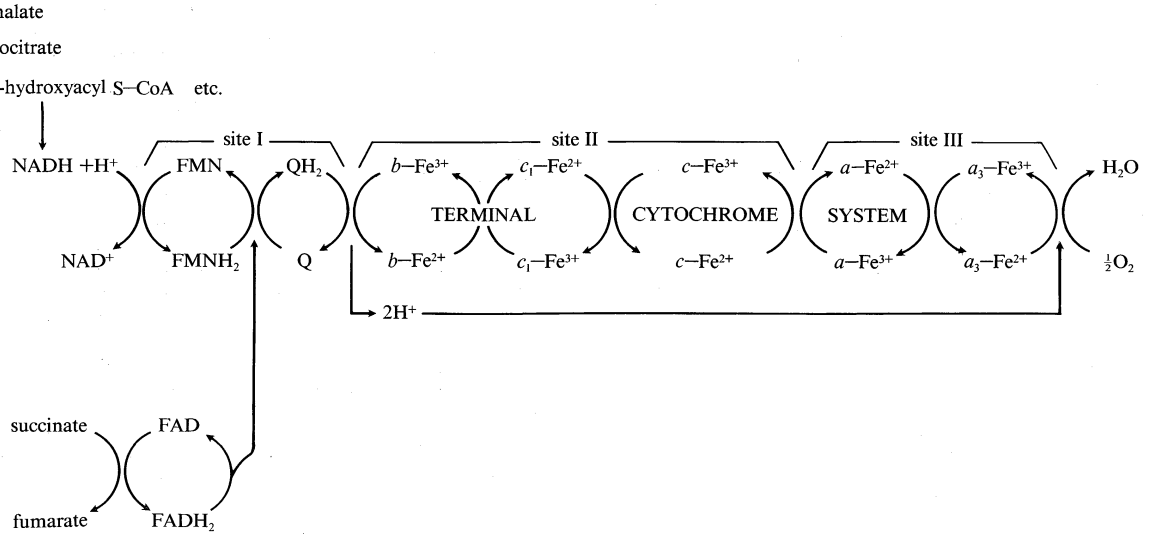


Figure 7: The respiratory chain (see text).

form membrane-bounded closed sacs (thylakoids) often arranged in stacks (grana). Solar energy splits two molecules of H_2O into molecular oxygen (O_2), four protons (H^+), and four electrons.

This is the source of oxygen evolution, clearly visible as bubbles from underwater plants in bright sunshine. The process involves a chlorophyll molecule, P_{680}, that changes its redox potential from +820 millivolts (in which there is a tendency to accept electrons) to about −680 millivolts (in which there is a tendency to lose electrons) upon excitation with light and acquisition of electrons. The electrons are subsequently passed along a series of carriers (plastoquinone, cytochromes b and f, and plastocyanin), analogous to the mitochondrial respiratory chain. This process pumps protons across the membrane from the outside of the thylakoid membrane to the inside. Protons (H^+) do not move freely across the membrane although chloride ions (Cl^-) do, creating a pH gradient. An ATP synthetase enzyme similar to that of the mitochondria is present, but on the outside of the thylakoid membrane. Passage of protons (H^+) through it from inside to outside generates ATP.

Hence, a gradient of protons (H^+) across the membrane is the high-energy intermediate for forming ATP in plant photosynthesis and in the respiration of all cells capable of passing reducing equivalents (hydrogen atoms or electrons) to electron acceptors.

The biosynthesis of cell components

THE NATURE OF BIOSYNTHESIS

The stages of biosynthesis. The biosynthesis of cell components (anabolism) may be regarded as occurring in two main stages. In the first, intermediate compounds of the central routes of metabolism are diverted from further catabolism and are channeled into pathways that usually lead to the formation of the relatively small molecules that serve as the building blocks, or precursors, of macromolecules.

In the second stage of biosynthesis, the building blocks are combined to yield the macromolecules—proteins, nucleic acids, lipids, and polysaccharides—that make up the bulk of tissues and cellular components. In organisms with the appropriate genetic capability, for example, all of the amino acids can be synthesized from ammonia and intermediates of the main routes of carbohydrate fragmentation and oxidation. Such intermediates act also as precursors for the purines, the pyrimidines, and the pentose sugars that constitute DNA and for a number of types of RNA. The assembly of proteins necessitates the precise combination of specific amino acids in a highly ordered and controlled manner; this in turn involves the copying, or transcription, into RNA of specific parts of DNA (see below *The synthesis of macromolecules*). The first stage of biosynthesis thus requires the specificity normally required for the efficient functioning of sequences of enzyme-catalyzed reactions. The second stage also involves—directly for protein and nucleic acid synthesis, less directly for the synthesis of other macromolecules—the maintenance and expression of the biological information that specifies the identity of the cell, the tissue, and the organism.

Requirements

Utilization of ATP. The two stages of biosynthesis—the formation of building blocks and their specific assembly into macromolecules—are energy-consuming processes and thus require ATP. Although the ATP is derived from catabolism, catabolism does not "drive" biosynthesis. As explained in the first section of this article, the occurrence of chemical reactions in the living cell is accompanied by a net decrease in free energy. Although biological growth and development result in the creation of ordered systems from less ordered ones and of complex systems from simpler ones, these events must occur at the expense of energy-yielding reactions. The overall coupled reactions are, on balance, still accompanied by a decrease in free energy and are thus essentially irreversible in the direction of biosynthesis. The total energy released from ATP, for example, is usually much greater than is needed for a particular biosynthetic step; thus, many of the reactions involved in biosynthesis release inorganic pyrophosphate (PP_i) rather than phosphate (P_i) from ATP, and hence yield AMP rather than ADP. Since inorganic pyrophosphate readily undergoes virtually irreversible hydrolysis to two equivalents of inorganic phosphate (see [21a]), the creation of a new bond in the product of synthesis may be accompanied by the breaking of two high-energy bonds of ATP—although, in theory, one might have sufficed.

The efficient utilization for anabolic processes of ATP and some intermediate compound formed during a catabolic reaction requires the cell to have simultaneously a milieu favourable for both ATP generation and consumption. Catabolism occurs readily only if sufficient ADP is available; hence, the concentration of ATP is low. On the other hand, biosynthesis requires a high level of ATP and consequently low levels of ADP and AMP. Suitable conditions for the simultaneous function of both processes are met in two ways. Biosynthetic reactions often take place in compartments within the cell different from those in which catabolism occurs; there is thus a physical separation of energy-requiring and energy-yielding processes. Furthermore, biosynthetic reactions are regulated independently of the mechanisms by which catabolism is controlled. Such independent control is made possible by the fact that catabolic and anabolic pathways are not identical; the pacemaker, or key, enzyme that controls the overall rate of a catabolic route usually does not play any role in the biosynthetic pathway of a compound. Similarly, the pacemaker enzymes of biosynthesis are not involved in catabolism. As discussed below (see *Regulation of metabolism: Fine control*), catabolic pathways are often regulated by the relative amounts of ATP, ADP, and AMP in the cellular compartment in which the pacemaker enzymes are located; in general, ATP inhibits and ADP (or AMP) stimulates such enzymes. In contrast, many biosynthetic routes are regulated by the concentration of the end products of particular anabolic processes, so that the cell synthesizes only as much of these building blocks as it needs.

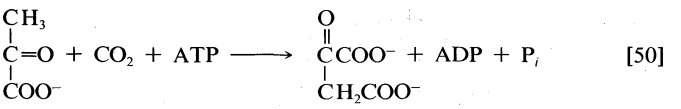
Regulation

THE SUPPLY OF BIOSYNTHETIC PRECURSORS

When higher animals consume a mixed diet, sufficient quantities of compounds for both biosynthesis and energy supply are available. Carbohydrates yield intermediates of glycolysis and of the phosphogluconate pathway, which in turn yield acetyl coenzyme A (see Figure 4); lipids yield glycolytic intermediates and acetyl coenzyme A (see Figure 2); and many amino acids form intermediates of both the TCA cycle and glycolysis. Any intermediate withdrawn for biosynthesis can thus be readily replenished by the catabolism of further nutrients. This situation does not always hold, however. Microorganisms in particular can derive all of their carbon and energy requirements by utilizing a single carbon source. The sole carbon source may be a substance such as a carbohydrate or a fatty acid, or an intermediate of the TCA cycle (or a substance readily converted to one). In both cases, reactions ancillary to those discussed thus far must occur before the carbon source can be utilized.

Anaplerotic routes. Although the catabolism of carbohydrates can occur via a variety of routes (see Figure 4), all give rise to pyruvate. During the catabolism of pyruvate, one carbon atom is lost as carbon dioxide and the remaining two form acetyl coenzyme A [37]; these two are involved in the TCA cycle ([41] and [42]). Because the TCA cycle is initiated by the condensation of acetyl coenzyme A with oxaloacetate, which is regenerated in each turn of the cycle, the removal of any intermediate from the cycle would cause the cycle to stop. Yet, as also indicated in Figure 4, various essential cell components are derived from α-oxoglutarate, succinyl coenzyme A, and oxaloacetate, so that these compounds are, in fact, removed from the cycle. Microbial growth with a carbohydrate as the sole carbon source is thus possible only if a cellular process occurs that effects the net formation of some TCA cycle intermediate from an intermediate of carbohydrate catabolism. Such a process, which replenishes the TCA cycle, has been described as an anaplerotic reaction.

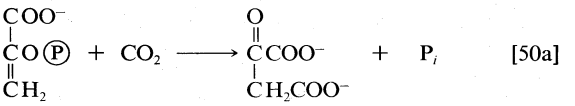
The anaplerotic function may be carried out by either of



pyruvate oxaloacetate [50]

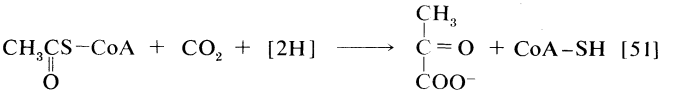
two enzymes that catalyze the fixation of carbon dioxide onto a three-carbon compound, either pyruvate [50] or phosphoenolpyruvate (PEP, [50a]) to form oxaloacetate, which has four carbon atoms. Both reactions require energy. In [50] it is supplied by the cleavage of ATP to ADP and inorganic phosphate (P_i); in [50a] it is supplied by the release of the high-energy phosphate of PEP as inorganic phosphate. Pyruvate serves as a carbon dioxide acceptor not only in many bacteria and fungi but also in the livers and kidneys of higher organisms, including man; PEP serves as the carbon dioxide acceptor in many bacteria, such as those that inhabit the gut.

Carbon dioxide acceptors



PEP oxaloacetate [50a]

Unlike higher organisms, many bacteria and fungi can grow on acetate or compounds such as ethanol or a fatty acid that can be catabolized to acetyl coenzyme A. Under these conditions, the net formation of TCA cycle intermediates proceeds in one of two ways. In obligate anaerobic bacteria, pyruvate can be formed from acetyl coenzyme A and carbon dioxide [51]; reducing equivalents [2H] are necessary for the reaction. The pyruvate so formed can then react via either step [50] or [50a].



acetyl coenzyme A pyruvate [51]

Reaction [51] does not occur in facultative anaerobic organisms or in strict aerobes, however. Instead, in these organisms two molecules of acetyl coenzyme A give rise to the net synthesis of a four-carbon intermediate of the TCA cycle via a route known as the glyoxylate cycle. In this route (Figure 8), the steps of the TCA cycle that lead to the loss of carbon dioxide (see [40], [41], and [42]) are bypassed. Instead of being oxidized to oxalosuccinate, as occurs in [40], isocitrate is split by isocitrate lyase [52] in a reaction similar to that of reactions [4] and [15] of

The glyoxylate cycle

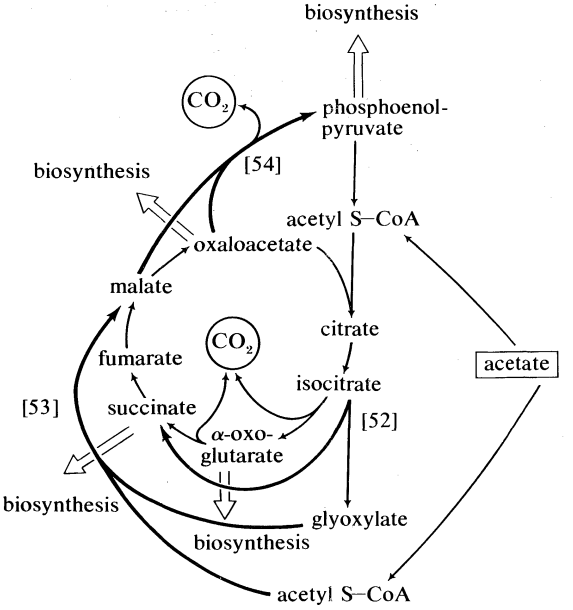
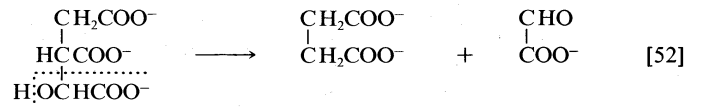
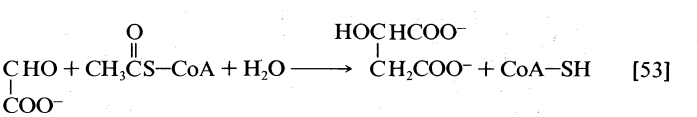


Figure 8: The glyoxylate cycle. Numbers refer to reactions described in text.



isocitrate succinate glyoxylate [52]

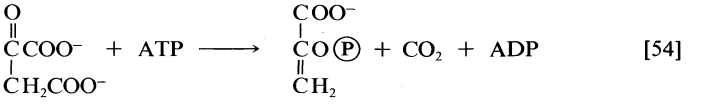
carbohydrate fragmentation. The dotted line in [52] indicates the way in which isocitrate is split. The products are succinate and glyoxylate. Glyoxylate, like oxaloacetate, is the anion of an *α*-oxoacid and thus can condense, in a reaction catalyzed by malate synthase, with acetyl coenzyme A; the products are coenzyme A and malate [53].



glyoxylate acetyl coenzyme A malate [53]

In conjunction with the reactions of the TCA cycle that effect the re-formation of isocitrate from malate (steps [46], [38], and [39]), steps [52] and [53] lead to the net production of a four-carbon compound (malate) from two two-carbon units (glyoxylate and acetyl coenzyme A). The sequence thus complements the TCA cycle, enabling the cycle to fulfill the dual roles of providing both energy and biosynthetic building blocks when the sole carbon source is a two-carbon compound such as acetate.

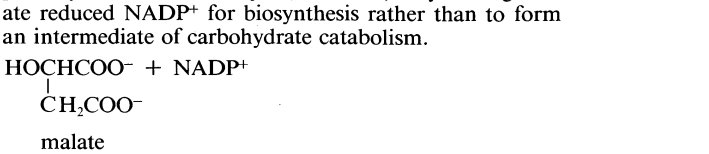
Growth of microorganisms on TCA cycle intermediates. Most aerobic microorganisms grow readily on substances such as succinate or malate as their sole source of carbon. Under these circumstances, the formation of the intermediates of carbohydrate metabolism requires an enzymatic step ancillary to the central pathways. In most cases this step is catalyzed by phosphoenolpyruvate (PEP) carboxykinase [54]. Oxaloacetate is decarboxylated (*i.e.*, carbon dioxide is removed) during this energy-requiring reaction. The energy may be supplied by ATP or a similar substance (*e.g.*, GTP) that can readily be derived from it via a reaction of the type shown in [43a]. The products are PEP, carbon dioxide, and ADP.



oxaloacetate PEP [54]

Another reaction that can yield an intermediate of carbohydrate catabolism is catalyzed by the so-called malic enzyme; in this reaction, malate is decarboxylated to pyruvate, with concomitant reduction of $NADP^+$ [55]. The primary role of malic enzyme, however, may be to generate reduced $NADP^+$ for biosynthesis rather than to form an intermediate of carbohydrate catabolism.

The malic enzyme



malate pyruvate [55]

THE SYNTHESIS OF BUILDING BLOCKS

Gluconeogenesis. The formation of sugars from noncarbohydrate precursors, gluconeogenesis, is of major importance in all living organisms. In the light, photosynthetic plants and microorganisms incorporate, or fix, carbon dioxide onto a five-carbon sugar and, via a sequence of transfer reactions, re-form the same sugar while also effecting the net synthesis of the glycolytic intermediate, 3-phosphoglycerate (see PHOTOSYNTHESIS). Phosphoglyc-

erate is the precursor of starch, cell-wall carbohydrates, and other plant polysaccharides. A situation similar in principle applies to the growth of microorganisms on precursors of acetyl coenzyme A (Figure 8) or on intermediates of the TCA cycle—that is, a large variety of cell components are derived from carbohydrates that, in turn, are synthesized from these noncarbohydrate precursors. Higher organisms also readily convert glucogenic amino acids (*i.e.*, those that do not yield acetyl coenzyme A as a catabolic product) into TCA cycle intermediates, which are then converted into glucose. The amounts of glucose thus transformed depend on the needs of the organism for protein synthesis and on the availability of fuels other than glucose. The synthesis of blood glucose from lactate, which occurs largely in liver, is a particularly active process during recovery from intense muscular activity.

Most of the steps in the pathway for the biosynthesis of glucose from pyruvate are catalyzed by the enzymes of glycolysis; the direction of the reactions is reversed. Three virtually irreversible steps in glucose catabolism (see [10], [3], and [1]) that cannot be utilized in gluconeogenesis, however, are bypassed by alternative reactions that tend to proceed in the direction of glucose synthesis (Figure 9).

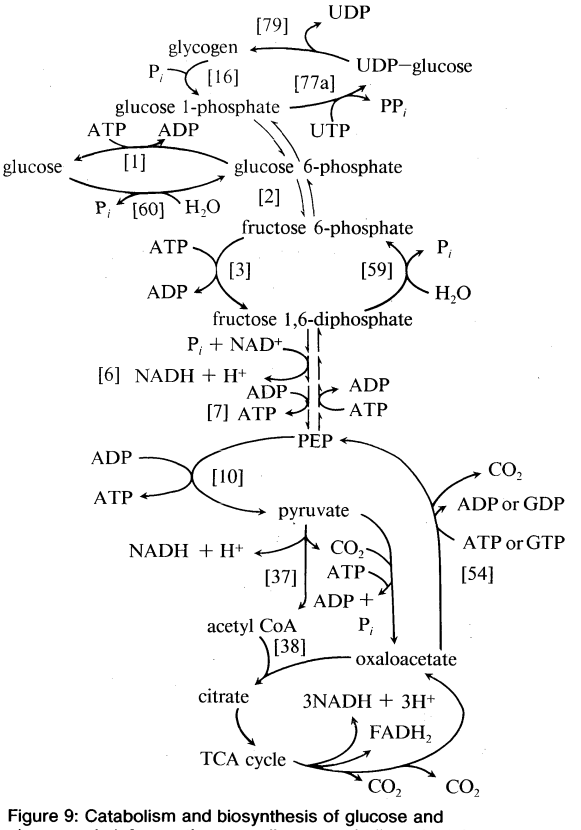
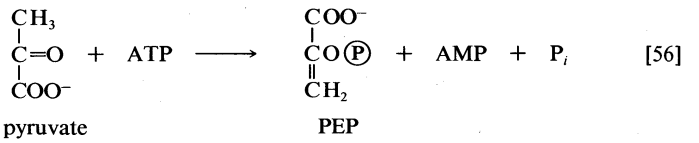


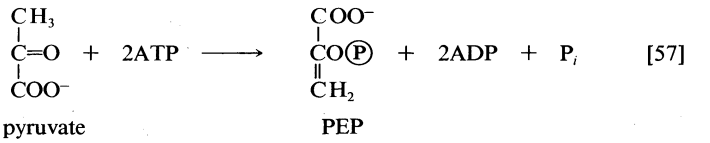
Figure 9: Catabolism and biosynthesis of glucose and glycogen. At left, reactions peculiar to catabolism. At right, reactions peculiar to anabolism. Numbers in brackets refer to reactions explained in text.

The first alternative reaction is the conversion of pyruvate to PEP. Three mechanisms for overcoming the energy barrier associated with the direct reversal of the pyruvate kinase reaction [10] are known. In some bacteria, PEP is formed from pyruvate by the utilization of two of the high-energy bonds of ATP; the products include, in addition to PEP, AMP and inorganic phosphate [56]. A variant of this reaction occurs in some bacteria, in which ATP and inorganic phosphate are reactants and AMP and inorganic pyrophosphate are products; as mentioned above, inorganic pyrophosphate is likely to be hydrolyzed to two equivalents of inorganic phosphate, so that the net balance of the reaction is identical with [56].

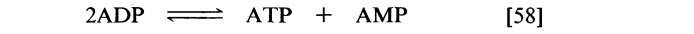


In other organisms, including many microorganisms, birds, and mammals, the formation of PEP from pyruvate is effected by the sum of reactions [50] and [54], each of which consumes one ATP; the overall balance is shown in [57], in which two molecules of ATP react with pyruvate

Formation of PEP from pyruvate

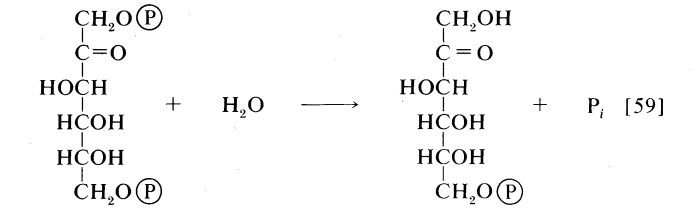


to form PEP, ADP, and inorganic phosphate. The enzyme adenylate kinase catalyzes the interconversion of the various adenine nucleotides, as shown in [58]. The combination of steps [57] and [58] yields the same energy balance as does the direct conversion of pyruvate to PEP [56].

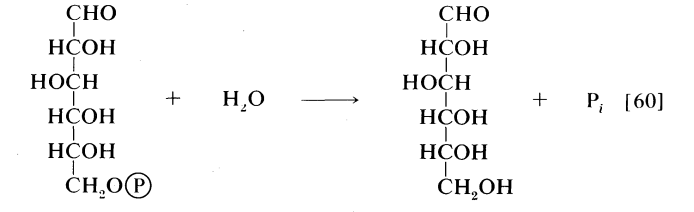


The second step of glycolysis bypassed in gluconeogenesis is that catalyzed by phosphofructokinase [3]. Instead, the fructose 1,6-diphosphate synthesized from dihydroxyacetonephosphate and glyceraldehyde 3-phosphate in the reaction catalyzed by aldolase is hydrolyzed, with the loss of the phosphate group linked to the first carbon atom. The enzyme fructose diphosphatase catalyzes the reaction [59], in which the products are fructose 6-phosphate

Phosphofructokinase



and inorganic phosphate. The fructose 6-phosphate thus formed is a precursor of mucopolysaccharides (polysaccharides with nitrogen-containing components). In addition, its conversion to glucose 6-phosphate provides the starting material for the formation of storage polysaccharides such as starch and glycogen, of monosaccharides other than glucose, of disaccharides (carbohydrates with two sugar components), and of some structural polysaccharides (*e.g.*, cellulose). The maintenance of the glucose content of vertebrate blood requires glucose 6-phosphate to be converted to glucose. This process occurs in the kidney, in the lining of the intestine, and most importantly in the liver. The reaction does not occur by reversal of the hexokinase or glucokinase reactions that effect the formation of glucose 6-phosphate from glucose and ATP [1]; rather, glucose 6-phosphate is hydrolyzed in a reaction catalyzed by glucose 6-phosphatase, and the phosphate is released as inorganic phosphate [60].

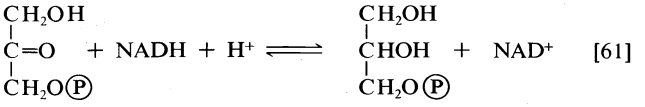


Lipid components. The component building blocks of the lipids found in storage fats, in lipoproteins (combinations of lipid and protein), and in the membranes of cells and organelles are glycerol, the fatty acids, and a number of other compounds (*e.g.*, serine, inositol).

Building blocks of lipids

Glycerol. Glycerol is readily derived from dihydroxyacetone phosphate, an intermediate of glycolysis (see [4]). In a reaction catalyzed by glycerol 1-phosphate dehydro-

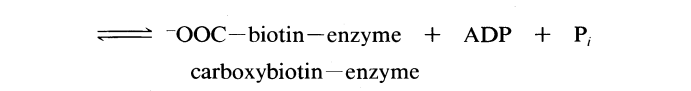
genase [61], dihydroxyacetone phosphate is reduced to glycerol 1-phosphate; reduced NAD^+ provides the reducing equivalents for the reaction and is oxidized. This compound reacts further (see below *Other components*).



[61]

dihydroxyacetone phosphate — glycerol 1-phosphate

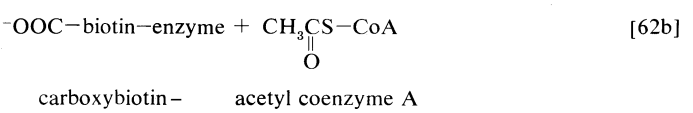
Fatty acids. Although all the carbon atoms of the fatty acids found in lipids are derived from the acetyl coenzyme A produced by the catabolism of carbohydrates and fatty acids (Figure 2), the molecule first undergoes a carboxylation, forming malonyl coenzyme A, before participating in fatty acid synthesis. The carboxylation reaction is catalyzed by acetyl CoA carboxylase, an enzyme whose prosthetic group is the vitamin biotin. The biotin–enzyme first undergoes a reaction that results in the attachment of carbon dioxide to biotin; ATP is required and forms ADP and inorganic phosphate [62a]. The complex product,



[62a]

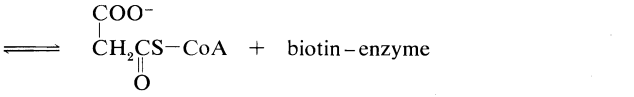
carboxybiotin–enzyme

called carboxybiotin–enzyme, releases the carboxy moiety to acetyl coenzyme A, forming malonyl coenzyme A and restoring the biotin–enzyme [62b].



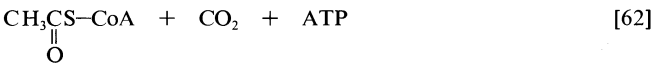
[62b]

carboxybiotin–enzyme — acetyl coenzyme A



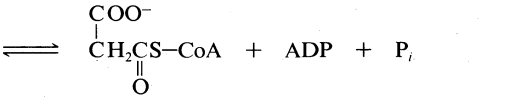
malonyl coenzyme A

The overall reaction [62] catalyzed by acetyl coenzyme A carboxylase thus involves the expenditure of one molecule of ATP for the formation of each molecule of malonyl coenzyme A from acetyl coenzyme A and carbon dioxide.



[62]

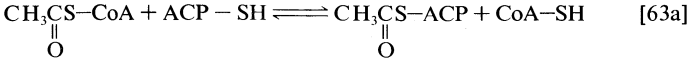
acetyl coenzyme A



malonyl coenzyme A

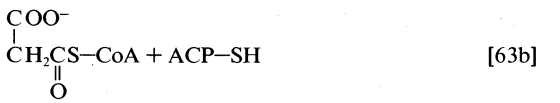
Malonyl coenzyme A and a molecule of acetyl coenzyme A react (in bacteria) with the sulfhydryl group of a relatively small molecule known as acyl-carrier protein (ACP–SH); in higher organisms ACP–SH is part of a multienzyme complex called fatty acid synthetase. ACP–SH is involved in all of the reactions leading to the synthesis of a fatty acid such as palmitic acid from acetyl coenzyme A and malonyl coenzyme A. The products of [63a] and [63b] are acetyl-S-ACP, malonyl-S-ACP, and coenzyme A. The enzymes catalyzing [63a] and [63b] are known as acetyl transacylase and malonyl transacylase, respectively. Acetyl-ACP and malonyl-ACP react in a reaction

Acyl-carrier protein



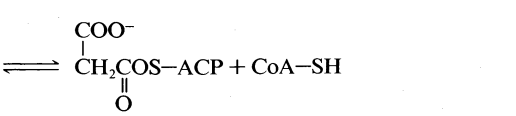
[63a]

acetyl coenzyme A — acetyl-S–ACP



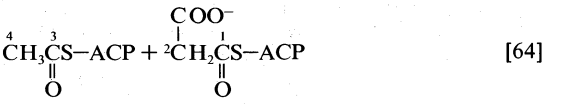
[63b]

malonyl coenzyme A



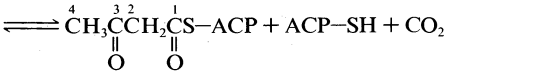
malonyl-S–ACP

catalyzed by β-ketoacyl-ACP synthetase so that the acetyl moiety (CH_3CO-) is transferred to the malonyl moiety ($^-OOCH_2CO-$). Simultaneously, the carbon dioxide fixed in step [62] is lost, leaving as a product a four-carbon moiety attached to ACP and called acetoacetyl-S-ACP [64].



[64]

acetyl-S–ACP malonyl-S–ACP

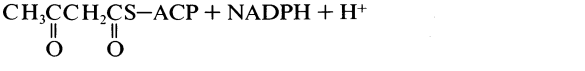


acetoacetyl-S–ACP

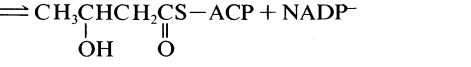
It should be noted that the carbon atoms of acetyl-S-ACP occur at the end of acetoacetyl-S-ACP (see carbon atoms numbered 4 and 3 in [64]) and that carbon dioxide plays an essentially catalytic role; the decarboxylation of the malonyl-S-ACP [64] provides a strong thermodynamic pull toward fatty acid synthesis.

Fatty acid synthesis and breakdown

The analogy between reaction [64] of fatty acid synthesis and the cleavage reaction [25] of fatty acid catabolism is apparent in the other reactions of fatty acid synthesis. The acetoacetyl-S-ACP, for example, undergoes reduction to β-hydroxybutyryl-S-ACP [65]; the reaction is catalyzed

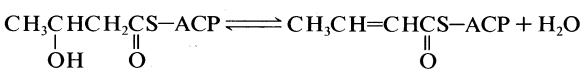


acetoacetyl-S–ACP [65]



β-hydroxybutyryl-S–ACP

by β-ketoacyl-ACP reductase. Reduced $NADP^+$ is the electron donor, however, and not reduced NAD^+ (which would participate in the reversal of reaction [24]). $NADP^+$ is thus a product in [65]. In [66] β-hydroxybutyryl-S-ACP is dehydrated (*i.e.*, one molecule of water is removed), in a reaction catalyzed by enoyl-ACP-hydrase, and then undergoes a second reduction [67], in which reduced $NADP^+$ again acts as the electron donor. The products of [66] are crotonyl-S-ACP and water. The products of [67], which is catalyzed by crotonyl-ACP reductase, are butyryl-S-ACP and $NADP^+$.



[66]

hydroxybutyryl-S–ACP — crotonyl-S–ACP

The formation of butyryl-S-ACP [67] completes the first of several cycles, in each of which one molecule of malonyl coenzyme A enters via reactions [62] and [63b]. In the cycle following the one ending with [67], the butyryl moiety is transferred to malonyl-S-ACP, and a molecule of carbon dioxide is again lost; a six-carbon compound results. In subsequent cycles, each of which adds two carbon atoms to the molecule via reaction [64], successively longer β-oxoacyl-S-ACP derivatives are produced.

$$CH_3CH{=}CHCS{-}ACP + NADPH + H^+ \quad [67]$$
(O)

crotonyl-S—ACP

$$\rightleftharpoons CH_3CH_2CH_2CS{-}ACP + NADP^+$$
(O)

butyryl-S—ACP

Ultimately, a molecule with 16 carbon atoms, palmityl-S-ACP, is formed. In most organisms a deacylase catalyzes the release of free palmitic acid; in a few, synthesis continues, and an acid with 18 carbon atoms is formed. The fatty acids can then react with coenzyme A (compare reaction [21]) to form fatty acyl coenzyme A, which can condense with the glycerol 1-phosphate formed in step [61]; the product is a phosphatidic acid. The overall formation of each molecule of palmitic acid from acetyl coenzyme A—via step [62] and repeated cycles of steps [63] through [67]—requires the investment of seven molecules of ATP and 14 of reduced $NADP^+$ (see [68]). The process is thus an energy-requiring one (endergonic) and represents a major way by which the reducing power generated in NADP-linked dehydrogenation reactions of carbohydrate catabolism is utilized (see above *The phosphogluconate pathway*).

$$CH_3CS{-}CoA + 7\left[\begin{matrix}COO^-\\ |\\ CH_2CS{-}CoA\\ \|\\ O\end{matrix}\right] \quad [68]$$
(O)

acetyl coenzyme A — malonyl coenzyme A

$$+ 14NADPH + 14H^+ \longrightarrow CH_3(CH_2)_{14}COO^-$$

palmitate

$$+ 7CO_2 + 8CoA{-}SH + 14NADP^+ + 6H_2O$$

Phospholipids

Other components. The major lipids that serve as components of membranes, called phospholipids, as well as lipoproteins, contain, in addition to two molecules of fatty acid, one molecule of a variety of different compounds. The precursors of these compounds include serine, inositol, and glycerol 1-phosphate. They are derived from

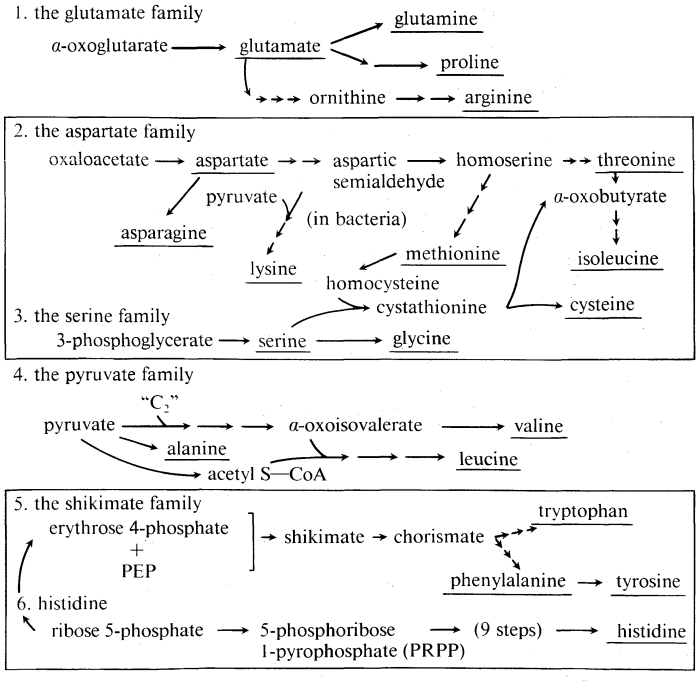


Figure 10: Family relationships in amino-acid biosyntheses. Components of proteins are underlined. Not all of the intermediates formed are named.

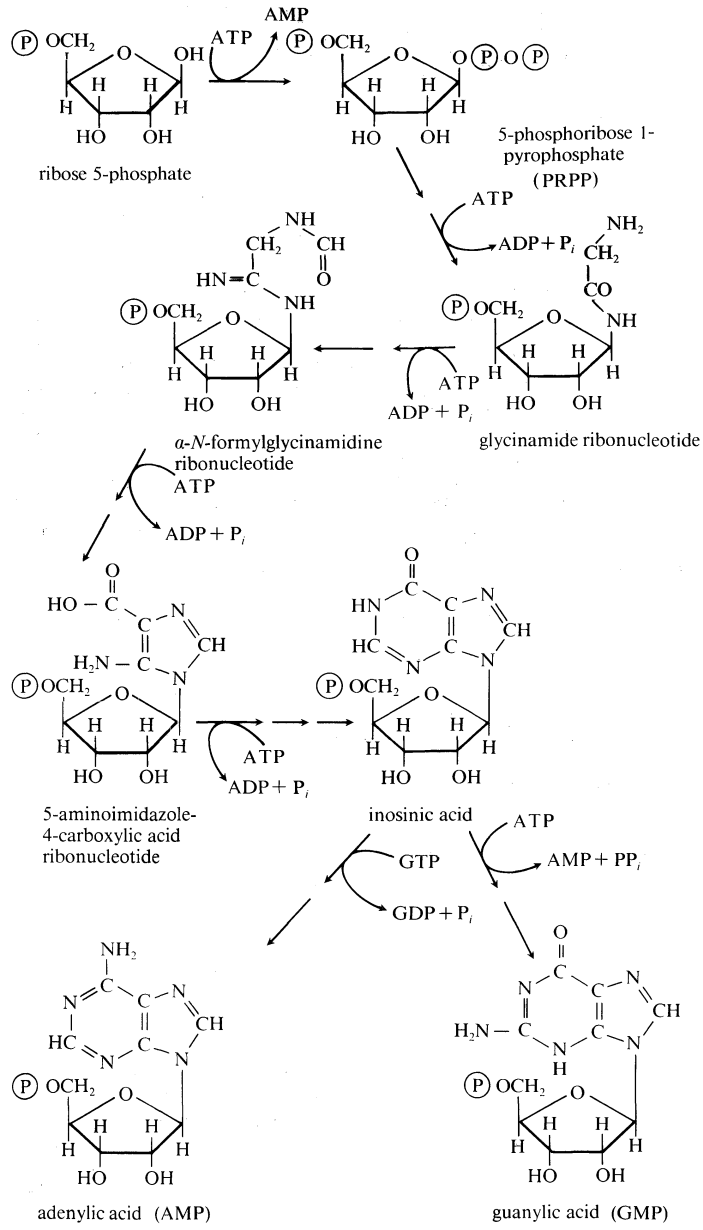


Figure 11: Biosynthesis of purine nucleotides. Not all of the intermediate compounds formed are shown.

intermediates of the central metabolic pathways (*e.g.*, Figure 10; reaction [61]).

Amino acids. Organisms differ considerably in their ability to synthesize amino acids from the intermediates of central metabolic pathways. Most vertebrates can form only the chemically most simple amino acids; the others must be supplied in the diet. Man, for example, synthesizes about 10 of the 20 commonly encountered amino acids; these are termed nonessential amino acids. The essential amino acids must be supplied in food.

Higher plants are more versatile than animals; they can make all of the amino acids required for protein synthesis, with either ammonia (NH_3) or nitrate (NO_3^-) as the nitrogen source. Some bacteria, and leguminous plants (*e.g.*, peas) that harbour such bacteria in their root nodules, are able to utilize nitrogen from the air to form ammonia and use the latter for amino-acid synthesis.

Bacteria differ widely in their ability to synthesize amino acids. Some species, such as *Escherichia coli*, which can grow in media supplied with only a single carbon source and ammonium salts, can make all of their amino acids from these starting materials. Other bacteria may require as many as 16 different amino acids.

Each of the 20 common amino acids is synthesized by

Amino-acid synthesis

a different pathway, the complexity of which reflects the chemical complexity of the amino acid formed. As with other compounds, the pathway for the synthesis of an amino acid is for the most part different from that by which it is catabolized. A detailed discussion of the pathway by which each amino acid is formed is beyond the scope of this article, but two salient features of amino-acid biosynthesis should be mentioned.

First, ammonia is incorporated into the intermediates of metabolic pathways mainly via the glutamate dehydrogenase reaction [28], which proceeds from right to left in biosynthetic reactions. Similarly, the transaminase enzymes (reactions [26a, b, and c]) enable the amino group (NH_2-) to be transferred to other amino acids.

Second, a group of several amino acids may be synthesized from one amino acid, which acts as a "parent" of an amino-acid "family." The families are also interrelated in several instances. Figure 10 shows, for bacteria that can synthesize 20 amino acids, the way in which they are derived from intermediates of pathways already considered. Alpha-oxoglutarate and oxaloacetate are intermediates of the TCA cycle; pyruvate, 3-phosphoglycerate, and PEP are intermediates of glycolysis; and ribose 5-phosphate and erythrose 4-phosphate are formed in the phosphogluconate pathway.

Mononucleotides. Most organisms can synthesize the purine and pyrimidine nucleotides that serve as the building blocks of RNA (containing nucleotides in which the pentose sugar is ribose, called ribonucleotides) and DNA (containing nucleotides in which the pentose sugar is deoxyribose, called deoxyribonucleotides) as well as the agents of energy exchange.

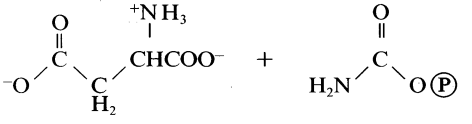
Purine ribonucleotides. The purine ribonucleotides (AMP and GMP) are derived from ribose 5-phosphate. The overall sequence that leads to the parent purine ribonucleotide, which is inosinic acid, involves 10 enzymatic steps.

Inosinic acid

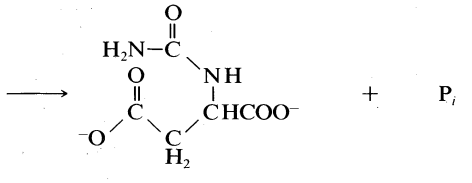
Figure 11 is an outline of the manner in which inosinic acid is synthesized. Inosinic acid can be converted to AMP and GMP; these in turn yield the triphosphates (*i.e.,* ATP and GTP) via reactions catalyzed by adenylate kinase [69] and nucleoside diphosphate kinase (see reaction [43a]).

$$GMP + ATP \rightleftharpoons GDP + ADP \qquad [69]$$

Pyrimidine ribonucleotides. The biosynthetic pathway for the pyrimidine nucleotides is somewhat simpler than that for the purine nucleotides. Aspartate (derived from the TCA cycle intermediate, oxaloacetate) and carbamoyl

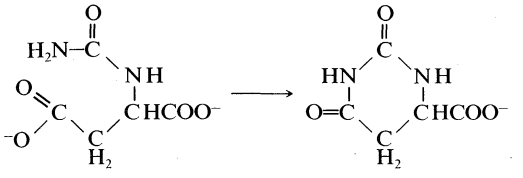


aspartate carbamoyl phosphate [70]



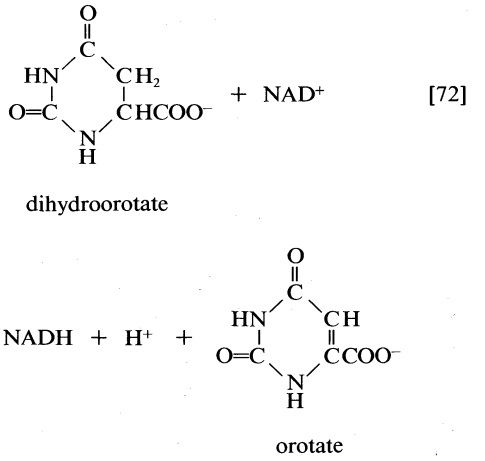
N-carbamoylaspartate

phosphate (derived from carbon dioxide, ATP, and ammonia via reaction [30]) condense to form *N*-carbamoylaspartate [70], which loses water [71] in a reaction catalyzed



N-carbamoylaspartate dihydroorotate [71]

by dihydroorotase; the product, dihydroorotate, is then oxidized to orotate in a reaction catalyzed by dihydroorotic acid dehydrogenase, in which NAD^+ is reduced [72]. Orotate



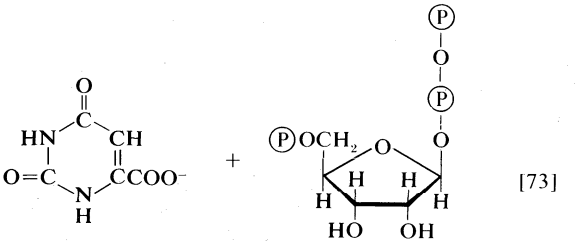
dihydroorotate [72]

orotate

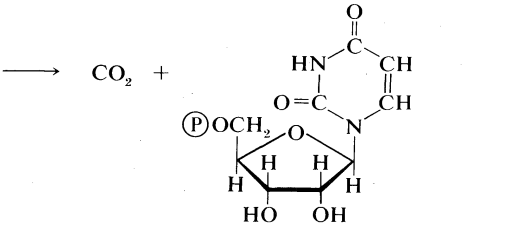
The orotate accepts a pentose phosphate moiety [73] from 5-phosphoribose 1-pyrophosphate (PRPP); PRPP, which is formed from ribose 5-phosphate and ATP, also initiates the pathways for biosynthesis of purine nucleotides (Figure 11) and of histidine (Figure 10). The product loses carbon dioxide to yield the parent pyrimidine nucleotide, uridylic acid (UMP; see [73]).

Analogous to the phosphorylation of purine nucleotides (steps [69] and [43a]) is the phosphorylation of UMP to UDP and thence to UTP by interaction with two molecules of ATP. Uridine triphosphate (UTP) can be converted to the other pyrimidine building block of RNA, cytidine triphosphate (CTP). In bacteria, the nitrogen for this reaction [74] is derived from ammonia; in higher animals, glutamine is the nitrogen donor.

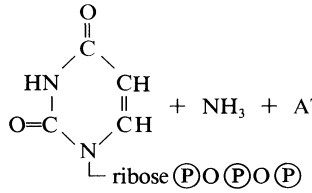
Deoxyribonucleotides. The building blocks for the synthesis of DNA differ from those for the synthesis of RNA in two respects. In DNA the purine and pyrimidine nucleotides contain the pentose sugar 2-deoxyribose instead of ribose. In addition, the pyrimidine base uracil, found in RNA, is replaced in DNA by thymine. The deoxyribonucleoside diphosphate can be derived directly from the corresponding ribonucleoside diphosphate by a process involving the two sulfhydryl groups of the protein, thioredoxin, and a flavoprotein, thioredoxin reductase, that can in turn be reduced by reduced $NADP^+$. Thus, for the reduction of XDP, in which X represents a purine base or cytosine, the reaction may be written as shown in [75a] and [75b]. In [75a] oxidized thioredoxin-S_2 is reduced to thioredoxin-$(SH)_2$ by NADPH, which is oxi-



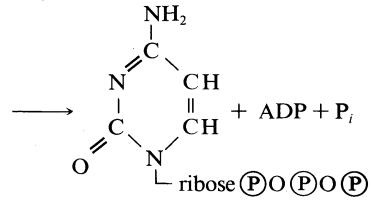
orotate PRPP [73]



UMP



UTP [74]



CTP

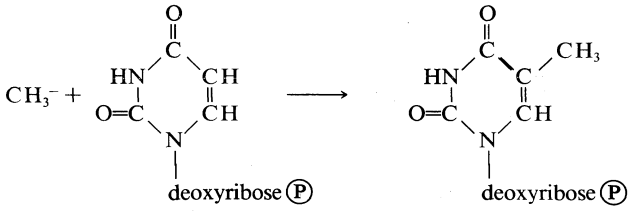
dized in the process. Thioredoxin-$(SH)_2$ then reduces XDP to deoxyXDP in a reaction [75b] in which thioredoxin is re-formed.

thioredoxin-S_2 + NADPH + H^+ [75a]
⟶ thioredoxin-$(SH)_2$ + $NADP^+$

thioredoxin-$(SH)_2$ + XDP [75b]
⟶ thioredoxin-S_2 + deoxyXDP

Deoxythymidylic acid

Deoxythymidylic acid (dTMP) is derived from deoxyuridylic acid (dUMP). Deoxyuridine diphosphate (dUDP) is first converted to dUMP, by reaction [69] proceeding from right to left. Deoxyuridylic acid then accepts a methyl group (CH_3—) in a reaction catalyzed by an enzyme (thymidylate synthetase) with the vitamin folic acid as a coenzyme; the product is dTMP [76].



dUMP dTMP [76]

THE SYNTHESIS OF MACROMOLECULES

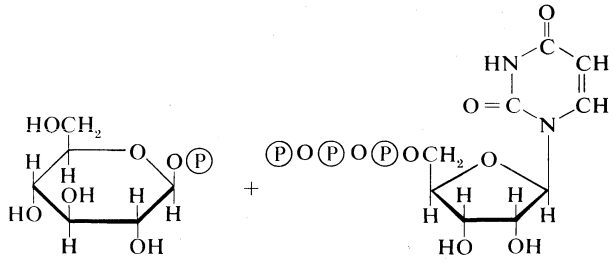
Carbohydrates and lipids. The formation of polysaccharides and of phospholipids from their component building blocks not only requires the investment of the energy of nucleoside triphosphates but uses these molecules in a novel manner. The biosynthetic reactions described thus far have mainly been accompanied by the formation of energy-rich intermediates (*e.g.*, PEP in [56]) with the formation of either AMP or ADP; however, nucleotides serve as intermediate carriers in the formation of glycogen, starch, and a variety of lipids. This unique process necessitates reactions by which ATP, or another nucleoside triphosphate, which can be readily derived from ATP via reactions of type [43a], combines with a phosphorylated reactant to form a nucleoside-diphosphate product. Although the change in standard free energy is small in this reaction, the subsequent hydrolysis of the inorganic pyrophosphate also released (reaction [21a]) effectively makes the reaction irreversible in the direction of synthesis. The nucleoside triphosphate is represented as NTP in [77], and the phosphorylated reactant as R—Ⓟ.

NTP + R—Ⓟ ⟹ NDP—R + PP_i [77]

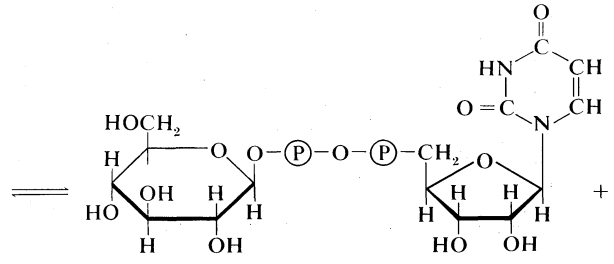
Reactions of type [77] are catalyzed by pyrophosphorylases, reaction [21a] by inorganic pyrophosphatase.

Glycogen and starch

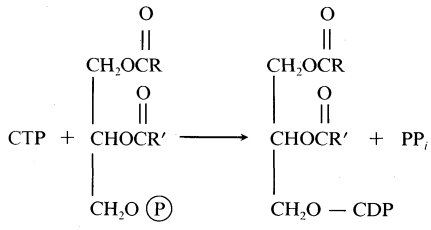
Formation of storage polysaccharides. In the formation of storage polysaccharides—*i.e.*, glycogen in animals,



glucose 1-phosphate UTP [77a]

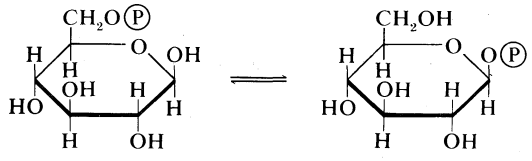


UDP-glucose



phosphatidic acid CDP-diglyceride [77b]

starch in plants—reaction [77] is preceded by the conversion of glucose 6-phosphate to glucose 1-phosphate, in a reaction catalyzed by phosphoglucomutase [78]. Glucose



glucose 6-phosphate glucose 1-phosphate [78]

1-phosphate functions as R—Ⓟ in reaction [77a]. UTP is the specific NTP for glycogen synthesis in animals [77a]; the products are UDP-glucose and pyrophosphate. In bacteria, fungi, and plants, ATP, CTP, or GTP serves instead of UTP. In all cases the nucleoside diphosphate glucose (NDP-glucose) thus synthesized can donate glucose to the terminal glucose of a polysaccharide chain, thereby increasing the number (n) of glucose molecules by one to $n + 1$[79]. UDP is released in this process, which is catalyzed by glycogen synthetase. Starch synthesis in plants occurs by an analogous pathway catalyzed by amylose synthetase; ADP-glucose rather than UDP-glucose is the preferred glucose donor [79a]. Similarly, cellulose, the major structural polysaccharide in plant cell walls, is synthesized in some plants by reaction [79a]; other plants undergo analogous reactions in which GDP-glucose or CDP-glucose acts as the glucose donor.

Cellulose synthesis

UDP-glucose + $(glucose)_n$ ⟶ UDP + $(glucose)_{n+1}$ [79]
ADP-glucose + $(glucose)_n$ ⟶ ADP + $(glucose)_{n+1}$ [79a]

Nucleoside diphosphate sugars also participate in the synthesis of disaccharides; for example, common table sugar, sucrose (consisting of glucose and fructose), is formed in sugarcane by the reaction sequence shown in [80] and [81];

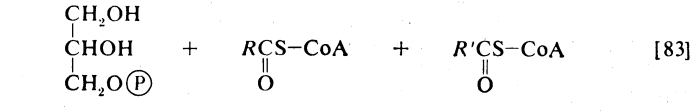
UDP-glucose + fructose 6-phosphate ⟶ UDP + sucrose 6′-phosphate [80]

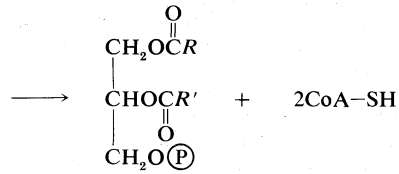
sucrose 6′-phosphate + H_2O ⟶ sucrose + P_i [81]

UDP-glucose and fructose 6-phosphate first form a phosphorylated derivative of sucrose, sucrose 6′-phosphate, which is hydrolyzed to sucrose and inorganic phosphate. Lactose, which consists of galactose and glucose, is the principal sugar of milk. It is synthesized in the mammary gland as shown in [82]; UDP-galactose and glucose react to form lactose; UDP is also a product.

UDP-galactose + glucose ⟶ UDP + lactose [82]

Formation of lipids. The neutral fats, or triglycerides, that constitute storage lipids, and the phospholipid components of lipoproteins and membranes, are synthesized from their building blocks by a route that branches after the first biosynthetic reaction. Initially, one molecule of glycerol 1-phosphate, the intermediate derived from carbohydrate catabolism, and two molecules of the appropriate fatty acyl coenzyme A (formed as described above, under *Biosynthesis of lipid components*) combine, yielding phosphatidic acid [83]. This reaction occurs preferentially with acyl coenzyme A derivatives of fatty acids containing 16 or 18 carbon atoms. In reaction [83], R and R' represent the hydrocarbon moieties ($Ch_3(CH_2)_n-$) of two fatty acid molecules. A triglyceride molecule (neutral fat) is formed from phosphatidic acid in a reaction catalyzed by a phosphatase that results in loss of the phosphate group [84]; the diglyceride thus formed can then accept a third molecule of fatty acyl coenzyme A (represented as $R''CS-CoA$ in [84a]).

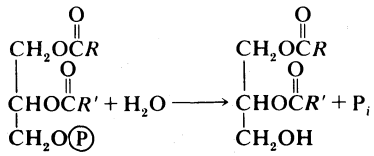




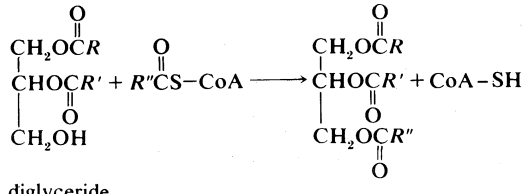
phosphatidic acid

Biosynthesis of phospholipids

In the biosynthesis of phospholipids, however, phosphatidic acid is not hydrolyzed; rather, it acts as the R—Ⓟ in reaction [77], the NTP here being cytidine triphosphate (CTP). A CDP-diglyceride is produced, and inorganic pyrophosphate is released [77b]. CDP-diglyceride is the common precursor of a variety of phospholipids. In subsequent reactions, each catalyzed by a specific enzyme, CMP is displaced from CDP-diglyceride by one of three compounds—serine, inositol, or glycerol 1-phosphate—to form CMP and, respectively, phosphatidylserine [85a], phosphatidylinositol [85b], or, in [85c], 3-phosphatidyl-glycerol 1′-phosphate (PGP). These reactions differ from those of polysaccharide biosynthesis ([79], [82]) in that phosphate is retained in the phospholipid, and the nucleotide product (CMP) is therefore a nucleoside monophosphate rather than the diphosphate. These compounds can react further: phosphatidylserine to give, sequentially, phosphatidylethanolamine and phosphatidylcholine; phosphatidylinositol to yield mono- and diphosphate derivatives that are components of brain tissue and of mitochondrial membranes; and PGP to yield the phosphatidylglycerol abundant in many bacterial membranes and the diphosphatidylglycerol that is also a major component of mitochondrial and bacterial membranes.



phosphatidic acid diglyceride [84]



diglyceride triglyceride [84a]

CDP-diglyceride + serine ⟶ phosphatidylserine + CMP [85a]

CDP-diglyceride + inositol ⟶ phosphatidylinositol + CMP [85b]

CDP-diglyceride + glycerol 1-phosphate ⟶ PGP + CMP [85c]

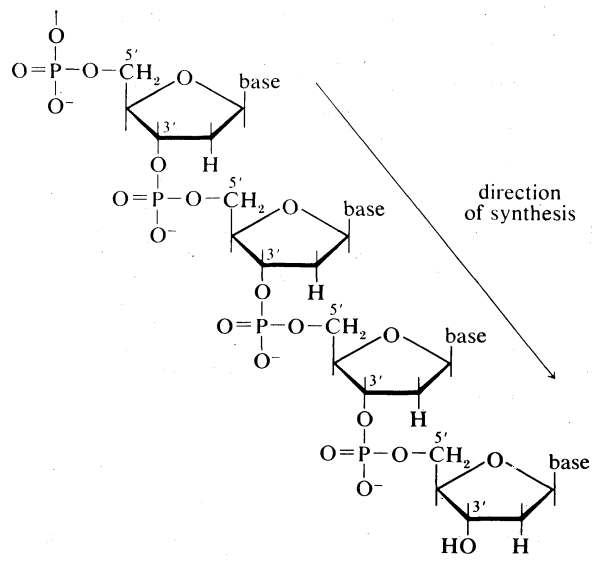
Nucleic acids and proteins. As with the synthesis of polysaccharides and lipids, the formation of the nucleic acids and proteins from their building blocks requires the input of energy. Nucleic acids are formed from nucleoside triphosphates, with concomitant elimination of inorganic pyrophosphate, which is subsequently hydrolyzed via reaction [21a]. Amino acids also are activated, forming, at the expense of ATP, aminoacyl-complexes. This activation process is also accompanied by loss of inorganic pyrophosphate. But, although these biochemical processes are basically similar to those involved in the biosynthesis of other macromolecules, their occurrence is specifically subservient to the genetic information in DNA. DNA contains within its structure the blueprint both for its own exact duplication and for the synthesis of a number of types of RNA, among which is a class termed messenger RNA (mRNA). A complementary relationship exists between the sequence of purines (*i.e.*, adenine and guanosine) and pyrimidines (cytosine and thymine) in the DNA comprising a gene and the sequence in mRNA into which this genetic information is transcribed. This information is then translated into the sequence of amino acids in a protein, a process that involves the functioning of a variety of other classes of ribonucleic acids (see GENETICS AND HEREDITY, THE PRINCIPLES OF).

Formation of nucleic acids

Synthesis of DNA. The maintenance of genetic integrity demands not only that enzymes exist for the synthesis of DNA but that they function so as to ensure the replication of the genetic information (encoded in the DNA to be copied) with absolute fidelity. This implies that the assembly of new regions of a DNA molecule must occur on a template of DNA already present in the cell. The synthetic processes must also be capable of repairing limited regions of DNA, which may have been damaged, for example, as a consequence of exposure to ultraviolet irradiation. The physical structure of DNA is ideally adapted to its biological roles. Two strands of nucleotides are wound around each other in the form of a double helix. The helix is stabilized by hydrogen bonds that occur between the purine and pyrimidine bases of the strands. Thus, the adenine of one strand pairs with the thymine of the other, and the guanine of one strand with the cytosine of the other. The base pairs may be visualized as the treads of a spiral staircase, in which the two chains of repeating units (*i.e.*, ribose-phosphate-ribose) form the sides.

Biosynthesis of DNA

During the biosynthesis of DNA, the two strands unwind, and each serves as a template for the synthesis of a new, complementary strand, in which the bases pair in exactly the same manner as occurred in the parent double helix. The process is catalyzed by a DNA polymerase enzyme, which catalyzes the addition of the appropriate deoxyribonucleoside triphosphate (NTP) in [86] onto one end, specifically, the free 3′-hydroxyl end (—OH) of the growing DNA chain (see diagram of DNA strand). In [86] the addition of a deoxyribonucleoside monophosphate



DNA strand

(dNMP) moiety onto a growing DNA chain (5′-DNA-polymer-3′-OH) is shown; the other product is inorganic pyrophosphate. The specific nucleotide inserted in the growing chain is dictated by the base in the complementary (template) strand of DNA with which it pairs. The functioning of DNA polymerase thus requires the presence of all four deoxyribonucleoside triphosphates (*i.e.,* dATP, dTPP, dGTP, and dCTP) as well as preformed DNA to act as a template. Although a number of DNA polymerase enzymes have been purified from different organisms, it is not yet certain whether those that have been most extensively studied are necessarily involved in the formation of new DNA molecules, or whether they are primarily concerned with the repair of damaged regions of molecules. A polynucleotide ligase that effects the formation of the phosphate bond between adjacent sugar molecules is concerned with the repair function but may also have a role in synthesis.

5′-DNA-polymer-3′-OH + dNTP
⟶ 5′-DNA-polymer-dNMP-3′-OH + PP_i [86]

The varieties of RNA

Synthesis of RNA. Various types of RNA are found in living organisms: messenger RNA (mRNA) is involved in the immediate transcription of regions of DNA; transfer RNA (tRNA) is concerned with the incorporation of amino acids into proteins; and structural RNA is found in the ribosomes that form the protein-synthesizing machinery of the cell. In cells of organisms with well-defined nuclei (*i.e.,* eukaryotes), a heterogenous RNA fraction of unknown function is constantly broken down and resynthesized in the nucleus of the cell but does not leave it. The different types of RNA are synthesized via RNA polymerases [87], the action of which is analogous to that of the DNA polymerases that catalyze reaction [86]. In [87] the growing RNA chain is represented by 5′-RNA-polymer-3′-OH, and the ribonucleoside triphosphate by NTP. One product (5′-RNA-polymer-NMP-3′-OH) reflects the incorporation of ribonucleoside monophosphate; the other product is, as in [86], inorganic pyrophosphate. Synthesis of RNA requires DNA as a template, thus ensuring that the base composition of the RNA faithfully reflects that of the DNA; in addition, as in DNA synthesis, all four nucleoside triphosphates must be present. The major differences between reactions [86] and [87] are that, in the latter, the nucleotides contain ribose instead of deoxyribose, and that, in RNA, uracil replaces the thymine of DNA.

It appears that, although only one strand of the DNA double helix serves as template during the formation of RNA, some regions are transcribed from one strand, some from the other.

5′-RNA-polymer-3′-OH + NTP
⟶ 5′-RNA-polymer-NMP-3′-OH + PP_i [87]

An important constraint on RNA synthesis is that the accurate copying of the appropriate DNA strand by RNA polymerase must start at the beginning of a gene—and not somewhere along it—and must stop as soon as the genetic information has been transcribed. The way in which this selectivity is achieved is not yet fully understood, although it has been established that *E. coli* contains a protein, the sigma factor, that is not required for the incorporation of the nucleoside triphosphates into the growing RNA chain but apparently is essential for binding RNA polymerase to the proper DNA sites to initiate RNA synthesis. After the initiation step, the sigma factor is released; the role of the sigma factor in transcription suggests that the DNA at the initiation sites must be unique in some way so as to ensure that the correct strand is used as the template. Evidence indicates further that other protein factors are involved in the termination of transcription.

Synthesis of proteins. Approximately 120 macromolecules are involved directly or indirectly in the process of the translation of the base sequence of a messenger RNA molecule into the amino-acid sequence of a protein. The relationship between the base sequence and the amino-acid sequence constitutes the genetic code. The basic properties of the code are: it is triplet—*i.e.,* a linear sequence of three bases in mRNA specifies one amino acid in a protein; it is nonoverlapping—*i.e.,* each triplet is discrete and does not overlap either neighbour; it is degenerate—*i.e.,* many of the 20 amino acids are specified by more than one of the 64 possible triplets of bases; and it appears to apply universally to all living organisms.

The main sequence of events associated with the expression of this genetic code, as elucidated for *E. coli,* may be summarized as follows (see also GENETICS AND HEREDITY, THE PRINCIPLES OF).

1. Messenger RNA binds to the smaller of two subunits of large particles termed ribosomes.

2. The amino acid that begins the assembly of the protein chain is activated and transferred to a specific transfer RNA (tRNA). The activation step, catalyzed by an aminoacyl-tRNA synthetase specific for a particular amino acid, effects the formation of an aminoacyl-AMP complex [88a] in a manner somewhat analogous to reaction [77]; ATP is required, and inorganic pyrophosphate is a product. The aminoacyl-AMP, which remains bound to the enzyme, is transferred to a specific molecule of tRNA in a reaction catalyzed by the same enzyme. AMP is released, and the other product is called aminoacyl-tRNA [88b]. In *E. coli* the amino acid that begins the assembly of the protein is always formylmethionine (f-Met). There is no evidence that f-Met is involved in protein synthesis in eukaryotic cells.

amino acid + ATP
⟶ aminoacyl—Ⓟ—ribose—adenine + PP_i [88a]
aminoacyl-AMP

aminoacyl—AMP + tRNA
⟶ aminoacyl—tRNA + AMP [88b]

3. Aminoacyl-tRNA binds to the mRNA-ribosomal complex in a reaction in which energy is provided by the hydrolysis of GTP to GDP and inorganic phosphate. In this step and in 5 below, the genetic code is translated. All of the different tRNAs contain triplets of bases that pair specifically with the complementary base triplets in mRNA; the base triplets in mRNA specify the amino acids to be added to the protein chain. During or shortly after the pairing occurs the aminoacyl-tRNA moves from the aminoacyl-acceptor (A) site on the ribosome to another site, called a peptidyl-donor (P) site.

Translation

4. The larger subunit of the ribosome then joins the mRNA-f-Met-tRNA-smaller ribosomal subunit complex.

5. The second amino acid to be added to the protein chain is specified by the triplet of bases adjacent to the initiator triplet in mRNA. The amino acid is activated and transferred to its tRNA by a repetition of reactions

[88a] and [88b]. This newly formed aminoacyl-tRNA now binds to the A site of the mRNA–ribosome complex, with concomitant hydrolysis of GTP.

6. The enzyme peptidyl transferase, which is part of the larger of the two ribosomal subunits, catalyzes the transfer of formylmethionine from the tRNA to which it is attached (designated $tRNA^{f\text{-}Met}$) to the second amino acid; for example, if the second amino acid were leucine, step 5 would have achieved the binding of leucyl-tRNA (Leu–$tRNA^{Leu}$) next to f-Met–$tRNA^{f\text{-}Met}$ on the ribosome–mRNA complex. Step 6 catalyzes the transfer reaction that is shown in [89], in which $tRNA^{f\text{-}Met}$ is released from formyl-methionine (f-Met), and Leu–$tRNA^{Leu}$ is bound to formyl-methionine.

$$\text{f-Met}-tRNA^{f\text{-}Met} + \text{Leu}-tRNA^{Leu} \longrightarrow \text{f-Met}-\text{Leu}-tRNA^{Leu} + tRNA^{f\text{-}Met} \quad [89]$$

7. In the next step three results are achieved. The dipeptide f-Met–Leu (a dipeptide consists of two amino acids) moves from the A (aminoacyl-acceptor) site to the P (peptidyl-donor) site on the ribosome; the $tRNA^{f\text{-}Met}$ is thereby displaced from the P site, and the ribosome moves the length of one triplet (three bases) along the mRNA molecule. The occurrence of these events is accompanied by the hydrolysis of a second molecule of GTP and leaves the system ready to receive the next aminoacyl-tRNA (by repetition of step 5). The cycle of events in 5, 6, and 7 is repeated until the ribosome moves to a triplet on the mRNA that does not specify an amino acid but provides the signal for termination of the amino-acid chain. Triplets of this type are represented by one uracil (U) preceding, and adjacent to, two adenines (UAA) or preceding one adenine and one guanosine in either order (UGA, or UAG).

Termination of the amino-acid chain

8. At the termination of synthesis the completed protein is released from the tRNA to which it had remained linked. Two further events then occur in *E. coli*. First, the formyl constituent of the f-methionyl moiety is hydrolyzed by the catalytic action of a formylase, producing a protein with methionine at the end. If the required protein does not contain methionine in this position (and the majority of proteins in *E. coli* appear to), the methionine and possibly other amino acids that follow it are removed by enzymatic reactions. Second, the ribosome–mRNA complex dissociates, and the ribosomal subunits become available for a new round of translation by binding another mRNA molecule, step 1.

For the sake of brevity, other ancillary protein factors that participate in this sequence 1 to 8 have been omitted; the role of many of these factors is as yet poorly understood.

Regulation of metabolism

FINE CONTROL

The flux of nutrients along each metabolic pathway is governed chiefly by two factors: (1) the availability of substrates on which pacemaker, or key, enzymes of the pathway can act and (2) the intracellular levels of specific metabolites that affect the reaction rates of pacemaker enzymes. Key enzymes are usually complex proteins that, in addition to the site at which the catalytic process occurs (*i.e.*, the active site), contain sites to which the regulatory metabolites bind. Interactions with the appropriate molecules at these regulatory sites cause changes in the shape of the enzyme molecule. Such changes may either facilitate or hinder the changes that occur at the active site. The rate of the enzymatic reaction is thus speeded up or slowed down by the presence of a regulatory metabolite.

In many cases, the specific small molecules that bind to the regulatory sites have no obvious structural similarity to the substrates of the enzymes; these small molecules are therefore termed allosteric effectors, and the regulatory sites are termed allosteric sites (see the article BIOCHEMICAL COMPONENTS OF ORGANISMS). Allosteric effectors may be formed by enzyme-catalyzed reactions in the same pathway in which the enzyme regulated by the effectors functions. In this case a rise in the level of the allosteric effector would affect the flux of nutrients along that pathway in a manner analogous to the feedback phenomena of homeostatic processes. Such effectors may also be formed by enzymatic reactions in apparently unrelated pathways. In this instance the rate at which one metabolic pathway operates would be profoundly affected by the rate of nutrient flux along another. It is this situation that, to a large extent, governs the sensitive and immediately responsive coordination of the many metabolic routes in the cell.

End-product inhibition. A biosynthetic pathway is usually controlled by an allosteric effector produced as the end product of that pathway, and the pacemaker enzyme on which the effector acts usually catalyzes the first step that uniquely leads to the end product. This phenomenon, called end-product inhibition, is illustrated by the multienzyme, branched pathway for the formation from oxaloacetate of the "aspartate family" of amino acids (Figure 10). The system of interlocking controls is described in greater detail in Figure 12. As mentioned previously in this article, only plants and microorganisms can synthesize many of these amino acids, most animals requiring such amino acids to be supplied preformed in their diets.

Figure 12 shows that there are a number of pacemaker enzymes in the biosynthetic route for the aspartate family of amino acids, most of which are uniquely involved in the formation of one product. Each of the enzymes functions after a branch point in the pathway, and all are inhibited specifically by the end product that emerges from the branch point. It is not difficult to visualize from Figure 12 how the supplies of lysine, methionine, and isoleucine required by a cell can be independently regulated. Threonine, however, is both an amino acid essential for protein synthesis and a precursor of isoleucine. If the rate of synthesis of threonine from aspartate were regulated as are the rates of lysine, methionine, and isoleucine, an imbalance in the supply of isoleucine might result. This risk is overcome in *E. coli* by the existence of three different aspartokinase enzymes, all of which catalyze the first step common to the production of all the products derived from aspartate. Each has a different regulatory effector molecule. Thus, one type of aspartokinase is inhibited by lysine, a second by threonine. The third kinase is not inhibited by any naturally occurring amino acid, but its rate of synthesis (see below) is controlled by the concentration of methionine within the cell. The triple control mechanism resulting from the three different aspartokinases ensures that the accumulation of one amino acid does not shut off the supply of aspartyl phosphate necessary for the synthesis of the others.

Another example of control through end-product inhibition also illustrates the manner in which the operation of two biosynthetic pathways may be coordinated. Both DNA and the various types of RNA are assembled from purine and pyrimidine nucleotides (see above *Nucleic acids and proteins*); these in turn are built up from intermediates of central metabolic pathways (see above *Mononucleotides*). The first step in the synthesis of pyrimidine nucleotides is that catalyzed by aspartate carbamoyltransferase [70a]. This step initiates a sequence of reactions that leads to the formation of pyrimidine nucleotides such as UTP and CTP [74]. Studies of aspartate carbamoyltransferase have

Aspartate carbamoyltransferase

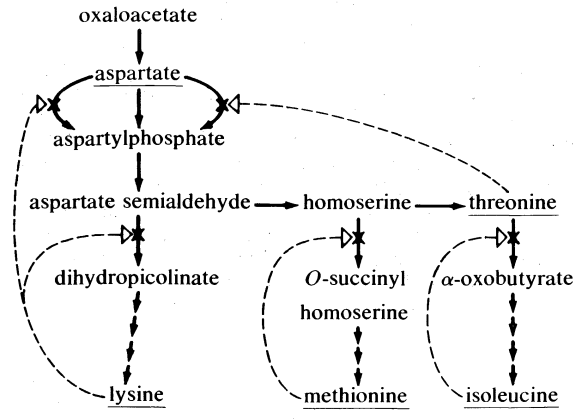


Figure 12: "Fine control" of the enzymes of the aspartate family in *E. coli* (see text).

revealed that the affinity of this enzyme for its substrate (aspartate) is markedly decreased by the presence of CTP. This effect can be overcome by the addition of ATP, a purine nucleotide. The enzyme can be dissociated into two subunits: one contains the enzymatic activity and (in the dissociated form) does not bind CTP; the other binds CTP but has no catalytic activity. Apart from providing physical evidence that pacemaker enzymes contain distinct catalytic and regulatory sites, the interaction of aspartate carbamoyltransferase with the different nucleotides provides an explanation for the control of the supply of nucleic acid precursors. If a cell contains sufficient pyrimidine nucleotides (*e.g.*, UTP), aspartate carbamoyltransferase, the first enzyme of pyrimidine biosynthesis, is inhibited. If, however, the cell contains high levels of purine nucleotides (*e.g.*, ATP), as required for the formation of nucleic acids, the inhibition of aspartate carbamoyltransferase is relieved, and pyrimidines are formed.

Positive modulation. Not all pacemaker enzymes are controlled by inhibition of their activity. Instead, some are subject to positive modulation—*i.e.*, the effector is required for the efficient functioning of the enzyme. Such enzymes exhibit little activity in the absence of the appropriate allosteric effector. One instance of positive modulation is the anaplerotic fixation of carbon dioxide onto pyruvate and phosphoenolpyruvate (PEP); this example also illustrates how a metabolic product of one route controls the rate of nutrient flow of another (see Figure 9).

The carboxylation of pyruvate in higher organisms [50] and the carboxylation of phosphoenolpyruvate in gut bacteria [50a] occurs at a significant rate only if acetyl coenzyme A is present. Acetyl coenzyme A acts as a positive allosteric effector and is not broken down in the course of the reaction. Moreover, some pyruvate carboxylases [50] and the PEP carboxylase of gut bacteria are inhibited by four-carbon compounds (*e.g.*, aspartate). These substances inhibit because they interfere with the binding of the positive effector, acetyl coenzyme A. Such enzymatic controls are reasonable in a physiological sense: it will be recalled that anaplerotic formation of oxaloacetate from pyruvate or PEP is required to provide the acceptor for the entry of acetyl coenzyme A into the TCA cycle. The reaction need occur only if acetyl coenzyme A is present in sufficient amounts. On the other hand, an abundance of four-carbon intermediates obviates the necessity for forming more through carboxylation reactions such as [50] and [50a].

Similar reasoning, though in the opposite sense, can be applied to the control of another anaplerotic sequence, the glyoxylate cycle (Figure 8). The biosynthesis of cell materials from the two-carbon compound acetate is, in principle, akin to biosynthesis from TCA cycle intermediates. In both processes, it is the availability of intermediates such as PEP and pyruvate that determines the rate at which a cell forms the many components produced through gluconeogenesis. Although in the strictest sense the glyoxylate cycle has no defined end product, PEP and pyruvate are, for these physiological reasons, best fitted to regulate the rate at which the glyoxylate cycle is required to operate. It is thus not unexpected that the pacemaker enzyme of the glyoxylate cycle, isocitrate lyase (reaction [52]), is allosterically inhibited by PEP and by pyruvate.

Rate of the glyoxylate cycle

Energy state of the cell. It is characteristic of catabolic routes that they do not lead to uniquely identifiable end products. The major products of glycolysis and the TCA cycle, for example, are carbon dioxide and water. Within the cell, the concentrations of both are unlikely to vary sufficiently to allow them to serve as effective regulatory metabolites. The processes by which water is produced (Figure 7) initially involve, however, the reduction of coenzymes, the reoxidation of which is accompanied by the synthesis of ATP from ADP. Moreover, as described in previous sections, the utilization of ATP in energy-consuming reactions yields ADP and AMP. At any given moment, therefore, a living cell contains ATP, ADP, and AMP; the relative proportion of the three nucleotides provides an index of the energy state of the cell. It is thus reasonable that the flux of nutrients through catabolic routes is, in general, impeded by high intracellular levels of both reduced coenzymes (*e.g.*, $FADH_2$, reduced NAD^+) and ATP, and that these inhibitory effects are often overcome by AMP.

The control exerted by the levels of ATP, ADP, and AMP within the cell is illustrated by the regulatory mechanisms of glycolysis and the TCA cycle (Figure 9); these nucleotides also serve to govern the occurrence of the opposite pathway, gluconeogenesis, and to avoid mutual interference of the catabolic and anabolic sequences. Although not all of the controls mentioned below have been found to operate in all living organisms examined, it has been observed that, in general:

1. Glucose 6-phosphate stimulates glycogen synthesis from glucose 1-phosphate [79] and inhibits both glycogen breakdown [16] and its own formation from glucose [1].
2. Phosphofructokinase, the most important pacemaker enzyme of glycolysis [3], is inhibited by high levels of its own substrates (fructose 6-phosphate and ATP); this inhibition is overcome by AMP. In tissues, such as heart muscle, which use fatty acids as a major fuel, inhibition of glycolysis by citrate may be physiologically the more important means of control. Control by citrate, the first intermediate of the TCA cycle, which produces the bulk of the cellular ATP, is thus the same, in principle, as control through ATP.
3. Fructose 1,6-diphosphatase [59], which catalyzes the reaction opposite to phosphofructokinase, is strongly inhibited by AMP.
4. Rapid catabolism of carbohydrate requires the efficient conversion of PEP to pyruvate. In liver and in some bacteria the activity of the pyruvate kinase that catalyzes this process [10] is greatly stimulated by the presence of fructose 1,6-diphosphate, which thus acts as a potentiator of a reaction required for its ultimate catabolism.
5. The oxidation of pyruvate to acetyl coenzyme A [37] is inhibited by acetyl coenzyme A. Because acetyl coenzyme A also acts as a positive modulator of pyruvate carboxylation [50], this control reinforces the partition between pyruvate catabolism and its conversion to four-carbon intermediates for anaplerosis and gluconeogenesis.

Inhibition of pyruvate oxidation

6. Citrate synthase [38], the first enzyme of the TCA cycle, is inhibited by ATP in higher organisms and by reduced NAD^+ in many microorganisms. In some strictly aerobic bacteria, the inhibition by reduced NAD^+ is overcome by AMP.
7. Citrate acts as a positive effector for the first enzyme of fatty acid biosynthesis [62]. A high level of citrate, which also indicates a sufficient energy supply, thus inhibits carbohydrate fragmentation (see [2]) and diverts the carbohydrate that has been fragmented from combustion to the formation of lipids.
8. Some forms of isocitrate dehydrogenase [40] are maximally active only in the presence of ADP or AMP and are inhibited by ATP. This is an example of regulation by covalent modification of an enzyme since the action of ATP here is to phosphorylate, and consequently to inactivate, the isocitrate dehydrogenase. A specific phosphatase, which is a different enzymatic activity of the protein that effects the phosphorylation by ATP, catalyzes the splitting-off by water of the phosphate moiety on the inactive isocitrate dehydrogenase and thus restricts activity. Again, the energy state of the cell serves as the signal regulating an enzyme involved in energy transduction.

COARSE CONTROL

Although fine control mechanisms allow the sensitive adjustment of the flux of nutrients along metabolic pathways relative to the needs of cells under relatively constant environmental conditions, these processes may not be adequate to cope with severe changes in the chemical milieu.

Such severe changes may arise in higher organisms with a change in diet or when, in response to other stimuli, the hormonal balance is altered (see BIOCHEMICAL COMPONENTS OF ORGANISMS). In starvation, for example, the overriding need to maintain blood glucose levels may require the liver to synthesize glucose from noncarbohydrate products of tissue breakdown at rates greater than can be achieved by the enzymes normally present in the liver. Under such circumstances, cellular concentrations of key enzymes of gluconeogenesis, such as pyruvate carboxylase

[50] and PEP carboxykinase [54], may rise by as much as 10-fold, while the concentration of glucokinase [1] and of the enzymes of fatty acid synthesis decreases to a similar extent. Conversely, high carbohydrate diets and administration of the hormone insulin to diabetic animals elicit a preferential synthesis of glucokinase [1] and pyruvate kinase [10]. These changes in the relative proportions and absolute amounts of key enzymes are the net result of increases in the rate of their synthesis and decreases in the rate of their destruction. Although such changes reflect changes in the rates of either transcription, translation, or both of specific regions of the genome, the mechanisms by which the changes are effected have not yet been clarified.

Microorganisms sometimes encounter changes in environment much more severe than those encountered by the cells of tissues and organs, and their responses are correspondingly greater. Mention has already been made of the ability of *E. coli* to form β-galactosidase when transferred to a medium containing lactose as the sole carbon source (see *Integration of catabolism and anabolism*); such a transfer may result in an increase of 1,000-fold or more in the cellular concentration of the enzyme. Because this preferential enzyme synthesis is elicited by exposure of the cells to lactose, or to non-metabolizable but chemically similar analogues, and because synthesis ceases as soon as the eliciting agents (inducers) are removed, β-galactosidase is termed an inducible enzyme. It has been established that a regulator gene exists that specifies the amino-acid sequence of a so-called repressor protein, and that the repressor protein binds to a unique portion of the region of DNA concerned with β-galactosidase formation. Under these circumstances the DNA is not transcribed to mRNA, and virtually no enzyme is made. The repressor, however, is an allosteric protein and readily combines with inducers. Such a combination prevents the repressor from binding to DNA and allows transcription and translation of β-galactosidase to proceed.

β-galactosidase

Although this mechanism for the specific control of gene activity may not apply to the regulation of all inducible enzymes—for example, those concerned with the utilization of the sugar arabinose—and is not universally applicable to all coarse control processes in all microorganisms, it can explain the manner in which the presence in growth media of at least some cell components represses (*i.e.*, inhibits the synthesis of) enzymes normally involved in the formation of such components by gut bacteria such as *E. coli*. Although, for example, the bacteria must obviously make amino acids from ammonia if that is the sole source of nitrogen available to them, it would not be necessary for the bacteria to synthesize enzymes required for the formation of amino acids supplied preformed in the medium. Thus, of the three aspartokinases formed by *E. coli* (Figure 12), two are repressed by their end products, methionine and lysine. On the other hand, the third aspartokinase, which (as described above) is inhibited by threonine, is repressed by threonine only if isoleucine is also present. This example of so-called multivalent repression is of obvious physiological utility. It is likely that the amino acids that thus specifically inhibit the synthesis of aspartokinases do so by combining with specific protein repressor molecules; however, whereas the combination of the inducer with the repressor of β-galactosidase inactivates the repressor protein and hence permits synthesis of the enzyme, the repressor proteins for biosynthetic enzymes would not bind to DNA unless they were also combined with the appropriate amino acid. Aspartokinase synthesis would thus occur in the absence of the end-product effectors and not in their presence.

This explanation applies also to the coarse control of the anaplerotic glyoxylate cycle (Figure 8). The synthesis of both of the enzymes unique to that cycle, isocitrate lyase [52] and malate synthase [53], is controlled by a regulator gene that presumably specifies a repressor protein unable to bind to DNA unless combined with pyruvate or PEP. Cells growing on acetate do not contain high levels of these intermediates because they are continuously being removed for biosynthesis. The enzymes of the glyoxylate cycle are therefore formed at high rates. If pyruvate or substances catabolized to PEP or pyruvate are added to the medium, however, further synthesis of the two enzymes is speedily repressed. (Ha.Ko.)

A regulator gene

Metabolic diseases

GENERAL CONSIDERATIONS

Metabolic diseases are conditions caused by an abnormality in one or more of the chemical reactions essential to producing energy, to regenerating cellular constituents, or to eliminating unneeded products arising from these processes. Depending on which metabolic pathway is involved, a single defective chemical reaction may produce consequences that are narrow, involving a single body function, or broad, affecting many organs and systems.

In addition to the fine and coarse mechanisms that affect individual chemical reactions (see above *Regulation of metabolism*), the rate and direction of metabolic processes are regulated by the availability of substrates (particularly essential amino acids or fatty acids), cofactors (whose presence is required to assist the enzyme), and key minerals; and by the influence of hormones secreted by endocrine glands and tissues. Abnormalities in any of these areas affect the many metabolic pathways that are driven by the catalytic activities of different enzymes. Nutritional and endocrine diseases are, in fact, the most common causes of metabolic dysfunction (see ENDOCRINE SYSTEMS), and diseases involving thyroid hormones and insulin, as well as caloric deprivation, illustrate the impact such abnormalities may have on the overall metabolic cycle.

The thyroid gland secretes two hormones, thyroxine and triiodothyronine, that regulate the overall rate of oxygen utilization and hence the production of useful energy from the combustion of food. When an excess of thyroid hormone is present (hyperthyroidism), the rate of oxidation of sugars, fatty acids, and amino acids is above normal. Simultaneously, as more oxygen is used, more carbon dioxide (CO_2) and urea are formed as the end products of metabolism. The demands for calories to supply the extra substrates and of vitamins as enzyme cofactors increase. Despite this, there is usually a loss of body weight from both fat and muscle tissue, the latter composed primarily of protein. Much of the extra energy generated is in the form of heat, and perspiration and evaporation of water are increased in order to maintain normal body temperature. Because the body requires more oxygen, the rates of respiration and blood flow increase to allow the extra oxygen to be delivered to tissues and to remove the extra carbon dioxide. A deficiency of thyroid hormone (hypothyroidism) produces exactly the opposite effects.

Another major hormone that influences metabolism is insulin, which is secreted by special cells of the pancreas. Insulin primarily regulates the direction of metabolism, shifting many processes toward the storage of substrates and away from their degradation. Insulin acts to increase the transport of glucose and amino acids as well as key minerals such as potassium, magnesium, and phosphate from the blood into cells. It also regulates a variety of enzymatic reactions within the cells, all of which have a common overall direction, namely the synthesis of large molecules from small units. A deficiency in the action of insulin (diabetes mellitus) causes severe impairment in (1) the storage of glucose in the form of glycogen and the oxidation of glucose for energy; (2) the synthesis and storage of fat from fatty acids and their precursors and the completion of fatty-acid oxidation; and (3) the synthesis of proteins from amino acids. As a result, high levels of glucose, fatty acids, keto acid products of incomplete fatty-acid oxidation, and essential amino acids are present in the blood. (The consequences of these abnormalities in metabolism are described in ENDOCRINE SYSTEMS.) Replacements for deficient hormones and methods to eliminate or combat excess hormones have been developed.

Insulin

A fundamental example of a nutritional cause for metabolic disease is caloric deprivation. In the most extreme case of fasting, the average human can survive about 60 days by oxidizing body stores of energy. Extremely obese persons can survive up to a year without food because of greater stores of fat and protein. During fasting, blood levels of glucose are low and the oxidation

of glucose is curtailed, even by the brain tissue for which it is required. The blood levels of fatty acids and of their products of combustion, keto acids, are raised; keto acids partially take the place of glucose as a substrate in the brain. As the proteins of organs are degraded and not replaced, organ function declines, eventually resulting in death. Partial deprivation of calories results in a more gradual weight loss, a decline in oxygen use and energy production, a decrease in muscle strength, and lassitude. In infants and children, growth is impaired, and they are susceptible to infection and early death. In those children in whom calories are adequate but protein is chronically deficient (kwashiorkor), growth virtually ceases and the synthesis of important proteins, such as albumin, is reduced. Low levels of protein in the blood allow water to accumulate in tissues, causing the face, extremities, and abdomen to swell (edema). The immune system produces fewer cells to combat invading organisms, leading to infection and possibly death. Less severe problems include changes in skin and hair colour and texture and delayed sexual maturation. Treatment with protein or with the missing amino acids reverses the disorder.

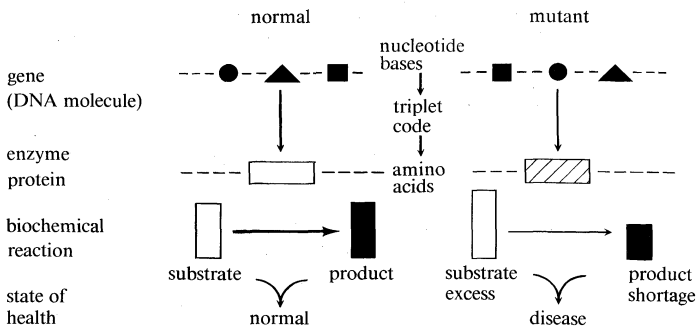


Figure 13: *Pathological sequence of inborn errors of metabolism.*
A change in the triplet nucleotide base sequence occurs at one point of the mutant gene, and the altered sequence codes for the "wrong" amino acid at a critical point in the structure of the enzyme protein. Because the mutant enzyme has little or no catalytic activity, its biochemical reaction is greatly diminished, causing a shortage of the product and an accumulation of the substrate.

THE DERIVATION OF SPECIFIC METABOLIC DISORDERS

Most metabolic diseases are caused by single defects in particular biochemical pathways, defects that are due to the deficient activity of individual enzymes. Around the turn of the century a British physician, Sir Archibald Garrod, proposed that such diseases were passed through generations, and he labeled them "inborn errors of metabolism." Subsequent studies have demonstrated the mechanisms by which a biochemical defect can be inherited. Beadle and Taum studied mutations in fruit flies and microorganisms and demonstrated that a single gene controls a single type of biochemical reaction. An American chemist, Linus Pauling, and a British biochemist, Vernon Ingram, showed that a mutation may be expressed as a specific alteration in the primary structure of a single type of protein.

Inborn errors of metabolism

Each enzyme is a protein and, like other proteins, has a unique sequence of amino acids. The order of the amino acids is directed by a segment of the DNA molecule called a gene. Within the gene, a sequence of three nucleotide bases directs each amino acid. If the nucleotide bases of the gene are rearranged, deleted, or substituted, the mutant gene directs the synthesis of a mutant enzyme protein, which may have only one amino acid substituted or large segments of amino acids rearranged. Changes in structure usually decrease or entirely eliminate an enzyme's catalytic activity, slowing the rate of the biochemical reaction catalyzed by the enzyme. As a result, the amount of the product of the reaction is low, while that of the starting material (substrate) is high (Figure 13).

The potential consequences of mutant enzymes on cellular or organ function are varied: (1) the product of the affected metabolic pathway may have a key role that is not fulfilled; (2) a metabolic pathway that is repressed by normal levels of the product may proceed, with toxic consequences, when that product is at low levels; (3) excess substrate from the affected reaction may be toxic; (4) the excess substrate may become involved in and move rapidly through an alternate metabolic pathway, and the product of this pathway may be toxic. About 170 errors of metabolism can be traced to mutations in the genetic code, and the consequences range from trivial to lethal.

The nucleus holds the DNA molecules in an amorphous strand called chromatin. When the cell is ready to divide and pass along the genetic code to a newly formed cell, the chromatin condenses into 46 gene-containing units, called chromosomes, which are grouped in 23 homologous pairs. One pair determines sex (the sex chromosomes), and the remaining 22 pairs (autosomes) determine the non-sex-related features of the progeny. In fertilization, 23 chromosomes (one-half of each pair) are donated by one parent, and 23 chromosomes are donated by the other parent; the zygote (fertilized egg) then contains 23 chromosome pairs (46 chromosomes).

Chromosomes

The effect of a mutant gene inherited from one parent depends on whether the other parent donates a similar mutant gene (homozygous) or a normal gene (heterozygous). In general, heterozygotes (those with only one mutant gene) have about 50 percent of the normal enzyme activity because only one gene is normal, although this may be sufficient to maintain good health. When, however, the gene directs the synthesis of an enzyme that is rate limiting (*i.e.*, a series of reactions will proceed at a normal rate only as long as this enzyme is available), then even a 50 percent reduction in enzyme activity may produce serious disease. This gene and the disease that results are called autosomal dominant. Homozygotes (those with two mutant genes) suffer from more severe adverse effects because the mutant enzyme has little or no normal activity. When two mutant genes must be paired to produce a disease, the gene and the disease are autosomal recessive. These basic points are summarized in Table 1.

If the mutant gene is part of the X chromosome (a sex chromosome; females have two X chromosomes per pair, and males have an X chromosome and a Y chromosome), the disease is sex-linked. All male offspring are affected because the Y chromosome of the XY pair does not have a compensating normal gene. Because the mutation is on the X chromosome, however, and males transmit only the Y chromosome to their sons during fertilization, males do not transmit the disease to them. They do, however, transmit the carrier state (*i.e.*, the mutant X chromosome) to their daughters. If the X-linked disease is dominant, both heterozygote and homozygote females are affected, the latter more severely; if the disease is recessive, only female homozygotes are affected. Female heterozygotes transmit an X chromosome with a mutant gene to half their male offspring, prompting a disease, and an X chromosome with a mutant gene to half of their female offspring, who may in turn transmit the disease to their sons. Occasion-

Table 1: Hereditary Aspects of Metabolic Disease

subject	gene 1	gene 2	enzyme activity	disease	parents
normal	normal	normal	100%	no	one, both, or neither may be heterozygote
heterozygote	normal	mutant	about 50%	no if recessive, yes if dominant	one must be at least heterozygote
homozygote	mutant	mutant	near 0%	yes	both must be at least heterozygote

ally, only one child in a family has a metabolic disorder and there is no evidence of an abnormal gene in either parent. The abnormal gene in the child must be assumed to have arisen by a spontaneous mutation.

Two different enzyme deficiencies, operating at separate steps in the same metabolic pathway and coded for by two entirely different genes, can produce similar diseases. Alternatively, more than one type of mutant gene (allele) may have arisen at the same gene location (locus) so that diseases may arise in heterozygotes who have two different mutant genes. Some of the generalized metabolic diseases, such as diabetes, may be polygenic (*i.e.*, the disease is not fully expressed unless two or more gene loci—even on different chromosomes—are affected). Finally, certain gene-related metabolic disorders may not be expressed without a particular type of diet, hormonal stimulus, or drug.

Investigation of an inborn error

If an inborn error of metabolism is suspected, a physician can investigate at several levels: (1) a thorough family history is obtained to analyze the pattern of inheritance; (2) blood or urine is analyzed to determine whether the substrate and the product of the suspect biochemical reaction are at normal levels; (3) the activity of the suspect enzyme may be measured in red or white blood cells, in tissue samples, or in skin cells grown in test tube cultures; (4) the number of enzyme molecules may be measured in blood or tissue samples; (5) the DNA molecule that codes for the suspected abnormal enzyme may be isolated from blood cells, and its nucleotide base sequence determined and identified as mutant; and (6) testing also may be applied to parents, siblings, or offspring of the affected person. The diagnosis of a treatable condition may be made through screening at birth. A severe metabolic disorder may be diagnosed in the fetus by using samples from the amniotic fluid surrounding the embryo. The information is then available for counseling or sometimes for initiating treatment before birth.

There are a variety of treatments for metabolic disorders: (1) If the product of the reaction is deficient, it can be provided; (2) if the high level of substrate is toxic, the substrate can be diminished by restricting the diet, by activating its removal from the body, or by preventing its synthesis at an earlier step; (3) enzyme activity may be increased by administering a cofactor or certain drugs; and (4) the enzyme deficiency may be replaced with normal enzyme, although this approach has been successful only rarely. The ideal treatment would be to insert the normal gene itself into the cells of the patient so that normal amounts of the enzyme can be synthesized and the metabolic defect corrected directly. There are techniques for transplanting genes into animals, and they are being studied for humans.

DISORDERS OF CARBOHYDRATE METABOLISM

Glycogen storage diseases. The role of glucose as a critical source of energy for the brain and as the initial source of energy for muscles requires that it be stored in an easily releasable form. The storage form of glucose is a large, branched molecule known as glycogen. The liver serves as the glycogen (and therefore glucose) depot for the brain, while muscles maintain a glycogen supply of their own. Glycogen is synthesized from, and broken down to, glucose as shown in Figure 14.

Types

There are at least 10 well-recognized varieties of glycogen storage diseases, all of which are autosomal recessive disorders (Figure 14). In each, glycogen accumulates in the liver or muscles and sometimes in the kidneys. In type I glycogen storage disease, the last step in glucose release from the liver is defective, leading to low levels of glucose in the blood (hypoglycemia). While hypoglycemia may not bring about symptoms in some persons, it may ultimately produce coma or convulsions in others. If the rate of conversion of the immediate precursor of glucose, glucose 6-phosphate, to lactic acid increases, acidosis (greater acidity of the blood) results. A compensatory mechanism that diverts amino acids away from protein synthesis and toward glucose synthesis (gluconeogenesis) leads to stunted growth. Treatment of type I disease consists of frequent daytime feedings augmented by glucose and protein delivery at night. This regimen protects the brain from hypoglycemia, reduces the size of the liver, and restores growth.

Defects in earlier steps in glycogen breakdown in the liver cause glycogen storage disease types III, VI, VIII, IX, and X (Figure 14), which usually lead to milder versions of type I disease. Each responds to similar treatment. In type IV disease, an abnormal unbranched glycogen molecule (amylopectin) accumulates and causes severe liver failure. In glycogen storage disease type II, an entirely separate pathway for degrading the glycogen present in cell storage organelles, called lysosomes, is defective, resulting in the accumulation of glycogen, in the heart, muscles, and liver; in severe cases this leads to early death from heart failure. In glycogen storage disease type V, the first step in muscle glycogen degradation is deficient, preventing the release of glucose. Strenuous exercise must be curtailed, and muscle cramps and even muscle damage can occur. In glycogen storage disease type VII, glycolysis (breakdown of glucose) is deficient in muscle, producing similar symptoms and an accumulation of glycogen.

Galactosemia. A classical autosomal recessive disorder, galactosemia usually is due to a defective component of the second major step in the metabolism of the sugar galactose (Figure 14). When galactose is ingested, as in milk, galactose 1-phosphate accumulates, causing severe liver damage as well as kidney abnormalities, mental retardation, and hypoglycemia. An alternative product of galactose metabolism, the sugar galactitol, is deposited in

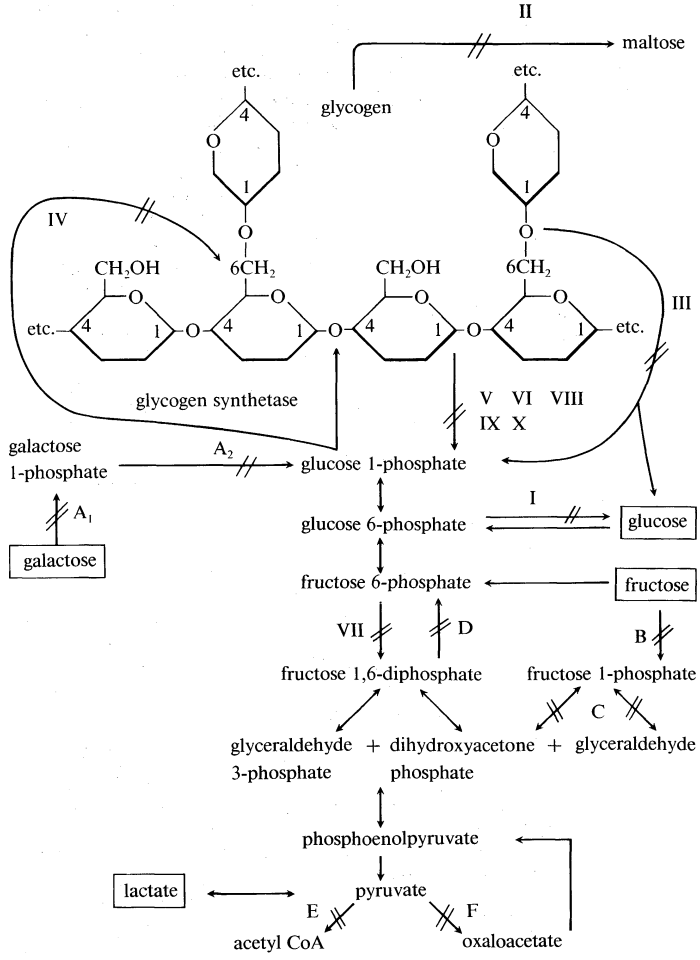


Figure 14: *(I–X) Enzyme defects in glycogen storage diseases and (A₁–F) enzyme defects in other carbohydrate diseases.* (I) Glucose-6-phosphatase (von Gierke). (II) Lysosomal α-1,4-glucosidase (Pompe). (III) Debranching enzyme (Forbes). (IV) Branching enzyme (Andersen). (V) Muscle phosphorylase (McArdle). (VI) Liver phosphorylase (Hers, VIII, IX, X). (VII) Muscle phosphofructokinase (Tauri). (A_1) Galactokinase. (A_2) Galactose 1-phosphate UDP transferase (galactosemia). (B) Fructokinase (fructosuria). (C) Aldolase (hereditary fructose intolerance). (D) Fructose 1,6-diphosphatase deficiency. (E) Pyruvate dehydrogenase complex deficiency. (F) Pyruvate carboxylase deficiency.

the lens of the eye, causing cataracts. A defect in the first step of galactose metabolism causes only the accumulation of galactose itself and cataracts. Early, complete, and permanent elimination of galactose from the diet prevents these adverse effects.

Fructose disorders. Fructosuria is an autosomal recessive disorder caused by a defect in the first step in the metabolism of the sugar fructose (Figure 14). Fructose is therefore excreted in the urine, but there are no other consequences and the condition is harmless. In contrast, in the autosomal recessive disorder hereditary fructose intolerance, a deficiency of a later key enzyme in fructose metabolism leads to the accumulation of fructose 1-phosphate, a toxic substance that causes vomiting and hypoglycemia initially and liver damage over a long period of time. Eliminating fructose, in the form of fruit or sucrose (table sugar), is effective treatment. Fructose 1,6-diphosphatase deficiency is an autosomal recessive disorder that prevents the synthesis of glucose from its precursors (Figure 14), causing hypoglycemia after a short period of fasting. Ingestion of fructose also causes hypoglycemia and lactic acidosis. Treatment consists of frequent feeding and avoiding fructose.

Hypoglycemia

Pyruvate disorders. Two autosomal recessive disorders, pyruvate dehydrogenase complex deficiency and pyruvate carboxylase deficiency (Figure 14), can cause the accumulation of a three-carbon metabolite, pyruvate, in the blood. As a result, the level of lactic acid in the blood rises, causing acidosis. Both disorders also cause abnormalities of the central nervous system and retardation. In addition, pyruvate carboxylase deficiency causes hypoglycemia because glucose is not synthesized (Figure 14).

Intestinal carbohydrate malabsorption. Carbohydrates are ingested primarily in the form of larger molecules, such as starch (the chief storage form of carbohydrates in plants), from which the smaller sugars eventually are produced by the actions of intestinal enzymes (oligosaccharidases), such as maltase, which splits maltose to glucose; lactase, which splits lactose (milk sugar) to glucose and galactose; and sucrase, which splits sucrose (table sugar) to glucose and fructose. Lactase deficiency is common, occurring in up to 20 percent of North American whites, and 50 to 90 percent of blacks and Asians. Although hereditary, this disorder usually does not show symptoms until after infancy. Abdominal fullness, nausea, and cramping pain with watery diarrhea follow the intake of milk products containing lactose. Similar symptoms occur with a deficiency of sucrase after the ingestion of table sugar. In these conditions, glucose levels do not increase in the blood after lactose or sucrose, respectively, are ingested. These symptoms are eliminated when lactose or sucrose, respectively, are avoided.

Mucopolysaccharidoses. Sugar molecules may come together in repeating units to form polymers known as mucopolysaccharides. When proportionately more sugar polymers are bound to a protein core, the molecule is known as a proteoglycan. When the molecule contains proportionately more protein than sugar polymers, it is called a glycoprotein. All of these perform diverse functions, particularly as basic components of the material between the tissue cells (the matrix). Mucopolysaccharides are regularly degraded as part of the process of tissue renewal. If any one of 15 enzymes responsible for this degradation is abnormal, partially degraded mucopolysaccharides accumulate in the tissues, causing recessive diseases with similar characteristics (Hurler, Hunter, Sanfilippo, and Morquio syndromes). These characteristics may include short stature, various deformities of the skeleton, stiffened joints, corneal clouding, deafness, dental abnormalities, hardening of the arteries, enlarged liver and spleen, and mental retardation. These syndromes are diagnosed by analyzing the incompletely degraded mucopolysaccharides in the urine and tissues. Prenatal diagnosis is also possible. No treatment is available for any of these disorders.

DISORDERS OF LIPID METABOLISM

Blood lipid disorders. Lipid molecules, such as triglycerides and cholesterol, circulate in the blood and are bound to carrier protein molecules. In certain hereditary metabolic disorders, the levels of triglycerides and cholesterol in the blood are high because the mechanisms that remove them are defective. As a result, triglycerides and cholesterol may be deposited in the skin, tendons, and walls of blood vessels, which can lead to coronary heart disease and early death.

In familial dysbetalipoproteinemia (type III hyperlipemia, autosomal recessive), a mutant carrier protein causes the accumulation of triglycerides and cholesterol. The disease is treated by restricting fat and cholesterol in the diet and with drugs that lower the level of lipids in the blood. In familial hypercholesterolemia (type II hyperlipemia, autosomal dominant), a protein receptor on the surface of all cells, which allows cholesterol to move into the cell, is either missing or defective. As a result, cholesterol is not moved into the cells, where mechanisms are in place to register the availability of cholesterol to metabolic processes. Under normal conditions, when enough cholesterol is registered, feedback mechanisms signal the cholesterol-synthesizing enzymes to cease cholesterol synthesis. In the disease state, the synthesizing enzymes are relieved of feedback inhibition, thus inducing the synthesis of still more cholesterol. Homozygotes (one per 1,000,000 persons) have much higher levels of cholesterol in the blood than heterozygotes (one per 500 persons) and may die before the age of 20. Restricting cholesterol in the diet and drug treatment sufficiently lower cholesterol in heterozygotes but not in homozygotes. Although relatively uncommon, familial hypercholesterolemia is an important model for the association between a high prevalence of coronary heart disease and modest elevations of blood cholesterol levels, which is often found in the West. In familial lipoprotein lipase deficiency (type I hyperlipemia, autosomal recessive), the enzyme responsible for transporting triglycerides into cells is deficient. Triglycerides build up in the blood, and the blood serum takes on a creamy consistency and colour after the ingestion of fat. There is abdominal pain, inflammation of the pancreas, and enlargement of the liver, but there is no increased risk of coronary heart disease. Restricting dietary fat eliminates the symptoms.

Familial dysbetalipoproteinemia

In familial betalipoprotein deficiency (autosomal recessive), the absence of a major transport protein leads to low levels of triglyceride and cholesterol in the blood. Dietary fat and lipid-soluble vitamins, particularly vitamin E, are not adequately absorbed. Vitamin E deficiency may be partly responsible for the resulting severe dysfunction of the nervous system, the eyes, and the muscles.

In familial high-density lipoprotein deficiency (Tangier's disease, autosomal recessive), the level of cholesterol in the blood is low, while that of triglycerides is high or normal. Cholesterol is deposited in many tissues, including those of the nerves, cornea, lymph nodes, and enlarged yellow tonsils. There is no treatment.

In Wolman's disease, cholesterol esters accumulate in many tissues because the specific enzymes that cleave them are absent. Massive enlargement of the liver and spleen, calcification of the adrenal glands, failure to grow, and death in infancy are characteristic of this disease.

Wolman's disease

Lipid oxidative disorders. The oxidation of fatty acids for energy requires that they be transferred into the mitochondria by associating with a cofactor, carnitine, which is synthesized in the body. Deficiency of carnitine or of the enzyme carnitine acyltransferase (which facilitates the movement of fatty acids across the mitochondrial membrane) in muscle tissue leads to weakness and cramping with exercise. A low level of carnitine in the blood and tissues produces severe brain, liver, and heart damage. The exact defect leading to this autosomal recessive disorder is not known. Treatment with carnitine is partially effective.

Tissue lipid disorders. Complex molecules composed of long-chain fatty acids linked with various sugars, phosphate, and sulfate groups form essential components of cell membranes and are vital to the normal functioning of the nervous system. These molecules are degraded and resynthesized in the normal course of tissue turnover. Defects in the enzymes that regulate this process cause these lipid molecules to accumulate, which is associated with a number of severe autosomal recessive disorders (Figure 15).

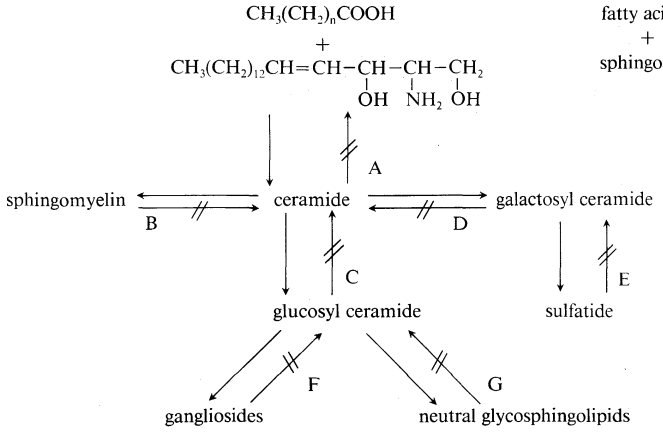


Figure 15: *Enzyme defects in tissue lipid disorders.* (A) Ceramidase (Farber). (B) Sphingomyelinase (Niemann-Pick). (C) Glucocerebrosidase (Gaucher). (D) Galactocerebroside β-galactosidase (Krabbe). (E) Arylsulfatase A (metachromatic leukodystrophy). (F) Hexosaminidase A (Tay-Sachs). (G) α-Galactosidase A (Fabry).

Among the symptoms are nervous system dysfunction and mental retardation; a unique cherry-red spot in the centre of the inner eye (Tay-Sachs disease); enlargement of the liver and spleen; characteristic foamy, lipid-filled cells in the bone marrow (Gaucher's and Niemann-Pick disease); or arthritis (Farber's disease). All but Gaucher's disease may cause early death. Fabry's disease is an X-linked disorder involving predominantly the blood vessels, kidneys, and eyes. For most of these conditions, the defective genes can be identified in carriers, and a prenatal diagnosis can be made. There is no successful treatment.

DISORDERS OF AMINO-ACID METABOLISM

Twenty amino acids, including nine that cannot be synthesized in humans and therefore must be obtained through food, are involved in metabolism. All are incorporated into proteins; some also are synthesized into important molecules in the body such as neurotransmitters, hormones, pigments, oxygen-carrying molecules, and the nucleotide bases of DNA and RNA. Each amino acid is further broken down to ammonia, carbon dioxide (CO_2), and water (H_2O). About 50 metabolic defects in amino-acid metabolism are known (Table 2), and their clinical consequences range from none to lethal.

Phenylalanine and tyrosine. The disorders involving the essential amino acids phenylalanine and tyrosine are examples of the many consequences of amino-acid dysfunction. Figure 16 demonstrates how these amino acids are synthesized and broken down into important molecules. Phenylketonuria is caused by the absence of the enzyme phenylalanine hydroxylase, which converts phenylalanine to tyrosine. Other variants (hyperphenylalaninemia IV and V) are due to defects in the supply of the cofactor required for this reaction. Phenylketonuria produces severe mental retardation, seizures, rash, and high levels of phenylalanine in the blood and of its metabolite, phenylpyruvic acid, in the urine. Screening of blood at birth and restriction of phenylalanine in the diet are preventive measures.

Effects of phenylketonuria

Tyrosinemia results from a deficiency of the enzyme that catalyzes the first, and rate-limiting, step in tyrosine degradation, tyrosine transaminase (also called tyrosine aminotransferase). Tyrosinemia causes erosions of the skin and cornea of the eye. Deficiency in the next step causes a transient and harmless elevation of the level of tyrosine in the blood. A defect in the third successive enzymatic step causes an accumulation of homogentisic acid (which darkens the urine) and is called alkaptonuria. Although rare, alkaptonuria is of historical importance because it was Garrod's studies of this condition that led to his one-gene : one-enzyme hypothesis (one gene codes for one enzyme). When homogentisic acid is oxidized, the products are deposited in cartilage and connective tissue, producing a blue-gray colour in some areas of the skin and black discoloration in the joints; in adults, a severe form of arthritis, called ochronosis, results.

A disorder that results from a defect in the last step of tyrosine degradation, tyrosinosis, is a more generalized condition involving elevations of the levels of the amino acids methionine and tyrosine in the blood. Sudden liver failure may lead to early death; alternatively, there may be progressive kidney dysfunction in the form of Fanconi's syndrome (see below *Disorders of amino-acid transport*).

Tyrosine is a parent amino acid for skin, hair, and eye pigments (Figure 16). There are at least 10 syndromes, known generally as oculocutaneous albinism, that are characterized by the failure to form melanin pigments, resulting in partial or complete albinism. In addition to having white hair and white skin, persons with these defects are hypersensitive to light, and they have decreased vision and aberrant control of eye muscles. Oculocutaneous albinism can be autosomal recessive, autosomal dominant, or X-linked. A defect in the first step leading toward melanin synthesis, catalyzed by the enzyme tyrosinase, accounts for some of the syndromes; in others, the site of the block in pigment production is not known. The only available treatment is to avoid sunlight and to correct vision.

Finally, tyrosine is also the precursor amino acid for hormones of the thyroid gland. Defects in this synthetic pathway result in familial hypothyroidism, compensatory enlargement of the thyroid gland (goitre), severe growth failure, and retardation of central nervous system development. Screening of newborns and replacement of deficient thyroid hormone in affected persons prevent all or most of this damage.

Urea cycle enzymes. When amino acids are degraded, the ammonia nitrogen at one end of the molecule is split off, incorporated into urea (the chief solid component of

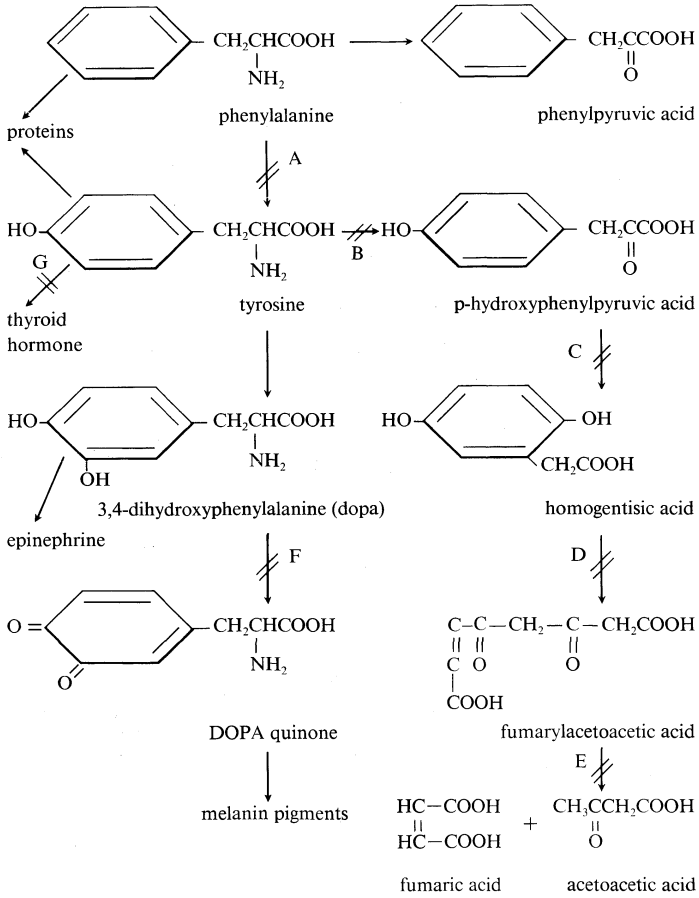


Figure 16: *Enzyme defects in disorders of the amino acids phenylalanine and tyrosine.* (A) Phenylalanine hydroxylase (phenylketonuria). (B) Tyrosine transaminase (tyrosinemia). (C) p-Hydroxyphenylpyruvate oxidase (tyrosinemia). (D) Homogentisic acid oxidase (alkaptonuria). (E) Fumarylacetoacetate hydrolase (tyrosinosis). (F) Tyrosinase (albinism). (G) Peroxidase and others (familial goitre).

Table 2: Disorders of Amino-Acid Metabolism

amino-acid disorder	enzyme deficiency	physical abnormalities	prenatal diagnosis feasible	early postnatal test available
Aromatic amino acids				
Phenylketonuria	phenylalanine hydroxylase	see text		X
Hyperphenylalanemia, types II and III	phenylalanine hydroxylase	see text		X
Hyperphenylalanemia, type IV	dihydropteridine reductase	see text		X
Hyperphenylalanemia, type V	dihydrobiopterin synthetase	see text		X
Tyrosinemia, type II	tyrosine transaminase	see text		X
Alkaptonuria	homogentisic acid oxidase	see text		X
Tyrosinosis	fumarylacetoacetate hydrolase	see text		
Oculocutaneous albinism, tyrosinase-negative type	tyrosinase	see text		X
Oculocutaneous albinism, tyrosinase-positive type S	unknown	see text		X
Familial goitre	peroxidase, deiodinase, and others	see text		X
Urea cycle				
Carbamoyl phosphate synthetase deficiency	carbamoyl phosphate synthetase	see text		X
Ornithine carbamoyltransferase* deficiency	ornithine carbamoyltransferase	see text	X	X
Citrullinemia	argininosuccinate synthetase	see text	X	X
Argininosuccic acidemia	argininosuccinate lyase	see text	X	X
Hyperargininemia	arginase	see text		X
Branched chain amino acids				
Hypervalinemia	valine transaminase	see text	X	X
Hyperleucine-isoleucinemia	leucine-isoleucine transaminase	see text		X
Maple syrup urine disease (branched-chain ketoaciduria)	branched-chain α-keto acid decarboxylase	see text	X	X
Isovaleric acidemia	isovaleryl CoA dehydrogenase	see text	X	X
Glutaric aciduria, type II	acyl CoA dehydrogenases	see text		
3-Hydroxy-3-methylglutaric aciduria	3-hydroxy-3-methylglutaryl CoA lyase	see text	X	X
3-Ketothiolase deficiency	acetoacetyl CoA 3-ketothiolase	see text		
Propionic acidemia	propionyl CoA carboxylase	see text	X	X
Methylmalonic acidemia	methylmalonyl CoA mutase	see text	X	X
Others				
Histidinemia	histidase	speech disorder, mental retardation	X	X
5-Oxoprolinuria	glutathione synthetase	nervous system disorder, acidosis	X	X
Hyperprolinemia	proline oxidase or Δ^1-pyrroline-5-carboxylate dehydrogenase	benign		X
Hyperhydroxyprolinemia	4-hydroxy-L-proline oxidase	benign		X
Prolidase deficiency (hyperimidodipeptiduria)	prolidase (imidodipeptidase)	rash, mental retardation		
Hyperornithinemia gyrate atrophy of the choroid and retina	ornithine δ-transaminase	visual disturbance, mental retardation	X	X
Hyperlysinemia, persistent	L-lysine: ketoglutarate reductase	uncertain	X	X
Homocystinuria	cystathionine β-synthetase	abnormal lens and bones, blood vessel clotting, mental retardation	X	X
Cystathioninuria	γ-cystathionase	benign	X	X
Cystinosis	unknown	kidney failure, cataracts	X	X
Hypersarcosinemia	sarcosine dehydrogenase complex	uncertain		X
Nonketotic hyperglycinemia	glycine cleavage enzyme complex	mental retardation, seizures		X
Primary oxaluria	α-ketoglutarate-glyoxylate carboligase	kidney stones and failure		X
Hyper-β-alaninemia	β-alanine transaminase	nervous system dysfunction		X
Carnosinase deficiency	serum carnosinase	uncertain		X
Homocarnosinosis	brain homocarnosinase	uncertain		X
Dihydrofolate reductase deficiency	dihydrofolate reductase	anemia, nervous system dysfunction		
Methylenetetrahydrofolate reductase deficiency	methylenetetrahydrofolate reductase	nervous system dysfunction		

*X-linked, dominant.

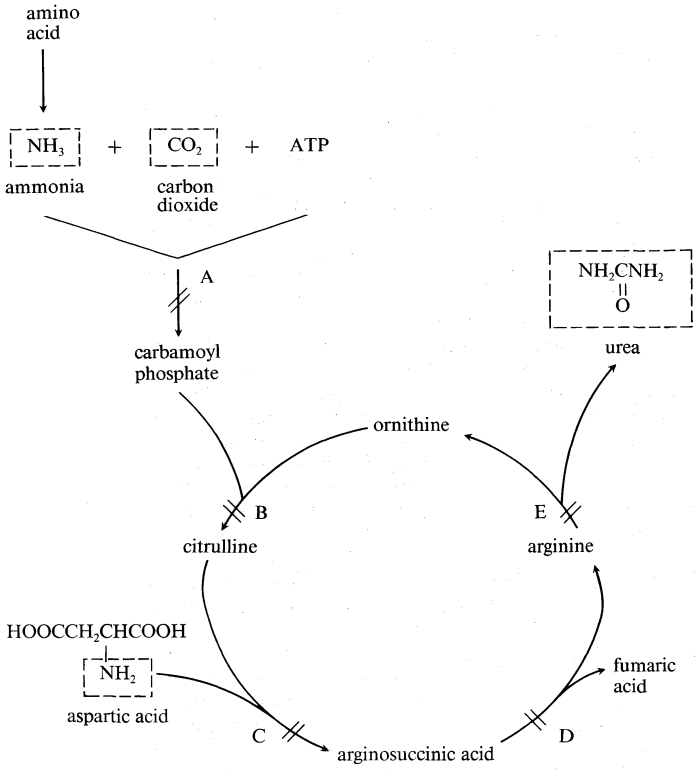


Figure 17: *Enzyme defects in urea cycle disorders.* (A) Carbamoyl phosphate synthetase. (B) Ornithine carbamoyltransferase. (C) Arginosuccinate synthetase. (D) Arginosuccinate lyase. (E) Arginase.

Formation of urea

urine), and excreted in the urine (Figure 17). This process starts when an ammonia group is combined with bicarbonate and phosphate to create carbamoyl phosphate, which hooks into a metabolic cycle involving five enzymes. The cycle takes in a second ammonia group from the amino acid aspartic acid, ultimately generating the amino acid arginine from which urea is split off. Defects at any of the first four steps of the cycle may cause a sudden increase in the level of ammonia in the blood, which in the newborn causes vomiting, coma, liver enlargement, and mental retardation. Early death can result when the enzyme is almost completely absent. A defect in the fifth step produces motor difficulties and seizures beginning in childhood. Specific amino-acid supplements, drugs that remove nitrogen by combining with the amino acids glycine and glutamine, and a diet low in protein are of limited benefit.

Branched chain amino acids. The amino acids valine, leucine, and isoleucine are degraded initially by removing the amino group and producing CO_2, followed by reduction. Once degraded, however, their paths diverge. There are at least 11 genetic defects in the shared and nonshared portions of these pathways (Table 2). The consequences often include acidosis, hypoglycemia, and developmental retardation. In four of these conditions, the presence of incompletely degraded products in the urine is associated with distinctive odours, similar to those of maple syrup and sweaty feet. The appropriate amino acid or acids are restricted to a level that relieves the symptoms but still permits growth. Alkali, in the form of sodium bicarbonate, is given when the acidity of the blood increases suddenly.

Disorders of amino-acid transport. Energy is required to move many amino acids from the intestinal tract into the blood or to reclaim them from the urine by special cells in the kidney. This transport of amino acids is mediated by specific carrier proteins whose structures, like those of enzymes, are determined by genes. One gene may contain the code for the carrier protein that transports several amino acids; in addition, a single carrier protein may transport the same amino acids in the kidney and the intestinal tract. Mutant proteins with decreased transport activities may prevent the absorption of dietary amino acids or cause their loss in the urine. For example, in cystinuria there is decreased intestinal absorption and increased urinary excretion of four amino acids (cystine, arginine, lysine, and ornithine). The only consequence, however, is the formation of kidney stones. The essential amino acid lysine is probably absorbed from the diet in other forms. Iminoglycinuria is a completely harmless disorder of the absorption tubules in the kidney; glycine, proline, and hydroxyproline are excessively excreted in the urine without apparent consequence. In Hartnup disease, 13 amino acids, six of them essential, are not adequately reabsorbed from the intestine and kidney because of defects in the mechanism that removes them. As a result, high levels of these amino acids are excreted in the urine. Because the levels of the amino acids are low and because they are converted to toxic products within the intestinal tract, persons with Hartnup disease suffer intermittently from rashes and neuropsychiatric disturbances. In Fanconi's syndrome, because amino acids as well as other substrates (glucose, phosphate, and minerals, for example) are not absorbed well from the kidney, they are excreted in the urine. The syndrome occurs with several metabolic diseases, including cystinosis (a disorder in which the amino acid cystine is stored in cells of most body tissues). Bone disease, disturbed growth, dehydration, potassium deficiency, and acidosis result.

DISORDERS OF PORPHYRIN METABOLISM

Porphyrin functions

Porphyrins are fluorescent cyclic molecules composed of four carbon-nitrogen rings that combine with specific metals, such as iron. These molecules, which carry oxygen in the blood (as hemoglobin) and in tissues, are also components of respiratory enzymes. Porphyrins are synthesized and degraded in the metabolic process shown in Figure 18. Defects in this process cause diseases, generally known as porphyrias, which primarily affect the skin (there is blistering on exposure to sunlight), the nervous system,

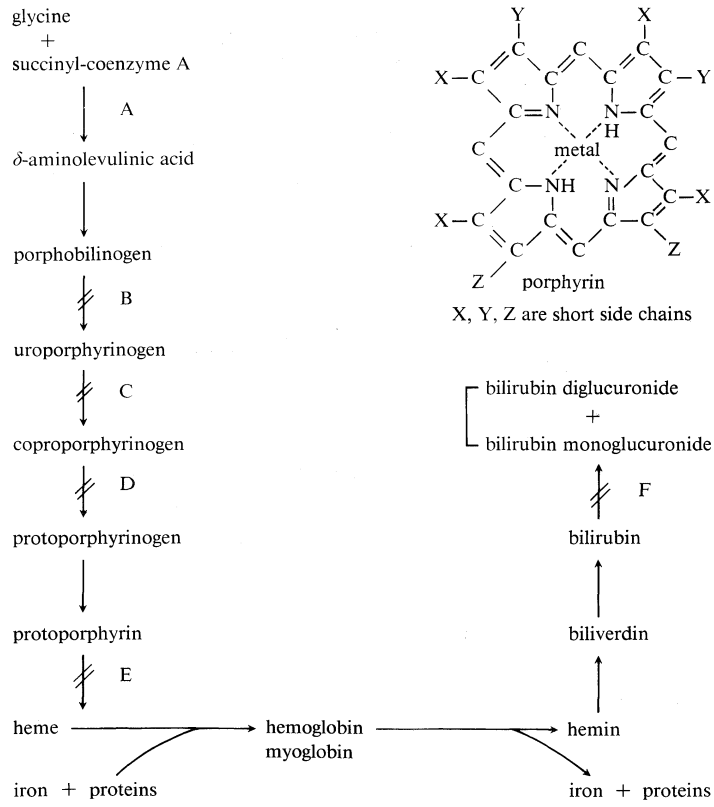


Figure 18: *Enzyme defects in porphyrin metabolism.* (A) δ-Aminolevulinic acid synthase. (B) Porphobilinogen deaminase, also called uroporphyrinogen I synthase (acute intermittent porphyria). (C) Uroporphyrinogen decarboxylase (porphyria cutanea tarda). (D) Coproporphyrinogen oxidase (hereditary coproporphyria). (E) Ferrochelatase (erythropoietic protoporphyria). (F) Bilirubin-UDP glucuronyltransferase (Crigler-Najjar and Gilbert).

and, in rare instances, the blood (rapid destruction of red blood cells). They are distinguished by the appearance of fluorescent products in the tissues and urine.

A dramatic form of porphyria is acute intermittent porphyria (AIP), which is caused by a deficiency of porphobilinogen deaminase, also called uroporphyrinogen I synthetase (Figure 18), an enzyme that catalyzes the second step in the pathway of porphyrin biosynthesis. Symptoms include abdominal pain, vomiting, muscle weakness, and mental disturbance. Porphobilinogen and its precursor, aminolevulinic acid, are found in the urine. AIP is an autosomal dominant disorder that occurs intermittently because the preceding reaction, catalyzed by the enzyme aminolevulinic acid synthetase, is the rate-limiting step in the pathway. Therefore, porphobilinogen does not reach toxic levels unless the activity of aminolevulinic acid synthetase is greatly increased by exposure to drugs, by starvation or excessive dieting, or by cyclic hormonal changes.

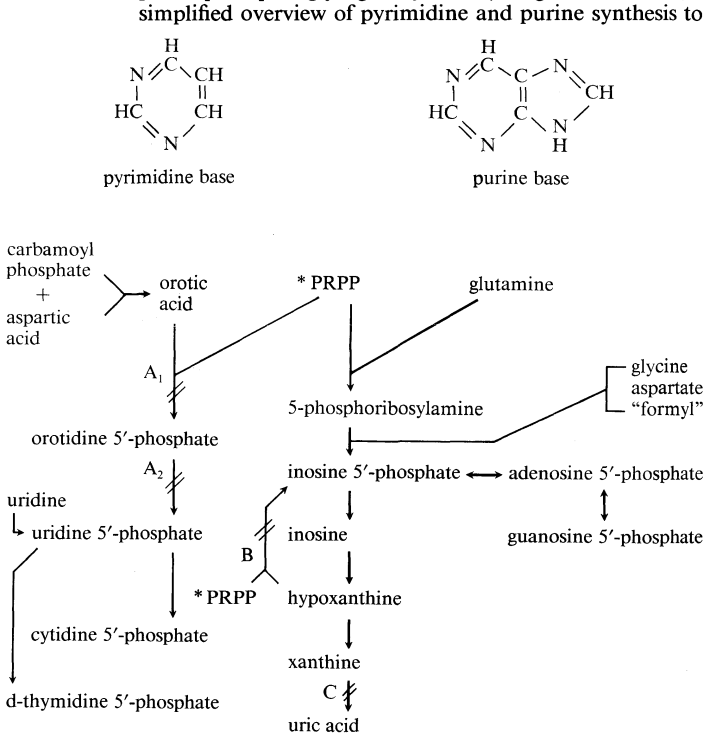
Red blood cells release hemoglobin (oxygen-carrying porphyrin molecules) when they die. The free porphyrins are degraded to a reddish yellow pigment known as bilirubin (Figure 18), which is transported to the liver. Defects in the mechanisms that remove and metabolize bilirubin increase the level of bilirubin in the blood and produce jaundice, a yellowing of the skin and eyes. A harmless defect in bilirubin metabolism is often seen for a short time in normal newborns and permanently in some adults. In infants without the enzyme bilirubin-UDP glucuronyl transferase (Crigler-Najjar syndrome), deposits of large amounts of bilirubin in the brain can cause severe and permanent damage to the nervous system. Treatment involves increasing enzyme activity by using the drug phenobarbitol or converting bilirubin to less toxic products by exposing the infant to blue light.

DISORDERS OF PURINE AND PYRIMIDINE METABOLISM

Nucleotide bases

Purines and pyrimidines are the nucleotide bases in DNA and RNA. They have important roles in energy transformations (for example, ATP in the TCA cycle) and in other metabolic processes (for example, uridine triphosphate [UTP] in glycogen synthesis). Figure 19 shows a simplified overview of pyrimidine and purine synthesis to the parent compounds, uridine monophosphate and inosine monophosphate, respectively, as well as the important degradative route for the latter. Orotic aciduria (autosomal recessive) results from a block in the last two steps of uridine monophosphate synthesis. It is unique as a metabolic disorder in that one gene appears to control the activity of two successive enzymatic steps. Arrested growth, mental retardation, and anemia result from the deficiency of uridine monophosphate. Treatment with the pyrimidine base uridine is effective because it can be converted to uridine monophosphate by an alternate pathway.

Lesch-Nyhan syndrome, an X-linked disorder, results from a deficiency of the enzyme hypoxanthine-guanine phosphoribosyltransferase, which reuses hypoxanthine, a product of inosine breakdown, for purine synthesis (Figure 19). If the enzyme is absent, the amount of unused key intermediate 5-phosphoribosyl 1-pyrophosphate (PRPP) is high, and the synthesis of inosine monophosphate continues (Figure 19), resulting in more hypoxanthine and, ultimately, uric acid. The nervous system is affected, resulting in mental retardation, abnormal movements, seizures, repeated self-mutilation, and death, usually by the age of 30. If the enzyme is only partially absent, gout results from the excess uric acid that settles in tissues. Gout also is a common disorder of purine metabolism in adults of the West, but in most cases the exact biochemical cause is unknown. The common denominator of all gouty conditions is high levels of uric acid in the blood. This leads to the formation of uric acid crystals in joint fluids, causing attacks of arthritis, and in the urine, causing kidney stones. The drug allopurinol, which inhibits the conversion of the immediate precursor, xanthine, to uric acid by inhibiting the enzyme xanthine oxidase (Figure 19), is effective in preventing these consequences. Hereditary xanthinuria results from a deficiency of xanthine oxidase, but is usually harmless or, at most, associated with kidney stones formed from xanthine. (S.Ge.)



* PRPP=5-phosphoribose 1-pyrophosphate

Figure 19: *Enzyme defects in disorders of pyrimidine and purine metabolism.*
(A_1) Orotate phosphoribosyltransferase (orotic aciduria). (A_2) Orotidine 5'-phosphate decarboxylase (orotic aciduria). (B) Hypoxanthine phosphoribosyltransferase (Lesch-Nyhan). (C) Xanthine oxidase (xanthinuria).

BIBLIOGRAPHY

General works: ALBERT L. LEHNINGER, *Principles of Biochemistry* (1982); LUBERT STRYER, *Biochemistry,* 2nd ed. (1981); EARLENE BROWN CUNNINGHAM, *Biochemistry: Mechanisms of Metabolism* (1978); JAY TEPPERMAN and HELEN M. TEPPERMAN, *Metabolic and Endocrine Physiology: An Introductory Text,* 5th ed. (1987); and S. DAGLEY and DONALD E. NICHOLSON, *An Introduction to Metabolic Pathways* (1970).

Cell metabolism: JAMES DARNELL, HARVEY LODISH, and DAVID BALTIMORE, *Molecular Cell Biology* (1986); T.A.V. SUBRAMANIAN (ed.), *Cell Metabolism, Growth and Environment,* 2 vol. (1986); W. BARTLEY, H.L. KORNBERG, and J.R. QUAYLE (eds.), *Essays in Cell Metabolism* (1970); JOHN M. DIETSCHY, ANTONIO M. GOTTO, JR., and JOSEPH A. ONTKO, *Disturbances in Lipid and Lipoprotein Metabolism* (1978); J. FRANK HENDERSON and A.R.P. PATERSON, *Nucleotide Metabolism: An Introduction* (1973); and DAVID A. BENDER, *Amino Acid Metabolism,* 2nd ed. (1985).

Regulation of metabolism: DANIEL E. ATKINSON, *Cellular Energy Metabolism and Its Regulation* (1977); E.A. NEWSHOLME and C. START, *Regulation in Metabolism* (1977); RONALD G. THURMAN, FREDERICK C. KAUFFMAN, and KURT JUNGERMANN (eds.), *Regulation of Hepatic Metabolism* (1986); and CHARLES ZAPSALIS and R. ANDERLE BECK, *Food Chemistry and Nutritional Biochemistry* (1985).

Genetics and metabolism: JAMES D. WATSON, *Molecular Biology of the Gene,* 4th ed. (1987); VINCENT M. RICCARDI, *The Genetic Approach to Human Disease* (1977); PHILIP F. BENSON and ANTHONY H. FENSOM, *Genetic Biochemical Disorders* (1985); AUBREY MILUNSKY (ed.), *Genetic Disorders and the Fetus: Diagnosis, Prevention, and Treatment,* 2nd ed. (1986); FORRESTER COCKBURN and RICHARD GITZELMAN, *Inborn Errors of Metabolism in Humans* (1982); and DAVID J. GALTON, *Molecular Genetics of Common Metabolic Disease* (1985).

Metabolic disorders: ARCHIBALD E. GARROD, *Inborn Errors of Metabolism* (1909, reprinted with a suppl. by H. HARRIS, 1963), a classic; JOHN B. STANBURY *et al.* (eds.), *The Metabolic Basis of Inherited Disease,* 5th ed. (1983); JOHN W. HARE (ed.), *Signs and Symptoms in Endocrine and Metabolic Disorders* (1986); PHILIP FELIG *et al.* (eds.), *Endocrinology and Metabolism* (1981); and F. DICKENS, P.J. RANDLE, and W.J. WHELAN (eds.), *Carbohydrate Metabolism and Its Disorders,* 3 vol. (1968–81). Special topics are treated in HENRIK GALBO, *Hormonal and Metabolic Adaptation to Exercise* (1983); and CHARLES S. LIEBER (ed.), *Metabolic Aspects of Alcoholism* (1977).

(Ha.Ko./S.Ge.)